Stereo- and Neural Network-Based Pedestrian Detection

Liang Zhao and Chuck Thorpe
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
email: {lzhao, cet}@cs.cmu.edu

Abstract

In this paper, we present a real-time pedestrian detection system that uses a pair of moving cameras to detect both stationary and moving pedestrians in crowded environments. This is achieved through stereo-based segmentation and neural network-based recognition. Stereo-based segmentation allows us to extract objects from a changing background; neural network-based recognition allows us to identify pedestrians in various poses, shapes, sizes, clothing, occlusion status. The experiments on a large number of urban street scenes demonstrate the feasibility of the approach in terms of pedestrian detection rate and frame processing rate.

1 Introduction

Object detection and pedestrian recognition is essential to avoid dangerous traffic situations. In a driver assistance system for warning the driver of potential collision with nearby objects—especially pedestrians, we have to accomplish the following two tasks in real time. The first is to separate foreground objects from the background; the second is to distinguish pedestrians from other objects in order to protect pedestrians in danger. The first task is a segmentation procedure, the second one a recognition procedure.

In this paper, we employ a video-rate stereo system [17] to provide range information for object detection and pedestrian recognition. Using stereo to guide pedestrian detection carries with it some distinct advantages over conventional techniques. First, it allows explicit occlusion analysis and is robust to illumination changes. Second, the real size of an object derived from the disparity map provides a more accurate classification metric than the image size of the object. Third, using stereo cameras can detect both stationary and moving objects. Fourth, computation time is significantly reduced by performing recognition where objects are detected; it is less likely to detect background area as pedestrian since detection is biased toward areas where objects are detected.

Neural networks have been successfully applied to many real-time intelligent vehicle systems [11, 12, 13]. In our system a neural network trained with the backpropagation algorithm is used to discriminate pedestrians from other objects. Unlike similar systems [1, 12, 13] which are limited to detecting walking people, our system uses shape features instead of motions cues to detect both moving and stationary pedestrians. Since neural networks can express highly nonlinear decision surfaces, they are especially appropriate to classify objects presenting high degree of shape variability. In this system, the trained neural network implicitly represents the appearance of pedestrians in various poses, postures, sizes, clothing, and occlusion situations; it performs pedestrian detection in real time. The experiments on a large number of urban street scenes demonstrate the feasibility of the approach in terms of recognition rate and frame processing rate.

This paper is organized as follows. Section 2 describes the related work on pedestrian detection. Section 3 presents the stereo guided object detection algorithm. The neural network based pedestrian recognition is presented in Section 4. Section 5 gives the experimental results, followed by a summary and conclusions in Section 6.

2 Related Work

Most human tracking and motion analysis systems [1, 7, 8] employ a simple segmentation procedure such as background subtraction or temporal differencing to detect pedestrians. A serious problem with these approaches is the dynamic background caused by illumination changes or background (or camera) motion. Some techniques such as Pfinder [2], W^4 [3], and path clustering [6], have been developed to compensate for small, or gradual changes in the scene or the lighting. However, they cannot deal with large, sudden changes in the background. Although optical flow [9] can be used to detect independently moving targets in the presence of camera motion, it is not feasible for non-rigid object extraction since the movements of the body parts are different. Above all, a common drawback with the above approaches is the assumption that all detected objects are pedestrians; this limits the generalization and application of these schemes.

More sophisticated pedestrian detection techniques include a recognition step to discriminate pedestrians from other objects. These techniques can be classified into motion-based, shape-based, and multi-cue-based methods. Most motion-based approaches [13, 24, 25, 26] use rhythmic features or motion patterns unique to human beings for pedestrian detec-

tion. However, there are several limitations with these schemes. First, the pedestrian's feet or legs should be visible in order to extract the rhythmic features. Second, the recognition procedure requires a sequence of images, which delays the identification until several frames later and increases the processing time. Third, the procedure cannot detect stationary pedestrians and pedestrians performing unconstrained and complex movement such as wandering around, turning, jumping, etc.

On the other hand, the shape-based approach relies on shape features to recognize pedestrians. Thus, this approach can detect both moving and stationary pedestrians from a single image. The primary difficulty in this approach is accommodating the wide range of variations in pedestrian appearance due to pose, non-rigid motion, lighting, clothing, occlusion, etc. [7, 8] use hand-crafted human models to detect pedestrians. An advantage of these methods is that they can analyze the motion of each body part; the disadvantage is that the segmentation of body parts is very difficult and even impossible in some situations. Lipton [23] depends on a dispersedness defined as the ratio $perimeter^2/area$ to classify human and vehicle. This classification metric is easy to calculate, but fails to distinguish humans from other objects with similar dispersedness and tends to misclassify pedestrians walking together as a vehicle. Papageorgious and Poggio [16] present a more robust pedestrian detection system based on wavelet analysis and the support vector machine (SVM) technique. However, the system has to search the whole image at multi-scales for pedestrians. This would be an extremely computationally expensive procedure, and it may cause multiple responses from a single pedestrian.

To increase reliability, some systems [10, 22] integrate multiple cues such as stereo, skin color, face, shape pattern to detect pedestrians. However, skin color is very sensitive to illumination changes [10]; face detection can identify only pedestrians facing the camera. These systems [10, 22] prove that stereo and shape are more reliable and helpful cues than color and face detection in general situations.

3 Stereo Guided Object Detection

Our driver warning system is equipped with a stereo-based object detection module that detects foreground objects in real-time. In the following subsections we first describe stereo analysis, then explain the range image segmentation for object detection.

3.1 Stereo Analysis

Real-time stereo systems [17, 19, 20, 21] have recently been available and applied to people detection [22, 18, 10]. Among these systems, we chose the Small Vision System (SVS) developed by SRI [17] to perform stereo analysis, because the SVS can run at video rates on a standard PC without specialized hardware. The SVS has the following four features. First, it computes a dense disparity map using area correlation after a Laplacian of Gaussian transform. Second, the disparity value can be searched at various disparity levels, e.g., 16, 24, or 32 pixels. Third, the SVS performs postfiltering with an interest operator, and a left- right consistency check. Fourth, the SVS does 4x range interpolation.

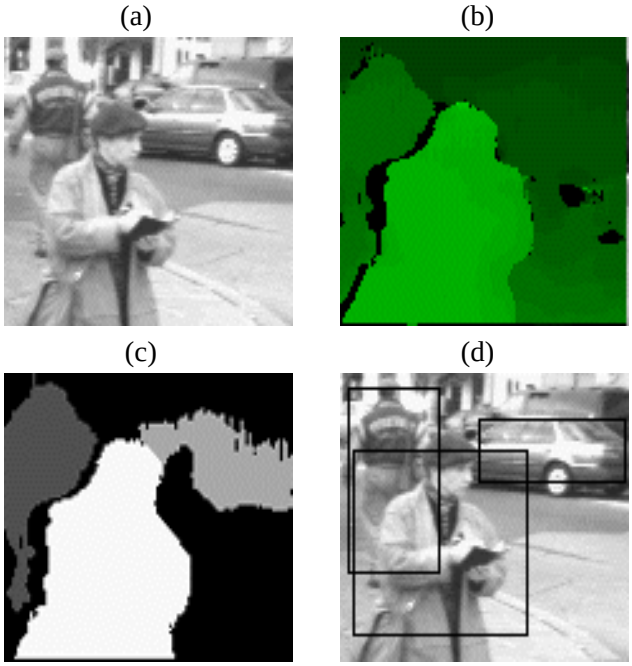

Figure 1: (a) the left image from stereo cameras (b) the disparity map (c) the segmentation result (d) the detected objects marked by boxes

Fig. 1(b) shows a typical disparity image produced by the SVS. Higher disparities (close objects) are brighter. There are 64 possible levels of disparity; disparity 0 (black areas) are regions where the range data is rejected by the post-processor interest operator due to insufficient texture.

3.2 Range Image Segmentation for Object Detection

Range information is a powerful cue for foreground/background segmentation. Compared with the intensity-based approach, range-based segmentation is less affected by light conditions, shadows and occlusion; compared with the 3D-based approach, it is less expensive computationally and suitable for real-time implementation. Most stereo-based segmentation algorithms [22, 18] assume a static background and a pair of stationary cameras, so they do not work properly with a changing scene.

In contrast, our segmentation and grouping technique requires no prior background model. The algorithm proceeds in several stages of processing as explained below. We first eliminate background objects from the disparity image by range thresholding. We then employ a morphological closing operator to remove the noise and to smooth the foreground regions. Then, a connected-component grouping operator is applied to find the foreground regions with smoothly varying range. Finally, small regions are eliminated

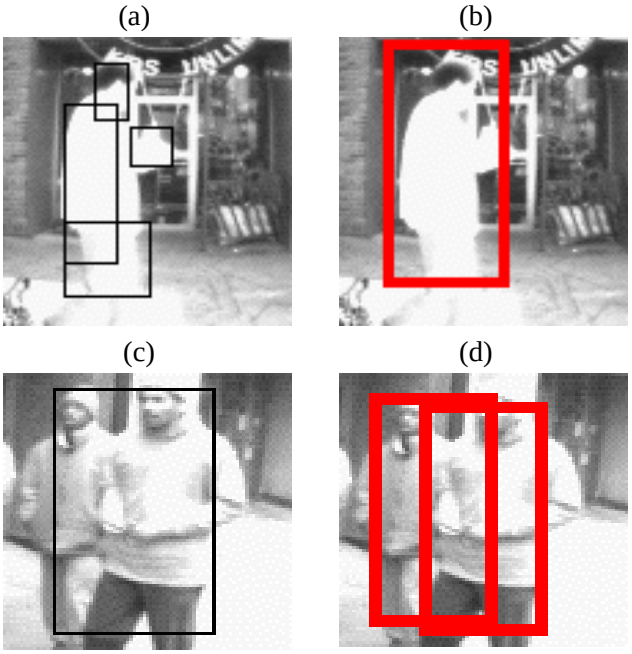

Figure 2: (a) The pedestrian is segmented into several parts as indicated by the black bounding boxes. (b) These parts are correctly grouped into a single one (indicated by the thick black box) through the hypothesis and verification procedure. (c) Two pedestrians are merged into a single one. (d) The pedestrians are correctly separated and identified.

through size thresholding. Fig. 1(c) illustrates the segmentation result on Fig. 1(a). From this result we can see that the overlapped objects are successfully separated due to their different distances from the cameras.

The segmentation step provides a rough estimate of the object position and size. Subsequently, a 2D box (shown in Fig. 1(d)) is fitted to each segmented region as an indication of an object. Then, all detected objects are fed into the trained neural network for pedestrian identification, which is presented in Section 4.

3.3 Postprocessing of the Object Detection Results

The major problem with the above object detection procedure is that each segmented region does not necessarily correspond to a single object. Due to noise, insufficient texture, and the limited pixel and range resolutions of a disparity map, a single object would be divided into multiple parts (e.g., Fig. 2(a)), while objects close to each other may be merged into a single one (e.g., Fig. 2(c)). [5] depends upon simple human size and motion cues to determine the correspondence between regions and pedestrians. In this system, we rely on spatial and shape information to achieve the same goal. Since we can derive the real size of an object from stereo analysis, our approach is much more reliable than that of [5].

This is a hypothesis and verification procedure. For small regions that are close to each other and have similar disparity values, they are temporally grouped into a single region. If the grouped region does not exceeds the size range of a normal person and can be classified as a pedestrian, then the grouping is confirmed, otherwise the small regions remain split. For a big region that exceeds the size range of a normal person, we use a window of a normal human size to search the whole region for pedestrians. If a subregion is identified as a pedestrian, it is separated from the original region; if no pedestrian is detected, the big region remains unchanged. Note that the verification is postponed after the pedestrian detection procedure. Fig. (b) and (d) show that the above procedure can correctly group regions belonging to a single object and split a region containing multiple objects.

4 Neural Network Based Pedestrian Recognition

In this paper a neural network approach is introduced which can be trained for different kind of scenes and can deal with noisy data robustly. In the following subsections we will describe the neural network input processing, training, and classification in details.

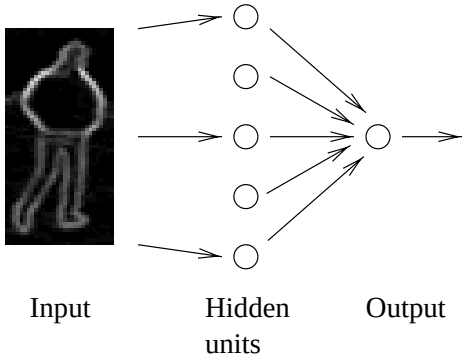

Figure 3: the architecture of the neural network

4.1 Preprocessing of the Input data

The design of the input data to the neural network is important; it directly affects the performance of the network. Our goal is to make the input data maintain the shape information for recognition while be reduced to a manageable amount.

An intensity image is often used as input in many neural network-based systems such as face detection [14] and autonomous vehicle steering [11]. However, image intensity is not appropriate to encode the consistent shape information of pedestrian since pedestrians present a much higher degree of variability in color and texture than human faces and road surfaces. The silhouette extracted from the segmented region is invariant to color and texture changes. However, range segmentation does not always provide useful silhouettes when severe noise or occlusion exists. An alternative choice is edge image, but the edge detection re-

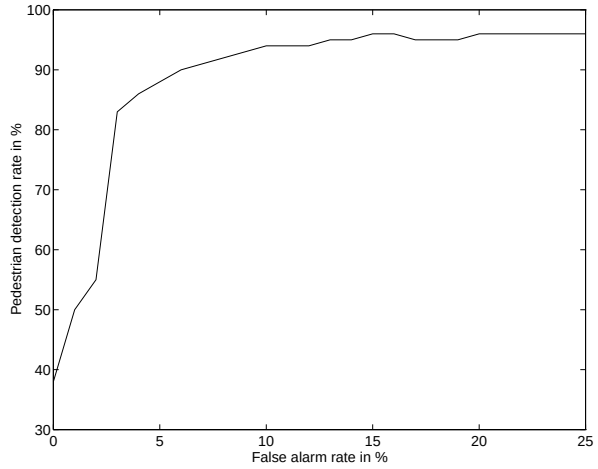

Figure 4: the ROC curve as a function of the threshold of the neural network output

sult is greatly influenced by the heuristic thresholding. In this system, we employ the intensity gradient image (as shown in Fig. 3) to encode the shape information of an object, because it has the same advantage as an edge image while it avoids the thresholding problem.

Each segmented region is normalized to a fixed 30x65 window. The gradient values ranging from 0 to 255 are linearly scaled to range from 0 to 1 so that the network inputs would have values in the same interval as the hidden unit and output unit activations.

4.2 The Neural network for Pedestrian Recognition

As shown in Fig. 3, we use a three-layer feed forward network for pedestrian recognition. There are 30x65 units in the input layer, five units in the hidden layer, and one unit in the output layer. Each layer is fully connected to the next, and each unit uses a sigmoid function for activation.

The network is trained by the back propagation algorithm. The initial training examples are generated from a set of manually labeled example images produced by the object detection module. Totally, we have 5318 training data — 1012 of pedestrians and 4306 of non-pedestrians ranging from traffic sign poles, parking meters, fire hydrants, vehicles, trees, to non-objects. Then, we use the "bootstrapping" strategy [15] to improve the system performance. For each training step one gradient image is chosen at random from the training set. The network parameters are initialized by small random numbers between 0.0 and 1.0, and are adapted during the training process. Therefore, the shape features to be extracted are learned from the training examples instead of being imposed *a priori*.

The network is trained to produce an output of 0.9 if a pedestrian presents, and 0.1 otherwise. Thus, we classify the detected object by thresholding the output value of the trained network: if the output is larger than the threshold, then the input object is classified as a pedestrian, otherwise as a non-pedestrian. The threshold determines the tradeoff between the rate of pedestrian detection and the rate of false alarm; we select this threshold by evaluating the *receiver operating characteristics* (ROC) curve shown in Fig. 4.

5 Experimental Results

The system has been implemented on a Pentium II 450 Mhz system under Microsoft Windows NT with an Imagination PXC200 digitizer. It has been tested extensively on large amounts of live video in urban areas obtained from a pair of cameras mounted on the top of a minivan. Over 8400 instances of pedestrians and other objects have been presented. By adjusting the threshold of the neural network output, we achieve a pedestrian detection rate of 85.2% and a false alarm rate of 3.1%. The system can detect and classify objects over a 320x240 pixel image pair at a frame rate ranging from 3 frames/second to 12 frames/second, depending on the number of objects presented in the field of view of the cameras.

In Fig. 5, we show the results of our pedestrian detection system on some typical urban street scenes under different weather conditions. Fig. 6 illustrates the results of processing the video of a walking pedestrian. Fig. 5 and Fig. 6 show that our system can detect pedestrians in different size, pose, gait, clothing, and occlusion status. However, there are some cases where the system fails. Most of the false alarms are objects with outlines similar to those of human beings. Other failures occur when a pedestrian is almost similar in color to the background, or two pedestrians are too close to each other to be separable by stereo and shape cues. We believe that including motion cues to the current system will improve the system performance.

6 Summary and Conclusions

This system is part of the Bus Driver Assistance project that aims at developing the next generation side collision warning system. To achieve this goal, foreground objects are first detected through foreground/background segmentation based on stereo vision. Each object is then classified as pedestrian or non-pedestrian by a trained neural network. The two key elements which make this system robust and realtime are stereo-guided object detection, and neural network-based pedestrian detection. The system effectively combines $2\frac{1}{2}$D information with the intensity information to detect pedestrians both walking and stationary. The neural network is trained on a large number of pedestrian and non-pedestrian data extracted from complex scenes, so it is applicable to various real world situations.

Acknowledgements

This research is sponsored in part by a contract titled "Development and Testing of Peformance Specifications for a Next Generation Side Collision Warning System". We gratefully acknowledge our partners, the USDOT Federal Transit Administration, PennDOT, and the Port Authority of Allegheny County.

References

[1] Y. Guo, G. Xu, S. Jsuji, "Understanding Human Motion Patterns," *Proc. of 12th Int. Conf. on Pattern Recognition*, Vol. 2, pp. 325-330, Jan. 1994.

[2] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, "Pfinder: Real-time Tracking of the Human Body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 780-785, July 1997.

[3] I. Haritaoglu, D. Harwood, L. Davis, "W^4—Real Time Detection and Tracking of People and their Parts," *Technical Report*, University of Maryland, Aug. 1997.

[4] H. Fujiyoshi, A. J. Lipton, "Real-time Human Motion Analysis by Image Skeletonization," *Workshop on Applications of Computer Vision*, 1998.

[5] O. Masoud, N. P. Papanikolopoulos, "Robust Pedestrian Tracking Using a Model-Based Approach," *IEEE Conf. on Intelligent Transportation Systems*, pp. 338-343, 1997.

[6] J. Segen, S. Pingali, "A Camera-Based System for Tracking People in Real Time," *Proc. of the 13th Int. Conf. on Pattern Recognition*, pp. 63-67, 1996.

[7] D. Hogg, "Model-based Vision: a Program to See a Walking Person," *Image and Vision computing*, Vol. 1, No. 1, pp. 5-20, 1983.

[8] K. Rohr, "Towards Model-based Recognition of Human Movements in Image Sequences," *CVGIP: Image Understanding*, Vol. 59, No. 1, pp. 94-115, Jan. 1994.

[9] P. J. Burt, J. R. Bergen, et al., "Object Tracking with a Moving Camera: An application of Dynamic Motion Analysis," *Proc. of IEEE Workshop on Visual Motion*, pp. 2-12, 1989.

[10] T. Darrell, G. Gordon, M. Harville, J. Woodfill, "Integrated Person Tracking Using Stereo, Color, and Pattern Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 601-608, 1998.

[11] D. A. Pomerleau, "Knowledge-Based Training of Artificial Neural Networks for Autonomous Robot Driving," *Robot Learning*, pp. 19-43, Boston, Kluwer Academic Publishers, 1993.

[12] M. Y. Siyal, M. Fathy, F. Dorry, "Neural-Vision Based Approach for Real-time road traffic applications," *Electric Letters*, Vol. 33, No 11, pp. 969-970, May 1997.

[13] C. Wohler, J. K. Aulanf, T. Portner, U. Franke, "A Time Delay Neural Network Algorithm for Real-time Pedestrian Recognition," *International Conference on Intelligent Vehicle*, Germany, 1998.

[14] H. A. Rowley, S. Baluja, T. Kanade, "Rotation Invariant Neural Network-Based Face Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 38-44, 1998.

[15] K. -K, Sung, T. Poggio, "Example-Based Learning for View-Based Human Face Detection," *A.I. Memo 1521*, AI Laboratory, MIT, Dec. 1994.

[16] C. Papageorgiou, T. Evgeniou, T. Poggio, "A Trainable Pedestrian Detection System," *1998 IEEE Int'l Conference on Intelligent Vehicles*, pp. 241-246, 1998.

[17] K. Konolige, "Small Vision Systems: Hardware and Implementation," *Proc. ISRR*, Hayama, 1997.

[18] C. Eveland, K. Konolige, R. C. Bolles, "Background Modeling for Segmentation of Video-Rate Stereo Sequences," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 266-271, 1998.

[19] T. Kanade, A. Yoshida, K. Oda, H. Hano, and M. Tanaka, "A Stereo Machine for Video-Rate Dense Depth Mapping and Its New Applications," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 196-202, 1996.

[20] J. Woodfill, B. Von Herzen, "Real-Time Stereo Vision on the the PARTS reconfigurable computer," *IEEE Workshop on FPGAs for Custom Computing Machines*, pp. 242-250, 1997.

[21] L. Matthies, A. Kelly, T. Litwin, "Obstacle Detection for Unmanned Ground Vehicles: a Program Report," *Proc. ISRR*, 1995.

[22] I. Haritauglu, D. Harwood, L. S. Davis, "W^4S: A Real-Time System for Detecting and Tracking People in $2\frac{1}{2}$D," *European Conference on Computer Vision*, pp. 877-892, 1998.

[23] A. J. Lipton, H. Fujiyoshi, R. S. Patil, "Moving Target Classification and Tracking from Real-Time Video," *Workshop on Applications of Computer Vision*, Princeton, NJ, Oct. 1998.

[24] H. Mori, N. M. Charkari, T. Matsushita, "On-Line Vehicle and Pedestrian Detection Based on Sign Pattern," *IEEE Trans. on Industrial Electronics*, Vol. 41, No. 4, pp. 384-391, Aug. 1994.

[25] S. A. Niyogi, E. H. Adelson, "Analyzing and Recognizing Walking Figures in *xyt*," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 469-474, 1994.

[26] S. A. Niyogi, E. H. Adelson, "Analyzing Gait with Spatiotemporal Surfaces," *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 64-69, Austin, 1994.

Figure 5: Results of pedestrian detection on typical urban street scenes

Figure 6: Pedestrian detection results of the video of a walking pedestrian