Chapter 16
Integrated Pedestrian Detection and Localization Using Stereo Cameras

Yu Wang and Jien Kato

Abstract Detecting and localizing other traffic participants, especially pedestrians, from a moving vehicle have many applications in smart vehicles. In this work, we address these tasks by utilizing image sensors, namely stereo cameras mounted on a vehicle. Our proposed method integrates appearance-based pedestrian detection and sparse depth estimation. To benefit from depth estimation, we map the prior distribution of a human's actual height onto the image to update the detection result. Simultaneously, the depth information that contributed to correct pedestrians' hypotheses is used for a better localization. The difference with other previous works is that we take the trade-off between accuracy and computational cost in the first place of consideration and try to make the most efficient integration for onboard applications.

Keywords Histogram of Oriented Gradients (HOG) • INRIA data • Pedestrian detection • Stereo cameras

16.1 Introduction

Pedestrian detection is a very fundamental component in many applications, such as smart vehicles and robot navigation. In this chapter, we address this task by using image sensor which has obvious advantages with regard to visibility and low setup cost. In utilizing an image sensor, the common method of finding pedestrians is to slide a window over all the scales and positions of the image, extract features from each window to match with a pretrained model, and return a set of detections with high-matching scores. Obviously, more distinctive features and more representative

Y. Wang (✉) • J. Kato
Nagoya University, Nagoya 464-8603, Japan
e-mail: ywang@mv.ss.is.nagoya-u.ac.jp; jien@is.nagoya-u.ac.jp

J.H.L. Hansen et al. (eds.), *Digital Signal Processing for In-Vehicle Systems and Safety*,
DOI 10.1007/978-1-4419-9607-7_16,

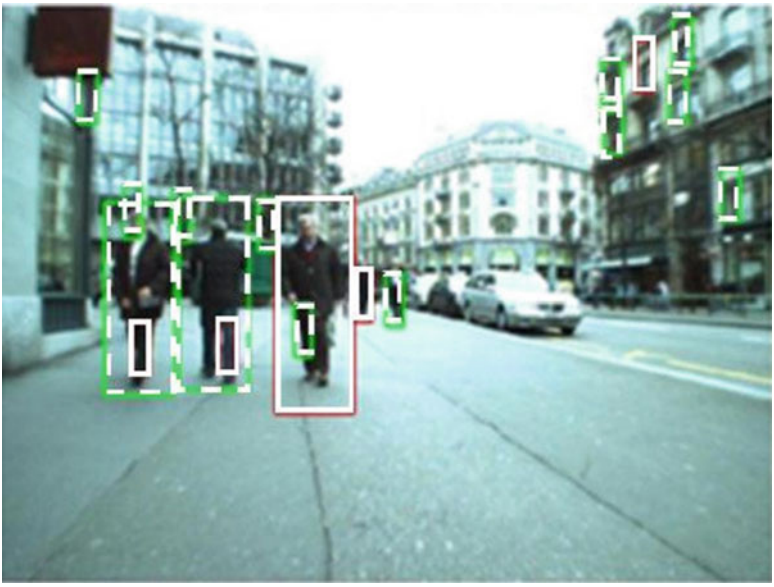

Fig. 16.1 Select candidates strictly (*continuous line bounding boxes*); use looser criterion, more candidates were found (*dashed line bounding boxes*)

models will lead to better accuracy. However, improvement in this approach sometimes comes with additional processing time which usually slows down the entire system's speed [1].

In most real-world applications, speed and accuracy are crucial issues and should be addressed simultaneously. Of course, time-consuming methods are not recommended. At the same time, the simplest and fastest methods are not robust enough by themselves. An example is illustrated in Fig. 16.1. We apply a very simple pedestrian detector described in [2] on the street view image. When selecting the candidates using strict standards, as shown with continuous line bounding boxes, many true occurrences for pedestrians were missed. As we make the selection standard a little looser, some missed true occurrences were successfully found. But the second approach has a drawback. The false number increased. This means that a detector using a simple feature and coarse model is not, by itself, discriminative enough. The inadequacy, however, could be compensated to some extent by using other cues from the image and background knowledge.

Several studies have tried to use other cues for pedestrian detection. Leibe et al. [3] proposed the use of scene geometry to improve object detection. By assuming that pedestrians can only be possibly supported by the ground plane, some false detection results could be filtered out. In another work, Gavrila and Munder [4] presented a system which involves a cascade of modules wherein each unit utilizes complementary visual criteria to narrow down the image searching space. These two were both excellent works; however, additional cues are mainly used to get rid of false results but unable to support a true one.

In a more recent publication, Hoiem et al. [5] showed how to take advantage of scene elements to jointly determine the camera's viewpoint, object identities, and surface geometry efficiently from a single image. By utilizing the probabilistic relationship among the scene elements, their integration makes a simple detector become much more discriminative. However, since the geometric estimation module costs too much time, their method has limited usage.

In this chapter, we build upon these ideas and expand them by integrating a simple appearance-based object detector with sparse depth estimation. By properly modeling the interdependence between object hypotheses and their location, our method could not only reject object hypotheses with unreasonable depth but also let sensible depth information to support a true one. In addition, the way we use depth is independent of prior assumptions and could be done quite fast.

16.2 Overall Strategy

Taking stereo images as input, our system mainly has two complementary modules which are able to run in parallel. The first one is a pedestrian detector which processes images from the left camera to find pedestrian hypotheses with image features only. For every single pedestrian hypothesis in the image, the detector will assign a bounding box around it and a detection score to indicate its confidence. The second module is sparse depth estimation which utilizes the stereo images together to estimate a sparse depth map of images from the left camera.

In order to integrate the two modules together, we use a probabilistic way. We assume that an object's imaged height is conditioned on the object category and its distance with respect to the camera. But the object identity and their distance are independent from each other. Using a graphical model, we can represent the conditional independence over the object identities o_i, their imaged height h_i, and the corresponding 3D distance d_i, as shown in Fig. 16.2. The I denotes the left camera image, and D means sparse depth map which could be estimated using the stereo image pair, both are observed evidences in our model. Typically, we have n object hypotheses in an image, where n varies by image.

With this model, the overall joint probability of the scene elements could be written in the following equation as

$$P(o,d,h,I,D) = \prod_i P(o_i)P(d_i)P(D|d_i)P(I|o_i)P(h_i|o_i,d_i) \tag{16.1}$$

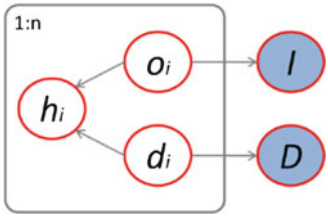

Fig. 16.2 Graphical model

With observed evidences I and D, we can use Bayes rule to give the likelihood of the scene elements conditioned on the evidences as

$$P(o, d, h|I, D) \propto \prod_i P(h_i|o_i, d_i)P(o_i|I)P(d_i|D) \tag{16.2}$$

The proportionality equation is with respect to I and$D$ which is constant evidence from stereo images. On the right-hand side, $P(o_i|I)$ means the confidence of an object hypothesis given image evidence, which could be estimated by our pedestrian detector. $P(h_i|o_i, d_i)$ indicates the probability of a hypothesis observed with imaged height h_i, conditioned on its category and 3D depth. In our case, it could be estimated by introducing a prior distribution of the pedestrians' actual height. That $P(d_i|D)$ is the confidence of depth estimation given the depth evidence from a depth map.

In this work, we estimate depth in an explicit way wherein the depth for each object hypothesis is exact and without any probabilistic description. This allows us to margin out the d on both left- and right-hand sides; for a single object hypothesis, we then get

$$P(o_i, h_i|I, D) \propto P(h_i|o_i, d_i)P(o_i|I) \tag{16.3}$$

where $P(o_i, h_i|I, D)$means, given the image evidences I and D, the probability of an object hypothesis o_i with its imaged height h_i. It is propagated with the $P(h_i|o_i, d_i)$ and $P(o_i|I)$, and could be considered as an improved confidence estimation of object hypothesis which not only takes into account the image evidence but also the depth information. We get the improved detection result by sorting the score of $P(o_i, h_i|I, D)$ for each object hypothesis and selecting the high ones. In the following paragraph, we will introduce the way we get $P(h_i|o_i, d_i)$and $P(o_i|I)$ from stereo images.

16.3 Pedestrian Detection

In order to obtain a set of pedestrian hypotheses, we built a baseline detector similar to the one described in [6]. As classifier, the Histogram of Oriented Gradients (HOG) feature and linear support vector machine was used. To distinguish this from the original 36-dimensional HOG feature used in [6], we employed an alternative 31-dimensional implementation from [1] to replace it. Also, to simplify the training process and speed up the runtime performance, a lower-dimensional feature set which could make a classifier with less parameters was utilized.

While training our detector, we used an existing package SVMPerf [7], which is highly optimized for training binary two-class classification SVMs with large data set. For this study, the INRIA person data set which has been organized into 3,610 positive samples of pedestrian with the size 70 by 134 was utilized. The negative

samples contain a fixed number of 15,000 patches that randomly selected from 1,239 person-free images of that data set. The training returns a 3,255-dimensional linear classifier (the size of 70 by 134 patch image's feature vector).

When a novel image emerges, we slide a window over the scales and positions to find the hypotheses. For each subwindow, we evaluate a score by doing dot product of the pretrained linear model and feature vector of the image patch. If the score is larger than the threshold, we either take it as a hypothesis or discard it. Typically, for an image portion that is likely to be a pedestrian instance, the score for the boxes around it will be very high. In order to eliminate any overlapped bounding boxes for the same instance, we perform non-maxima suppression to select only one box for each instance.

In this way, we get a set of hypotheses which is expected to have a pedestrian instance, each one with a bounding box and a classification score. However, the classification score is within the interval $(-\infty, +\infty)$. Since our graphical model wants a probabilistic input $P(o_i|I)$ which should be in the interval (0, 1), we therefore transform the SVM output into a probability form with logistic regression:

$$P = \frac{1}{1 + e^{Ax+B}} \tag{16.4}$$

where x is the classification score output from the dot product, P is the corresponding probability form of the score, and A and B are parameters which could be estimated by collecting a set of x and p. With novel classification score x′, we take the corresponding p′ as $P(o_i|I)$.

16.4 Localization of Pedestrian Instance

The use of a descriptor-based matching approach to obtain a sparse depth map distinguishes our work from the previous studies on how to estimate depth in a dense way. Though it could only provide a sparse representation of the scene, it is less ambiguous than dense matching which suffers from occlusion and nontexture regions. To make the depth map not "too sparse," we use two different kinds of key points as in [8] to relate the stereo images (Fig. 16.3).

We extract scale-invariant key points using Difference-of-Gaussian operator [10] and corner key point with Harris operator. For the scale-invariant key points, we utilize a GPU implementation of SIFT to compute their descriptors and match them by measuring the Euclidean distance. This implementation benefits from the Nvidia's CUDA technology and can get a speed of 25 Hz when processing images with size 640 by 480, which we think is enough for general real-world applications. The corner points are matched with a correlation window by normalized cross-correlation. Using two kinds of key points could help establish sufficient raw correspondences fast.

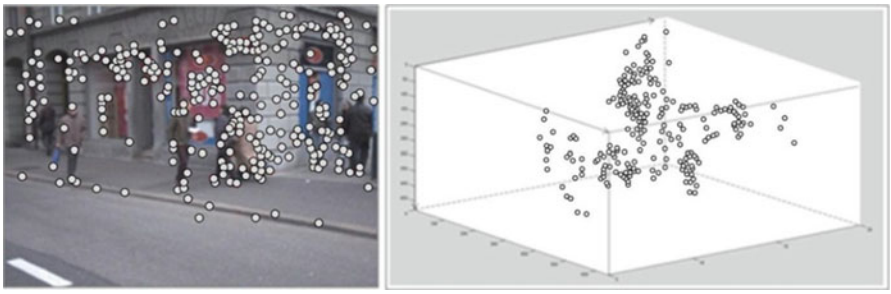

Fig. 16.3 Key points (*left*) and their 3D coordinates

With the raw matching result, we further refine them by enforcing Epipolar constraint and perform linear triangulation to get their 3D coordinates through precalibrated camera matrices. We set the left camera's optical center as the world origin, and then the z coordinate is the depth of each matched key point.

For each object hypothesis that we obtained, we collect all the matched key points inside its bounding box and select one representative for that bounding box's depth. Here we use a simple way to select the representative point by finding the nearest feature point around the diagonals' intersection and take the depth as the hypothesis' depth d_i.

Despite its simplicity, this solution performs reasonably well compared with other approaches such as using mean-shift to directly find the coordinates of the mass center. The reason may be that a lot of matched point is found around the object's boundary, and the mean-shift stops at local maxima frequently.

16.5 Utilize a Prior Height Distribution

The probability for the imaged height of a pedestrian hypothesis $P(h_i|o_i, d_i)$ is obtained by a product of the observed height of its bounding box h_i and a distance-conditioned height distribution $P(h_i|o_i, d_i)$. The later one is obtained using depth d_i. and a prior distribution of human's actual height.

Given a class-conditioned object hypothesiso_i, its distance d_i, and the camera's focal length f which we already know from the camera's calibration, we further model the height of an adult human using a simple Gaussian. The parameters of this Gaussian could be estimated from statistical data. We follow [5] to use a mean of 1.7 m and a standard derivation of 0.085 m for the pedestrian height distribution; therefore, we have the height distribution as $H \sim N\left(1.7, 0.085^2\right)$.

Given the prior distribution of pedestrian's actual height H, by using similarity relation, we can represent the imaged pedestrian's height as $h = Hf|D$. Because of $H \sim N\left(1.7, 0.085^2\right)$, h is also a simple Gaussian with $1.7f/d_i$ as mean and $0.085f/d_i$ as standard derivation. Therefore, we get

$$P(h|o_i, d_i) \sim N\left(1.7\frac{f}{d_i}, \left(0.085\frac{f}{d_i}\right)^2\right) \tag{16.5}$$

With this imaged height distribution and the observed height h_i of each bounding box, the confidence of every single hypothesis could be updated by taking the product of the detector output $P(o_i|I)$ and the $P(h|o_i, d_i)$. The updated confidence obtained in this way has thus taken into account the depth information and is expected to be more discriminative than the visual-features-only estimated result.

16.6 Experimental Results

We now present the experiment to show the performance of our method. The test data we used is collected from the ETHZ pedestrians' data set [9], which contains 5,235 pairs of stereo images that have been taken from either moving vehicles or mobile robots. All these images are from precalibrated cameras, with pedestrians on the left camera images annotated with bounding boxes as ground truth. The data were taken as sequences, so there are some continuous frames with almost the same scene. Since our work is only trying to evaluate the detection performance of single frame, we rearrange the data set by picking out image pairs with different scene structures. The final test set contains 133 pairs of stereo images with 798 annotations as ground truth.

In our experiment, we test three detection systems. First is our baseline detector, which uses HOG feature and linear support vector machine. The second is our proposed system which integrates this baseline detector and sparse depth estimation. The third one is UoCTTI detector [1], which employs mixtures of multiscale deformable part models. This is one of the best detectors in the PASCAL object detection challenge.

Some example detection results of the three systems on difficult images from our 133 stereo pairs' data set are shown in Fig. 16.4. The three columns from left to right show the output from our baseline detection system, proposed integration system, and UoCTTI system, respectively, on the same image. For a fair comparison, only the detections within the top ten confidences of each system are treated as output.

In general, the UoCTTI detector performed the best, as a result of more advanced modeling. Besides robust low-level feature, this detector uses a hierarchical structure model called deformable part model to represent the object category. In general, their detector finds pedestrians not only because they look like a person but also because they have parts (such as head, hands, legs), and these parts have appropriate positions. This makes the detector especially robust against occlusion. When distinguishing different human parts in crowded scene and large pedestrian volume conditions, the UoCTTI performs much better than our baseline detector system and our integration.

Fig. 16.4 Experimental results: (*left*) baseline system, (*middle*) integration system, (*right*) UoCTTI system

When compared to the raw output of our baseline detector, our integration system did quite well and shows significant improvement in the different scenarios. The reason is that we integrated the depth cue. Through it, the system could find pedestrians better by taking into account the observed height of detections and update the detection confidence to become more reasonable.

From the experiment, in some board scene images, as shown in the second row of Fig. 16.4, our integration system could perform better than the UoCTTI detector. We think the reason is the trade-off between different sources of information. While the UoCTTI detector utilizes both a deformable part model and the position of body parts to improve the detection, at the same time, there are drawbacks to this approach. Because the final detection result is partially based on the parts and the corresponding locations, in cases of low image resolution (parts are not visually clear) or the pedestrian instance is small (parts are not distinguishable), their model will penalize the detection and result in a low detection score. In contrast, our integration system uses depth information which is not dependent on any kind of condition (as long as the depth is accurately estimated). For the pedestrian instances that are small in the image, depth will help more because depth information itself does not depend on the resolution of the image.

Our quantitative experiment uses precision–recall (PR) curve to measure how a detection system performs in practice. It says a big deal about how the objects are

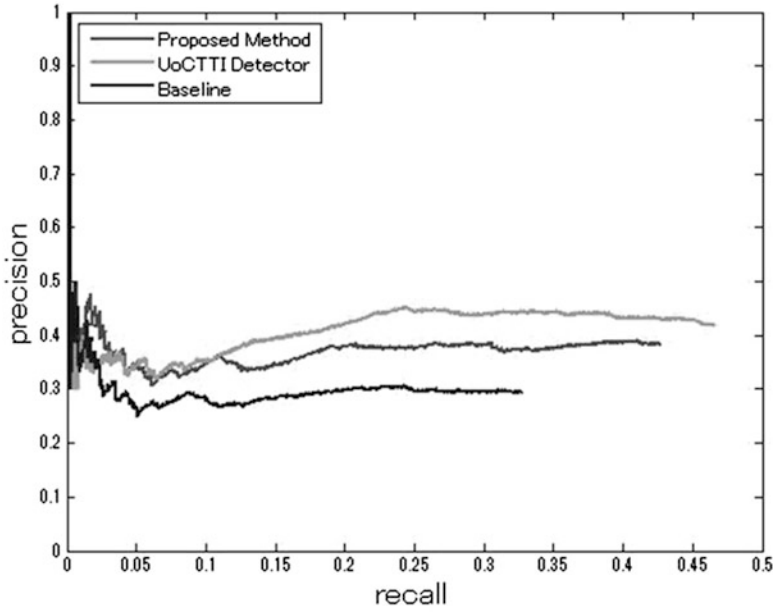

Fig. 16.5 PR curve for the detection performance

detected in practice. For a fair comparison, we also take top ten ranked hypotheses as system output. The comparison of the three systems' performance on the 133 stereo pairs is plotted in Fig. 16.5.

In most cases, the detector with deformable-part-based model has maintained a precision near 0.5. By integrating depth information, our proposed system outperforms the baseline detector significantly and closes to the best one.

We also compute an average precision for the three methods to show the overall performance. The results are 0.2325 (our method), 0.1738 (baseline), and 0.2530 (UoCTTI), respectively.

Without any optimization of speed, on a 2.83-G Intel Core 2 Quad CPU with 4 G RAM, the average speed of the three methods are 1.73 s (our method), 1.7 s (baseline), and 8.4 s (UoCTTI) on a single 640 by 480 image. The UoCTTI detector is quite time-consuming. It uses nearly five times more than our baseline detector. Since the UoCTTI detector also uses HOG as low-level feature set, its disadvantage in runtime may mainly boil down to the complicated model it uses. Therefore, even if it is powerful, it could not be used in some applications before the runtime issue could be resolved.

By carefully selecting the efficient cues, our integration system could also be very fast. Though this runtime performance is not good enough for some applications, it still has room for improvement. Currently, in our system, the most time-consuming part is the HOG feature pyramid computation and sliding window searching. Since these two kinds of processing can be done much more faster by using GPU programming, our integration system still have the potential to be used in real-time applications.

16.7 Conclusion

In this chapter, we proposed a method for pedestrian detection in traffic areas. We integrate typical object detection method with sparse depth estimation. This enables us to use 3D depth information naturally and improve detection accuracy by taking into account the human knowledge that "things become smaller when they move farther."

The efficiency of our integration was shown in our experiment. Without adding too much processing time, our method could improve the performance of our baseline detection system to a significant level, even close to a state-of-the-art detection system [1]. For the latter one, processing time for the detection with the same image size will cost nearly five times. Besides efficiency, another thing that we found out from the experiment is that the utilization of depth is independent of image resolution and instance size. This leads to stable improvement over the baseline system for all different kinds of scenes.

However, some issues still exist in the current system. First, the depth information that we introduced is obtained in an explicit way. This will, in some level, make the system sensitive against error in depth estimation. Secondly, our system is not good in handling occlusion and therefore quite weak in some crowd scenes. In the future work, we will mainly focus on robust depth estimation and occlusion handling.

References

1. Felzenszwalb P, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part based models. IEEE Trans Pattern Anal Mach Intell 32 (9):1627–1645
2. Torralba A, Murphy KP, Freeman WT (2004) Sharing features: efficient boosting procedures for multiclass object detection. In: IEEE Conference on computer vision and pattern recognition, Washington, 2004
3. Leibe B, Schindler K, Cornelis N, Van Gool L (2008) Coupled object detection and tracking from static cameras and moving vehicles. IEEE Trans Pattern Anal Mach Intell 30(10):1683–1698
4. Gavrila DM, Munder S (2007) Multi-cue pedestrian detection and tracking from a moving vehicle. Int J Comput Vision 73(1):41–59
5. Hoiem D, Efros AA, Hebert M (2008) Putting objects in perspective. Int J Comput Vision 80 (1):3–15
6. Dalal N, Bill Triggs B (2005) Histograms of oriented gradients for human detection In:. IEEE conference on computer vision and pattern recognition, San Diego, 2005
7. Joachims T (1999) Making large-scale support vector machine learning practical, MIT Press Cambridge, MA, USA
8. Wang Y, Kato J (2010) Reference view generating of traffic intersection. ICIC Expr Lett 4(4):1083–1088
9. Ess A, Leibe B, Schindler K, Van Gool L (2009) Robust multi-person tracking from a mobile platform. IEEE Trans Pattern Anal Mach Intell 31(10):1831–1846
10. Lowe DG (2008) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110