Integrate Sparse Depth Information into Pedestrians Detection

Yu Wang, Jien Kato and Kenichiro Ishii
Nagoya University, Japan
ywang@nagoya-u.jp, {jien, kishii}@is.nagoya-u.ac.jp

Abstract

In this paper, we propose to integrate sparse 3D depth information into pedestrian detection task, in order to achieve a fast boost in performance. Our proposed method uses a probabilistic way to integrate image-feature-based detection and sparse depth estimation together. The depth information is used as a cue, and provides additional discriminative ability for the detection. There are two contributions in this paper: 1) a simplified graphical model which could efficiently integrate depth cue into detection; and 2) a sparse depth estimation method which could provide fast and reliable estimation of depth information. The experiment shows that our method could provide promising enhancement over baseline detector with minimal additional time.

1 Introduction

Pedestrian detection is a very fundamental component in many applications, such as smart vehicle and robot navigation. Typical method towards this task is to slide a window over all the scales and positions of the image, extract image features from each detection window, and apply a pre-trained classifier to do the pedestrian/non-pedestrian classification. For this kind of method, image features are very important for the performance. A robust feature set is the key to discriminate pedestrians from background and other objects. Recent researches suggest that gradient-based features, such as Histogram of Gradient and edgelets, work very well in human detection, because they have strong ability in catching the silhouette information.

However, for many real world scenes where complex background and occlusion exist, gradient-based features also encounter difficulties in perceiving sufficient robustness. To deal with such kind of scenes, researchers tried different ways, such as using multiple image features and adding expressive object models, to build a more discriminative detector. R. Schwartz et al. [4] proposed to combine different types of local features such as color, gradient, and textures together. They extract an 170,820 dimensional feature vector from each detection window and use it for detection. In another work, P. Felzenszwalb et al. [3] introduced a deformable part model for detection. In their work, they classify an instance as human not only because it looks like a human, but also because it has parts (such as head and feet), and these parts are in appropriate positions. They showed that an informative model could also help to improve the detection accuracy effectively. These two works are both very excellent works. However, since both long feature vector and additional object models bring severe burdens on computation, it is not easy to adapt these methods in applications which require a fast processing speed.

Actually in many applications, the detection needs to be done not only accurately but also fast. So the method that used to improve detection performance should also preserve a fast processing speed. From this point of view, in this paper, we propose to use the 3D depth information as an additional cue to do the pedestrian detection task. In our method, the depth of each detection window is computed and used to map a prior distribution of human's actual height onto the image plane. The resulted imaged height distribution is then used to update the image-feature-based detection result for the corresponding detection window. The final detection result is contributed by both image features and depth information, and could provide stronger ability to discriminate pedestrians from other objects. There are mainly two contributions in our work: 1) a probabilistic model for the efficient use of depth information in detection; 2) a sparse depth estimation method for a fast and reliable estimation of depth information. We show that our method could provide over 33% enhancement in detection accuracy comparing to the baseline detector, with minor additional processing time.

2 Related Work

Depth information is valuable for human detection and has been explored in many previous works. Earlier works, such as [5] and [6], group the depth value of neighbouring pixels to generate the region of interest (ROI) in the image. Only the ROIs are expected to have pedestrians' existence and are further to be applied with a pedestrian detector. In these works, depth information was mainly used to do pre-processing to reduce the image searching space.

In [6], D. M. Gavrila et al. also implemented a way to use depth information to verify detector's output. They assumed pixels of a true detection should have similar depth values, and introduced a rejecting mechanism to get rid of detection windows which have large deviation of depth inside. This could help to filter out detections which contain an appreciable amount of background. However, because the depth was only used for post-verification on detector's output and could not contribute to the detection accuracy, such kind of usage was still limited and did not make the full use of depth information.

Recently, A. Ess et al. [2] presented a system which integrate dense depth estimation, visual odometry and pedestrian detection together for an on-board tracking purpose. In their system, the detector's output is integrated with depth information in a probabilistic way, which is similar with our proposed method. However, their approach is quite different with us. In A. Ess' work, the depth information is estimated for every single pixel by doing dense matching (find pixel wise correspondence) between stereo images. Though the resulted dense depth map is very informative, the dense matching itself is time consuming and sensitive with some image conditions (such as image noise, textureless regions, and occlusions). In a simple word,

they sacrifice the robustness and speed to obtain more complete depth information of the scene. Contrast to them, we put the speed and robustness in the first place of consideration and use sparse matching to obtain the depth information of the scene. Though our method could only obtain the sparse depth information of the scene, it is fast, reliable, and sufficient for our purpose.

3 Approach

Take stereo images as input, we have two complementary modules which are able to run in parallel. One is a image-feature-based pedestrian detector, which applies on left camera image to generate a set of pedestrian hypotheses. The other is depth estimation, which applies on the stereo images together to estimate a sparse depth map of the scene. For every pedestrian hypothesis output from the detector, a distance will be computed from the depth map. The distance is further used to update the hypothesis' corresponding image-feature-based confidence. In this work, we use a graphical model to integrate these two modules together, and introduce a prior height distribution of adult human to enable the confidence updating.

We assume object's imaged height is conditioned on its category and the distance with respect to the camera, but the object identity and the distance are independent from each other. Using a graphical model, we can represent the conditional interdependence over n pedestrian identities o_i, their imaged height h_i, and the corresponding 3D distance d_i, as shown in Fig.1. The I denotes the left camera image and the D indicates the sparse depth map estimated from the stereo image pair, both are observed evidence in the model.

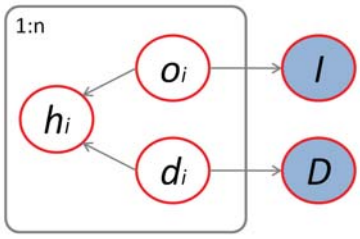


Figure 1. Graphical model.

With the model, the overall joint probability could be written in the following equation as

$$P(o,d,h,I,D) = \prod_i P(o_i)P(d_i)P(D \mid d_i)P(I \mid o_i)P(h_i \mid o_i d_i). \quad (1)$$

Using Bayes rule, we can give the likelihood of the properties of pedestrian hypotheses that conditioned on the image and depth evidences as

$$P(o,d,h \mid I,D) \propto \prod_i P(o_i \mid I)P(h_i \mid o_i d_i)P(d_i \mid D). \quad (2)$$

In this work, we determine the depth in an explicit way, where the depth for each pedestrian hypothesis is exact given the depth evidence. This allows us to margin out the d on both left and right hand side, for a single object hypothesis, we then get

$$P(o_i,h_i \mid I,D) \propto P(h_i \mid o_i d_i)P(o_i \mid I), \quad (3)$$

where in the left hand side, $P(o_i,h_i \mid I,D)$ indicates given the image evidence I and D, the probability of an pedestrian hypothesis o_i exits with its imaged height h_i. It is propagated with the $P(h_i \mid o_i d_i)$ and $P(o_i \mid I)$, and is a updated confidence estimation of pedestrian hypothesis which not only take into account the image evidence but also the depth information. We get updated confidence for every pedestrian candidates by propagating the $P(o_i,h_i \mid I,D)$ from $P(h_i \mid o_i d_i)$ and $P(o_i \mid I)$. The $P(o_i \mid I)$ and $P(h_i \mid o_i d_i)$ are estimated from the detection and depth estimation respectively, in the following paragraph, we will introduce the way we estimate them in detail.

4 Baseline Pedestrian Detector

In order to obtain a set of pedestrian hypothesis o_i and their corresponding confidence $P(o_i \mid I)$, we trained a baseline detector similar with the one described in [1]. We also use the Histogram of Oriented Gradients (HOG) as local feature and linear support vector machine as classifier. However, for the implementation efficiency, we replace the original 36-dimensional HOG feature with a novel proposed 31-dimensional one [3]. The training data we used is the INRIA person data set, from which we arranged 3,610 positive samples and 15,000 negative samples, both are of the size 70×134. The training returns a 3,255 dimensional linear classifier (the size of 70×134 patch image's feature vector).

While applying the trained detector to generate pedestrian hypothesis, the classification score output from the linear classifier is within the interval $(-\infty, +\infty)$. Since our graphical model wants a probabilistic input $p(o_i \mid I)$ which should in the interval $(0, 1)$, we therefore transform the SVM output into a probability form using a logistic function defined as

$$p = \frac{1}{1 + e^{Ax+B}}, \quad (4)$$

where x is the classification score, p is its probability form, 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 $p(o_i \mid I)$.

5 Utilize Depth Information

The probability for the imaged height of a pedestrian hypothesis $P(h_i \mid o_i d_i)$ is estimated by observing the height h_i of its bounding box in a distance-conditioned height distribution $p(h \mid o_i d_i)$. The later one is obtained with the distance and a prior distribution of pedestrian's actual height.

5.1 Sparse Depth Estimation

Different with many previous works uses dense matching, we adapt sparse matching to obtain the depth information of the scene. To make the depth map not "too sparse", we used a novel multiple operator key point matching approach to obtain the raw matching result of the stereo pairs.

From the stereo images, we extract scale invariant key points with Difference-of-Gaussian operator and corner key points with Harris operator. We compute a 128-dimensional SIFT descriptor for each scale invariant key point, and find its corresponding point in the other image by measuring the Euclidean distance. For the corner key points, we extract their surrounding 11×11 pixels and do the normalized cross-correlation matching. The matches of these two kinds of key points

are then fused to our raw matching result. Because these two kind of key points have quite different properties, using them together could help to establish sufficient raw correspondences that covers most significant portion of the image.

With the raw matching result, we further refine them by enforcing Epipolar constrain to remove correspondences that exist apart from their Epipolar line more than a threshold (ex. 2 pixels). This could further remove outlier matches and guarantee the quality of matching. Then we do linear triangulation to get the 3D coordinates of matched key points with precalibrated camera matrices.

For each object hypothesis o_i that we obtained with the detector described in the previous section, we collect all the matched key points inside its bounding box and select one that is representative for it. The distance d_i between the representative point and camera center is taken as the distance of the hypothesis. Here we use a simple way to select the representative point. We find the nearest k feature points $P_t(t = 1, ..., k)$ around the diagonals' intersection of the hypothesis' bounding box, and select the point P_i which has the minimum sum of distance in depth with other points. We think it is not a good solution and have tried to use mean-shift to directly find the coordinates of the 3D points' mass center. However, it did not perform well enough even comparing to our simplest solution. 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.

5.2 Map the Prior Height Distribution

With class conditioned object hypothesis o_i, its distance d_i and known camera's focal length f, we map a prior height distribution H of pedestrians to the imaged one $p(h \mid o_i d_i)$.

We specify that the height H of adult pedestrian is normally distributed with a mean of 1.7 meters and a standard deviation of 0.085 [8], therefore we have $H \sim N(1.7, 0.085^2)$. Using the similarity relation of the two triangles, we can represent the imaged pedestrian's height as $h = Hf/d$. Since $H \sim N(1.7, 0.085^2)$, h is also a simple Gaussian with $1.7f/d_i$ as mean and $0.085f/d_i$ as standard derivation. Therefore we get $p(h \mid o_i d_i) \sim N\left(1.7f/d_i, (0.085f/d_i)^2\right)$.

With this imaged height distribution and the observed height h_i of each bounding box in the image, confidence of every single hypothesis could be updated by propagating from $p(o_i \mid I)$ and $p(h_i \mid o_i d_i)$. 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.

6 Experimental Results

The purpose of the experiment is to see if using depth information in our proposed way could efficiently improve image-feature-based pedestrian detection in complex scenes. For this end, we prepared a very difficult dataset by selecting images from the ETHZ tracking sequence [2]. Our dataset contains 133 pairs of stereo images of complex street view scenes, with 798 annotations as ground truth. All of the experiments were done on an 2.83G Intel CPU with 4G RAM. The sparse depth estimation was partially supported by a NVIDIA GeForce 9800GT GPU with 512M VRAM.

In the experiment, we have three systems for comparison: the baseline detector, our proposed method (baseline + depth) and a UoCTTI detector (baseline + part model) proposed in [3]. The UoCTTI detector adapts a very expressive part model, and uses the part information as additional cue for detection. It is one of the best detectors in the PASCAL object detection challenge.

6.1 Quantitative Results

Because all the three systems work by scanning image windows and doing pedestrian/non-pedestrian classification in each window, we use the false positives per image (FPPI) as a metric to evaluate the combinational performance of detection and scanning. The plot is shown in Fig.2. We can see our method has made significant improvement over the baseline, and comparable with the UoCTTI detector.

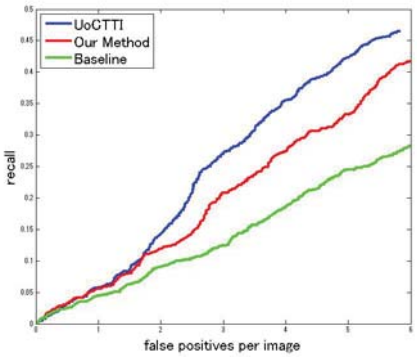


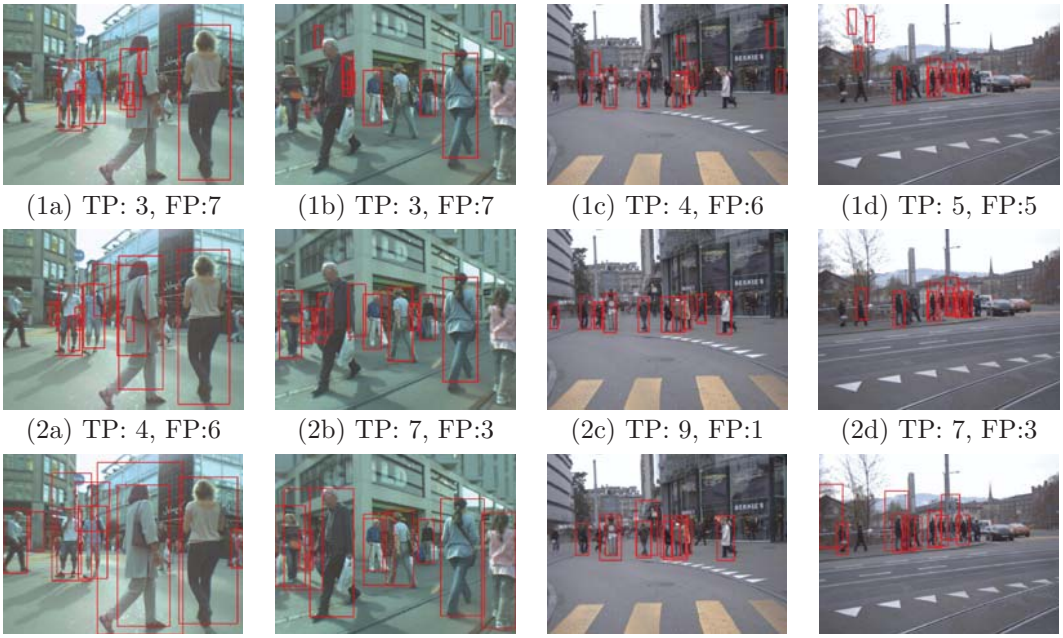
Figure 2. The FPPI plot of the three systems.

We also compute the interpolated Average Precision (AP) [7] to summarize the overall accuracy of each system. AP penalizes methods which achieve low total recall as well as those with consistently low precision, and is ideal for measuring the overall accuracy. The AP for the baseline detector, UoCTTI detector and our proposed methods are 0.1738, 0.2530 and 0.2325, respectively. Our depth added method is very close to the UoCTTI detector and brings near 33% improvement to the baseline detector.

The speed of the three systems is listed in Table.1. In the current implementation of our depth added method, the detection and depth estimation are done in serial. It cost 0.18 second per frame to update the detection result from the baseline detector, thus the overall speed is 1.88 seconds per frame. Since the detection and depth estimation are designed to be done in parallel, in that case, the speed will be 1.73 per frame. This overall runtime performance is not good enough and it still have space to improve. Because our depth added method is free with the choice of baseline detector, using other faster detector or implementing current detector with GPU programming could make the overall speed be able to meet more applications.

6.2 Discussion

In Fig.3, we display some example detection results outputted from the three systems. Comparing to the raw output of the baseline detector, both our depth added method and the UoCTTI detector have made



(1a) TP: 3, FP:7 (1b) TP: 3, FP:7 (1c) TP: 4, FP:6 (1d) TP: 5, FP:5

(2a) TP: 4, FP:6 (2b) TP: 7, FP:3 (2c) TP: 9, FP:1 (2d) TP: 7, FP:3

(3a) TP: 6, FP:4 (3b) TP: 8, FP:2 (3c) TP: 6, FP:4 (3d) TP: 7, FP:3

Figure 3. Some results of the three systems. The three rows correspond to the results from the baseline detector, our method and UoCTTI detector respectively. (TP: true positive, FP: false positive.)

Table 1. Detection speed of the three systems.

Baseline	1.7s	
UoCTTI	8.4s	
Our Method	baseline detector: 1.7s	1.88s
	depth estimation: 0.15s	
	integration: 0.03s	

significant improvement. It is to say, the depth cue and the part cue could provide additional discriminative ability to the image-feature-based detector. In some crowded scenes, the UoCTTI detector did even better than our method. This may mainly benefit from a very expressive part model it used. The detector uses part information as additional cue, which leads to the detector be able to preserve strong robustness even when heavy occlusion exist.

However, in some board scene images such like the last two columns of Fig.3, our depth added method sometimes could do even better than the UoCTTI detector. We think this is because the UoCTTI detector may face the trade-off between different sources of information. While it uses a deformable part model and utilize the position of parts to improve the detection, it may also suffer from that model. Because the final detection result is partially based on the parts and their corresponding locations, in case the parts are not visually clear enough, their model will penalize that detection and result in a low detection score. Contrast to it, our method does not have such kind of issues. It brings stable improvements over the baseline detector in different kind of scenes.

7 Conclusion

In this paper, we propose to integrate sparse 3D depth information in pedestrian detection. We introduce a novel descriptor based key point matching approach to obtain the sparse depth information of the scene, and a simplified graphical model to use the resulted depth information to update the image-feature-based classification result. We show in our experiment that, with minimal additional processing time, our method could improve the detection accuracy of the baseline detector significantly, and comparable to a start-of-the-art detection system [3]. While for the later one, processing a same size image will cost near five times of the time. Currently, our method does not have any occlusion handling mechanism, therefore is quite weak in some crowded scenes. In the future work, we will focus on extending current method to become more robust against occlusion

References

[1] N. Dalal and B. Triggs: "Histograms of Oriented Gradients for Human Detection" in CVPR 2005.

[2] A. Ess, B. Leibe, K. Schindler and L. V. Gool: "Robust Multi-Person Tracking from a Mobile Platform" TPAMI, Vol.31, No.10, 2009.

[3] P. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan:"Object Detection with Discriminatively Trained Part Based Models" TPAMI, Vol.32, No.9, 2010.

[4] W. R. Schwartz, A. Kembhavi, D. Harwood, L. S. Davis: "Human Detection Using Partial Least Squares Analysis" in ICCV 2009.

[5] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan and L. H. Matthies: "Results from a Real-time Stereo-based Pedestrian Detection System on a Moving Vehicle" in ICRA 2009.

[6] D. M. Gavrila and S. Munder: "Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle" IJCV, Vol.73, No.1, 2007.

[7] M. Everingham et al.: "The 2005 PASCAL Visual Object Classes Challenge" Selected Proceedings of the First PASCAL Challenges Workshop, Springer, 2006.

[8] D. Hoiem, A. A. Efros and M. Hebert: "Putting Objects in Perspective" IJCV, Vol.80, No.1, 2008.