IEICE TRANS. INF. & SYST., VOL.E94–D, NO.8 AUGUST 2011

PAPER

Pedestrian Detection with Sparse Depth Estimation

Yu WANG[†a)], *Nonmember and* **Jien KATO**[†], *Member*

SUMMARY In this paper, we deal with the pedestrian detection task in outdoor scenes. Because of the complexity of such scenes, generally used gradient-feature-based detectors do not work well on them. We propose to use sparse 3D depth information as an additional cue to do the detection task, in order to achieve a fast improvement in performance. Our proposed method uses a probabilistic model to integrate image-feature-based classification with sparse depth estimation. Benefiting from the depth estimates, we map the prior distribution of human's actual height onto the image, and update the image-feature-based classification result probabilistically. We have two contributions in this paper: 1) a simplified graphical model which can efficiently integrate depth cue in detection; and 2) a sparse depth estimation method which could provide fast and reliable estimation of depth information. An experiment shows that our method provides a promising enhancement over baseline detector within minimal additional time.

key words: *pedestrian detection, depth estimation, stereo matching*

1. Introduction

Pedestrian detection is a fundamental component in many applications, such as smart vehicle, robot navigation and various first person vision applications. Typical methods for this task 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 studies suggest that gradient-based features (such as Histogram of Gradient [1] and edgelets [2]) work very well in human detection, because they have strong ability in catching the silhouette information in image.

However, in many real world scenes where complex background and occlusion exist, such gradient-based image features encounter difficulties in achieving sufficient robustness. Take Fig. 1 for an example, we apply the pedestrian detector proposed by N. Dalal et al. [1] to find pedestrians in the street view image. The detector uses HOG (Histogram of Gradient) as feature and linear Support Vector Machine as classifier. When we use it to select the pedestrian candidates strictly, outputs shown in red, we found that many true pedestrian instances were not detected. As we make the selecting standard a little looser, more candidates could be found as shown by the green boxes. Inside these augmented candidates, we can see some missed true instances were successfully detected. However, the number of false detections was also increased simultaneously.

Fig. 1 Detect pedestrians using a typical HOG based detector. Select candidates strictly as shown in red, many true instances were missed; use looser criterion, more candidates were found in green, but the number of false positives also increased.

In order to meet the requirement of real world applications, researchers have tried different ways to build a more discriminative detector. W.R. Schwartz et al. [7] proposed to combine different types of local image features together to become a strong feature set. In their work, gradient-based feature is augmented with color, textures and their co-occurrence statistics. They compute a 170,820 dimensional feature vector in each detection window for classification, and showed that the high dimensional feature set could bring significant improvement to detection accuracy. For their method, speed is a remaining issue. Because such a high dimensional feature set brings severe burdens on computation, it is not easy to adapt them in applications which require a fast processing speed.

In another work, instead of developing a stronger feature set, P. Felzenszwalb et al. [5] proposed to use object's part information for detection. They introduced a deformable part model to represent object, and use it to do additional analysis in each detection window. In their work, a detection window is classified as human not only because it looks like a human (based on image feature), but also because it has parts (such as head and feet), and these parts are in the appropriate positions. They show that using part information could also effectively improve the detection accuracy. However, doing additional analysis with such an elegant object model also comes with high computational cost, thus make the system become relative slow. As a result, even though it can improve detection accuracy, such

Manuscript received December 10, 2010.
Manuscript revised April 6, 2011.

[†]The authors are with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.
a) E-mail: ywang@nagoya-u.jp

DOI: 10.1587/transinf.E94.D.1690

kind of object model is also hard to adopt in many applications.

Actually in many applications, such as those in smart vehicle and robot navigation, 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 beside 2D image features 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 pedestrian 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 Works

Depth information is valuable for human detection and has been explored in many previous works. Earlier works, such as [8] and [9], 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 preprocessing to reduce the image searching space.

In [9], 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. [4] 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 computationally complex (because it relies on global optimization) and sensitive with some image conditions (such as image noise, textureless regions, and occlusions). Contrast to them, we put the computational efficiency 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.

There also exist some other works that use range sensors to get the depth information and use in detection task. For example, S. Ikemura and H. Fujiyoshi [3] proposed to obtain the depth information estimated from a TOF (Time of Flight) camera, and use as a feature in human detection. The TOF camera measures the depth information by calculating the time it takes for the light reflect by object surface to arrive at the camera. It could provide quite accurate measurement of the depth. However, the TOF cameras are still limited to use in indoor environment and have a limited working range. As a result, such kind of method is not easy to be adapted in different applications, such as the outdoor pedestrian detection task we are dealing with in this work.

The purpose of our work is to use depth information to achieve better detection accuracy, while preserving a fast processing speed. To do this, it is necessary to appropriately deal with the following two problems: 1) how to efficiently integrate the depth information and detector's output together, and 2) how to fast and reliably estimate the depth information of the scene. For the first problem, we introduce a simplified graphical model to probabilistically integrate these two kinds of estimation together. Different with many prior works which use the depth for pre-processing or post-verification purpose, we use the depth as additional cue for discriminating pedestrians out of background and other objects. For the second problem, we introduce a sparse matching method to estimate the depth information of the scene. Our method uses key point descriptor to measure the similarity of points and select only confident matches for the depth computation. This makes the resulted depth information reliable for use.

3. Approach

Our basic idea is to use the 3D depth information as an additional cue to do the pedestrian detection. The detection results in our work are contributed by both image features and depth evidence, and could better discriminate pedestrians from background and other objects. In order to efficiently use the depth information, we apply a probabilistic way in this work.

3.1 Overall Strategy

Take stereo images as input, our method mainly consists of two complementary modules which are able to run in parallel. The first one is an image-feature-based pedestrian detector, which applies on one of the stereo images to generate a set of pedestrian hypotheses. For every pedestrian hypothesis in that image, the detector will assign a bounding box

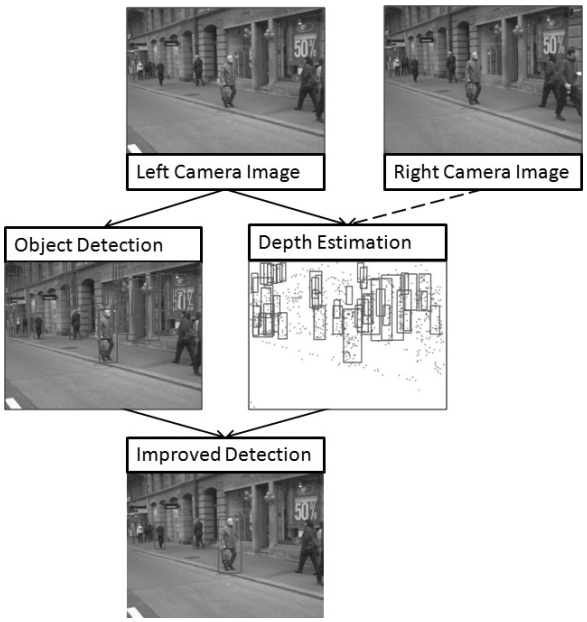

Fig. 2 System overview.

to indicate the location and a detection score to indicate its confidence. The second one is depth estimation, which applies on both stereo images to estimate a sparse depth map of the scene. The resulted sparse depth map contains a set of sparely distributed points corresponding to the key points in the stereo images, each attaches a 3D depth value. In our current system, the detection is performed on the left camera's image, and the system overview is shown in Fig. 2.

For every pedestrian hypothesis output from the detector, a distance will be computed by using the depth value from the depth map. The distance is further used to update the hypothesis' corresponding confidence. With human knowledge, such kind of updating is not difficult because we know how the size of a human should be and "things become smaller when get further". Take the pedestrian instance in the middle of Fig. 2 for example, we know what size it should be in the image given the corresponding distance. If the observed size is close to it, we will be more confident, and vice versa. Inspired by such human experience, we use a graphical model to probabilistically integrate the detection with depth estimation, and introduce a prior height distribution of adult human to enable the confidence updating.

3.2 Graphical Model

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 the pedestrian identities o_i, their imaged height h_i, and the corresponding 3D distance d_i, as shown in Fig. 3. The I denotes the left camera image and the D indicates the sparse depth map estimated from the stereo

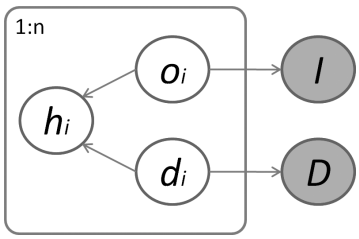

Fig. 3 Graphical model that represent the interdependence among hypothesis' properties, image and depth evidences.

image pair, both are observed evidence in the model. We have n pedestrian hypotheses for each stereo pairs.

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:

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

The proportionality equation in Eq. (2) is with respect to I and D which are constant given a stereo pair. In the right hand side, $P(o_i \mid I)$ means the confidence of a pedestrian hypothesis o_i given image evidence, which we estimate using a pedestrian detector. $P(h_i \mid o_i d_i)$ means the probability of a hypothesis observed with imaged height h_i, conditioned on its category and 3D distance. We estimated it by mapping a prior distribution of pedestrians' actual height to the image plane. The $P(d_i \mid D)$ is the confidence of the depth estimation given the evidence from depth map.

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 kind of solution could significantly simplify the computation while preserving the effectiveness of the use of depth. 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 resulted confidence are then be resorted and high ones are selected as the novel detection output. In the following paragraph, we will introduce the way we estimate the $P(o_i \mid I)$ and $P(h_i \mid o_i d_i)$ in detail.

4. Generate Pedestrian Hypotheses

Pedestrian hypotheses o_i are generated by applying a pedestrian detector on left camera images. Each hypothesis has a bounding box to indicate the location and a classification score x_i to indicate the confidence. In order to integrate this detection result with depth evidence, we convert each raw classification score x_i to its corresponding probabilistic form $P(o_i \mid I)$ with logistic regression.

4.1 Baseline Pedestrian Detector

The baseline pedestrian detector in our work is similar with the one proposed by N. Dalal et al. [1]. We also use the Histogram of Oriented Gradients (HOG) as the local image feature and linear support vector machine as the classifier. Different with the proposed method, we replace the original 36-dimensional feature set with a novel 31-dimensional one that described in [5]. The lower dimensional feature set could make the classifier with less parameter therefore simplify the training process and speed up the runtime performance.

The overall processing flow of the detector is outlined in Fig. 4. It could be coarsely divided into learning phase where a linear classifier function is being learned from labelled image samples, and detection phase where trained classifier function is to be applied on novel image to find object hypotheses.

Our baseline detector was trained on the INRIA person data set [1]. From the dataset, we arranged 3610 positive samples of adult pedestrian and 15000 person-free negative samples. All these samples are of the size 70×134. Using the 31-dimensional HOG proposed by P. Felzenszwalb et al. [5], we computed a 3255 dimensional feature vector x for each sample.

In the learning phase, we learn a linear model β from these training samples using SVM [11]. So that for the positive samples, their feature vector x have $\beta^T x > b$ and negative ones have $\beta^T x < b$. The training returns a 3255 dimensional linear classifier which has the same size with the samples' feature vector.

When novel image comes, we slide a window over its scales and positions, and use the linear model β to evaluate a classification score for each sub window. Because we use a linear model, the searching could be simplified. The classification score could be computed by doing a dot product of the pre-trained linear model and the feature vector of the image patch. Therefore, searching over a single scale of the image is equivalent to convolve the β with a single scale feature map of the image. Additionally, searching over scales of the image could be implemented by doing single scale searching over the image's pyramid.

In practice, we implement the sliding window searching by convolving the linear model β with the feature pyramid. By doing this, we obtain a classification score for each position with a scale. In general, for an image portion that is likely to be a pedestrian instance, the classification score for its surround bounding boxes will be all very high. We therefore perform non-maxima suppression to eliminate the overlapped bounding boxes by selecting only one box for each instance.

4.2 Confidence Converting

From the results of the detector described above, we filter out detections windows which have very low confidence, and leaves relative high scored hypotheses o_i which are expected to be pedestrian instances. However, their corresponding classification score x_i that output from the linear model is within the interval $(-\infty, +\infty)$, and 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 with logistic regression.

In our case, because the classification score is the only explanatory variable, we define the logistic function in the following form:

$$p = \frac{1}{1 + e^{Ax+B}}. \tag{4}$$

where, x is the classification score that output from our classifier, p is the corresponding probability form. A and B are the parameters, which are estimated by using a set of collected classification score and the correspond-

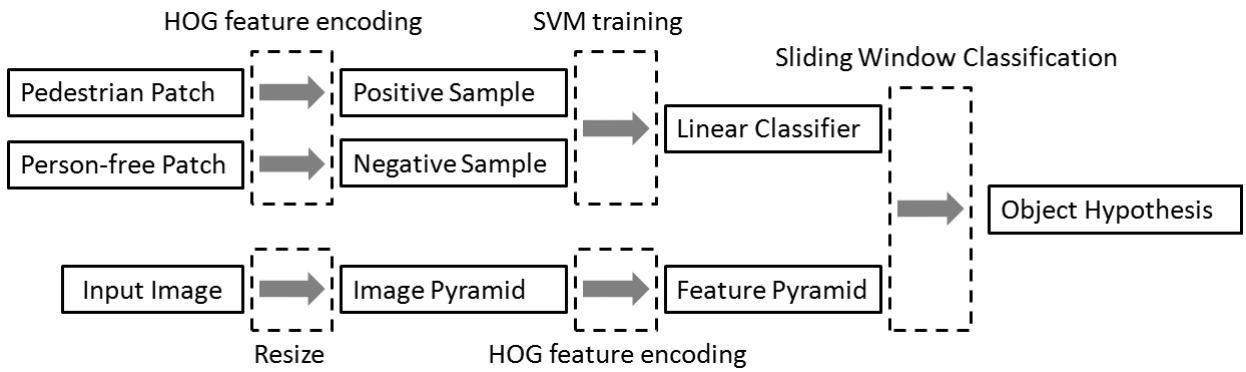

Fig. 4 Processing flow of training and applying the baseline pedestrian detector.

ing pedestrian/non-pedestrian label. The resulting function takes x within $(-\infty, +\infty)$ and outputs the corresponding p within $(0, 1)$.

In this way, for a given image, we generate a set of pedestrian hypotheses and its corresponding probability given the image evidence. Our method is free for the choice of pedestrian detector as long as they could output a score to indicate the confidence, thus is easy to be generalized to other detectors.

5. Utilizing Depth Evidence

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

In many prior works, depth is estimated by performing dense matching on stereo images. In general, dense matching methods use pixel intensity as similarity measurement, neighbouring smoothness as prior, and apply inference method (such as belief propagation or graph cuts) to find the pixel wise correspondence between stereo images. Such kind of methods could provide dense depth map which is quite informative, but also has some shortcomings. One is because they rely on the smoothness prior, in case the scene is complicated, the smoothness prior will lose its power and causes erroneous correspondences. Also, doing global inference is computationally complex, and not easy to be implemented in parallel for a faster speed.

Different with these works, we adapt sparse matching to obtain the depth information of the scene. Our approach involves detecting distinctive key points in stereo images, computing the descriptor of each point, and finding matches by measuring the similarity of their descriptors. Similar method was mainly used for camera calibration or structure initialization in 3D reconstruction. It has not yet been used in detection tasks because the sparse correspondence sometimes was not sufficient and the computational speed is not fast enough. In our work, to make the depth map not "too sparse", we adapt the multiple operator key point matching approach proposed in [12] to obtain the raw matching result of the stereo pairs. In the other hand, we implemented the key point detection and descriptor computation in parallel, and used a GPU to support the computation for a faster processing speed.

In Fig. 5, we summarized the sparse matching algorithm. In this work, we used two operators, namely the Difference-of-Gaussian operator [13] and Harris operator [10]. For each key point that founded by the DOG operator, a 128-dimensional SIFT descriptor is computed as its descriptor. We use the Euclidean distance as the distance function to measure the similarity between two SIFT descriptors. The Harris operator is used to find corner key points. For this kind of key points, we extract their surrounding 11×11 pixels, and concatenate to an 121 dimensional vector as their descriptor. The distance of two such descriptors is defined as their normalized cross-correlation.

Step.0 Given

- images $I_l(x, y)$ and $I_r(x, y)$
- key point operators $F_i(x, y)(i = 1, ..., n)$
- descriptor computation function $D_i(x, y)(i = 1, ..., n)$
- distance function $E_i(\mathbf{m}, \mathbf{n})(i = 1, ..., n)$

Step.1 Raw Matching

- For i = 1 to n:
  - Compute key point responses $F_i(x, y)$ for each pixel in both I_l and I_r.
  - Set thresholds τ_i and ψ_i.
  - For every $(x, y) \in I_l$, if it has $F_i(x, y) \geq \tau_i$, put it in key point set Ω_l; for every $(x, y) \in I_r$, if it has $F_i(x, y) \geq \tau_i$, put it in key point set Ω_r.
  - For every $(x_k, y_k) \in \Omega_l \bigcup \Omega_r$, compute descriptor $\mathbf{k} = D_i(x, y)$.
  - For every $(x_m, y_m) \in \Omega_l$, find a point $(x_{m'}, y_{m'}) \in \Omega_r$ that minimize $E_i(\mathbf{m}, \mathbf{m'})$.
  - If $E_i(\mathbf{m}, \mathbf{m'}) \leq \psi_i$, put $(x_m, y_m, x_{m'}, y_{m'})$ into matching set $\mathbf{M}$.
- End for.
- Output raw matches $\mathbf{M}$.

Step.2 Refining

- Set threshold ω.
- For every match $(x, y, x', y') \in \mathbf{M}$, compute the Epipolar line of (x, y) as $L_{x,y} \in I_r$.
- If (x', y') and $L_{x,y}$ has a distance larger than ω, remove (x, y, x', y') from $\mathbf{M}$.

Fig. 5 Multiple operator key point matching algorithm.

In the raw matching step, both key point detection and descriptor computation consist of a lot of independent computation. We therefore implement them in parallel for the efficiency purpose. With the raw matching result, Epipolar constrain is applied to remove correspondences that exist apart from their Epipolar line more than a threshold ω (ex. 2 pixels). This could further remove outlier matches and guarantee the quality of matching. In our method, the matching quality is controlled by setting three thresholds (two for matching, one for refining), and is not dependent on the complexity of the scene. This makes it possible to provide reliable measurement for different kind of scenes.

The refined matches are used to do linear triangulation and get the 3D coordinates of each match, with precalibrated camera matrices. We set the left camera's optical center as the world origin, then the z coordinate is the depth d. Figure 6 visualizes a sample sparse depth map and their corresponding key points in the image. For each object hypothesis o_i that we obtained, we collect all the matched key points inside its bounding box and select one that is representative for its depth. Here we use a simple way to select the representative point. We find the nearest k feature points $P_t(t = 1, \ldots, k)$ around the diagonals' intersection of the

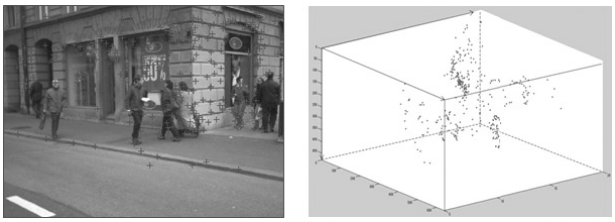

Fig. 6 Keypoints (left) and their 3D coordinates.

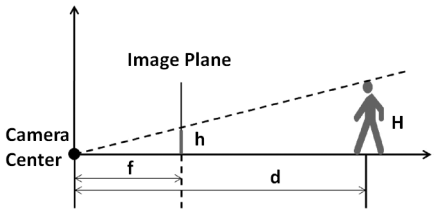

Fig. 7 Similarity relation between pedestrian' height and its imaged one.

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.

With the representative point P_i, the distance d_i could be computed by taking into account the point's imaged deviation with respect to the optical center. In case the camera does not have a wild field of view, d_i approximates to the depth.

5.2 Mapping 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, therefore we have $H \sim N(1.7, 0.085^2)$. This statistic is firstly estimated by D. Hoiem [17] based on the data from the National Center for Health Statistics (www.cdc.gov/nchs/), and has been proved to be effective. However, a limitation is that such distribution does not work for children. And also, our baseline detector is trained for adult pedestrian of the size 70×134 which has a quite different ration with respect to kids. To overcome this issue, one possible way is to train one more baseline detector with different size and utilize a different prior height distribution to detect children only.

Using the similarity relation of the two triangles as shown in Fig. 7, 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 Result

In this section, we present the experimental results. We start by introducing the dataset and our evaluation criterion, then give the quantitative results. After all, we will show concrete results and give some discussion.

6.1 Dataset and Evaluation Criterion

The purpose of the experiment is to see if estimating and using depth information in our proposed way can efficiently improve image-feature-based pedestrian detection in complex scenes. For this, we prepared a difficult dataset by selecting images from the ETHZ tracking sequence [4]. Our dataset contains 133 pairs 640×480 stereo images of complex street view scenes, with 798 annotations as ground truth. It could be found on our project page [14].

In the experiment, we have four systems for comparison: the baseline detector, the UoCTTI detector [5] (baseline + deformable part model), our proposed detection system (baseline + sparse depth) and a simplified implementation of the ETHZ system [4] (baseline + dense depth). As we have mentioned previously, the UoCTTI detector adapts a very expressive model called mixtures of multiscale deformable part model, and uses the part information for detection. The main difference between our proposed system and the UoCTTI detector is that our system uses depth information for detection, while the UoCTTI detector uses part information for detection. In the other hand, the ETHZ system [4], which also uses depth information for detection, is very close to our proposed system. The crucial difference between it and our proposed system lies in the depth estimation method: ETHZ system uses dense depth estimation but our proposed system uses sparse depth estimation. In order to investigate how different kind of depth estimation affect the practical performance, we implemented a simplified ETHZ system by integrating to our proposed system the dense stereo algorithm [16] that used in [4].

In our experiment, every detection system outputs a set of predicted pedestrian candidates for the left camera image. For each predicted bounding box, to be considered as a correct detection, the area of overlap a_o between itself B_p and ground truth bounding box B_{gt} must exceed 50% by the formula:

$$a_o = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}. \quad (5)$$

Multiple detections of the same instance in an image are

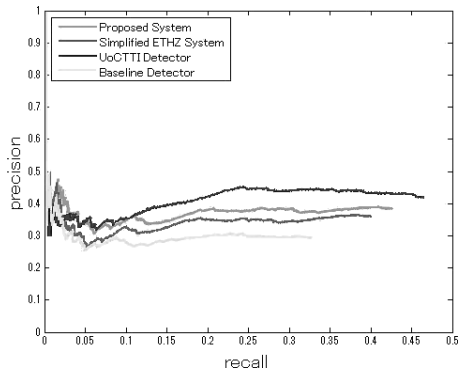

Fig. 8 PR curve for the detection performance.

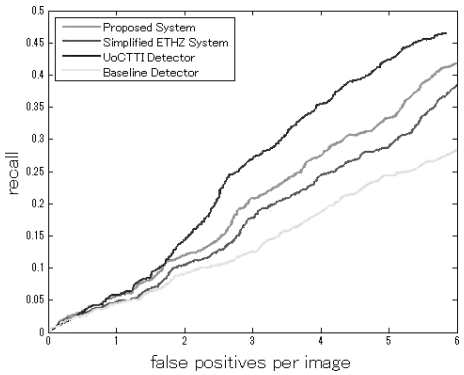

Fig. 9 The plot shows the false positives per image of the four systems.

Table 1 Average Precision for the detection performance.

Proposed System	0.2325
Simplified ETHZ System	0.2232
UoCTTI Detector	0.2530
Baseline Detector	0.1738

Table 2 Details of the one frame detection speed of the four systems.

Proposed System	baseline detector: 1.7 s	1.88 s
	sparse depth estimation: 0.15 s	
	integration: 0.03 s	
Simplified ETHZ System	baseline detector: 1.7 s	11.22 s
	dense depth estimation: 9.49 s	
	integration: 0.03 s	
UoCTTI Detector	8.4 s	
Baseline Detector	1.7 s	

considered as false detections. For example, four detections of a pedestrian instance is counted as one correct detection and three false detections.

All our experiment was done on a 2.83 G Intel Core 2 Quad CPU with 4 G RAM. The sparse depth map computation was partially supported by a NVIDIA GeForce 9800GT GPU with 512 M VRAM.

6.2 Quantitative Evaluation

Our quantitative experiment uses precision-recall (PR) curve [15] to measure how a detection system performs in practice. The PR curve makes it easy to observe the trade-off between the accuracy and how many instances in the image that have been detected. The result is plotted in Fig. 8. Under most recall, the UoCTTI detector, which is shown in blue curve, has maintained a precision near 0.5. By integrating sparse depth information, our proposed system outperforms the baseline detector, and slightly better than the simplified ETHZ system.

Besides PR, we also compute an interpolated Average Precision (AP) [15] to summarize the overall performance of each detection system. It measures the mean precision at a set of equally spaced recall levels, therefore includes measurements of precision across the full range of recall. It penalizes methods which achieve low total recall as well as those with consistently low precision, therefore is ideal for measuring the overall performance in experiments. The AP for the four systems is summarized in the Table 1. Our proposed system brings near 33% improvement to the baseline detector.

Because all the four systems work by scanning image windows and applying classification on each window, in order to evaluate the combinational performance of detection and scanning, we additionally use the false positives per image (FPPI) as a metric. The plot is shown in the Fig. 9. We can see with the same number of false positives, our proposed system is comparable with the UoCTTI detector and better than the other two systems.

The speed of the four systems is listed in Table 2. The baseline detector runs the fastest, only cost about 1.7 second on a single 640×480 image. The UoCTTI detector is quite time consuming and cost near five times as much time as the baseline detector. Since the UoCTTI detector also uses the HOG feature as low level image feature, such disadvantage in speed may mainly boil down to the computational cost from the expressive model it used.

In the current implementation of our proposed method, the detection and depth estimation are done in serial. It costs 0.18 second per frame to update the detection result from the detector, thus the overall speed is 1.88 seconds per frame. Since the two modules in our method are able 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 has space to improve. Because our proposed system is free with the choice of baseline detector, using other faster detector or implementing current detector with GPU processing could make the overall speed be able to meet more applications.

Comparing to our proposed system, the simplified ETHZ system is relatively slow due to the dense depth estimation module. In a sparse depth estimation setting, for one kind of key point operator, our matching algorithm runs in $O(n_l n_r)$ time, where n_l is the number of key points in the left image and n_r is the number of key points in the right image. In case of dense depth estimation, the computation is usually formulated as an image labelling problem, which aims to assign one optimal depth label to each pixel. Standard

Fig. 10 Some results of the four systems. The four rows from top to bottom correspond to the results from the baseline detector, the UoCTTI detector, the simplified ETHZ system and our proposed system, respectively. (TP: true positive, FP: false positive.)

Belief Propagation based algorithm for solving this problem runs in $O(nk^2T)$ time, where n is the number of pixels in the image, k is the number of possible labels for each pixel and T is the number of iterations for optimization. That is to say, for a same pair of stereo image, dense depth estimation is computationally much more complex than sparse depth estimation (especially when the images are of high resolution). However, with the rapid development of GPU computing, it has been reported that a GPU approximation of the dense depth estimation [16] could run at $300\,ms$ per frame on 640×480 images, which is fast enough for practical applications. In this means, we believe that there are still rooms for us to improve our sparse depth estimation method to become faster in the future work.

6.3 Results and Discussion

In Fig. 10 we display some example detection results outputted by the four systems on difficult images from our dataset. The top ten ranked detections are shown in red bounding boxes in each image. The four rows from top to bottom displays the outputs from the baseline detector, the UoCTTI detector, the simplified ETHZ system and our proposed system, respectively.

Compare to the raw output of the baseline detector, our proposed system has made significant improvement in different kinds of scenes. The reason is that we integrated the depth information, and with this cue, the system could better discriminate pedestrians by taking into account the observed height of detection window and update their confidence to become more reasonable. With the confidence updating, the rank of some detections which could get support from their observed height will rise. At the same time, some detection will lose confidence for their inappropriate height observation. This leads to an all-side improvement from the baseline detector.

On the whole, the UoCTTI detector did better than all other systems, especially for some crowded scenes. This may mainly benefit from a very expressive part model it used. The detector uses part information as additional cue,

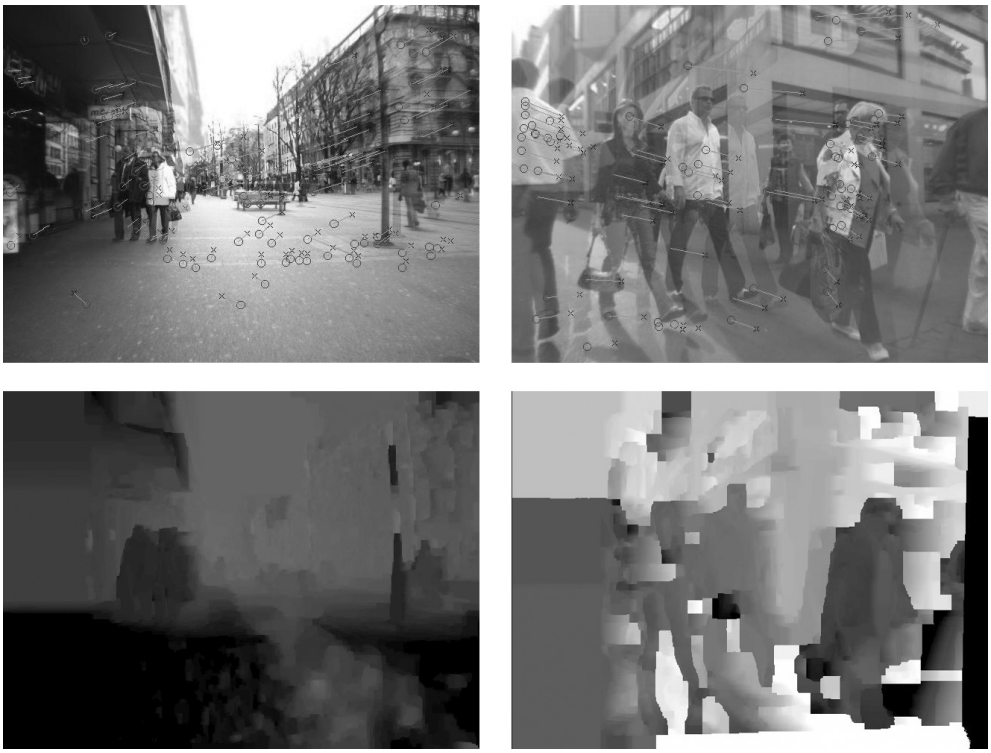

Fig. 11 Examples of sparse matching (top row) and dense matching (bottom row) results. In the top row, we produce a color composite of the left (cyan) and right (red) image, and randomly plot 1/10 of the matches in the composite image (red circles: key points in left images; blue crosses: key points in right images). There is hardly any mismatches. In the bottom row, we show the dense matching result of the same stereo pairs. In case of complex scene, the dense matching sometimes output noisy result.

which leads to the detector be able to preserve sufficient robustness even when heavy occlusion exist. As we can see in the first two columns of the example results, even when the scenes are crowed, it could achieve very good results. However, in some board scene images such like the last two columns of Fig. 10, our proposed system and the simplified ETHZ system sometimes could do even better. 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, using depth information does not have such kind of issues. It brings stable improvements over the baseline detector in different kind of scenes.

Our proposed system and the simplified ETHZ system have similar performance in most images. However, in some complex scenes, our proposed system did a better job. This is mainly because of the difference in depth estimation method. Though the dense depth estimation can provide pixel-wise depth maps which are very informative, it sometime fails in complex scenes (Fig. 11 bottom right). Different with it, the sparse depth estimation in our proposed system only provides confident depth information for a small set of key points. Though the resulted sparse depth map is less informative, it is stable in difference kind of scenes.

7. Conclusion

In this paper, we proposed an efficient approach to pedestrian detection in outdoor scenes by using 3D sparse depth information as additional cue. Our method applies a simplified graphical model to update the detector's output using depth information. It leads to the nature use of context information and effectively improves the performance of baseline detector. To reliably obtain the depth information, we use a novel descriptor based key point matching approach to arrange sparse depth information of the scene. Our depth estimation technique is more stable and computational efficient than dense estimation. The efficiency of our method was shown in our experiment. With minimal additional processing time, our method could improve the detection accuracy of the baseline detector significantly.

However, there are also issues exist in our current method. First is that though the sparse depth estimation in our work is reliable, the depth for every single hypothesis is determined in an explicit way. This may cause the resulting detection system to be sensitive to the error in depth determination. A better way should be used to eliminate the risk from such kind of determination. Secondly, our method currently does not have any occlusion handling mechanism,

therefore is still weak in some crowded scenes. In order to be more applicable in real world applications, such aspect also should be improved. Since the part information [5] is very robust for occlusion, extending our method by using such information will be a very interesting topic. In the future work, we will mainly deal with these two aspects.

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2005.
[2] N. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," Proc. Int'l Conf. Computer Vision, 2005.
[3] S. Ikemura and H. Fujiyoshi, "Real-time human detection using local features based on depth information," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J93-D, no.3, pp.355–364, March 2010.
[4] A. Ess, B. Leibe, K. Schindler, and L.V. Gool, "Robust multi-person tracking from a mobile platform," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.10, pp.1831–1846, 2009.
[5] P. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, pp.1627–1645, 2010.
[6] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," Int. J. Comput. Vis., vol.77, no.1-3, pp.259–289, 2008.
[7] W.R. Schwartz, A. Kembhavi, D. Harwood, and L.S. Davis, "Human detection using partial least squares analysis," Proc. Int'l Conf. Computer Vision, 2009.
[8] 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," Proc. Int'l Conf. Robotics and Automation, 2009.
[9] D.M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," Int. J. Comput. Vis., vol.73, no.1, pp.41–59, 2007.
[10] C. Harris and M.J. Stephens, "A combined corner and edge detector," Proc. Alvey Vision Conference, 1988.
[11] T. Joachims, Making Large-Scale Support Vector Machine Learning Practical, MIT Press, 1999.
[12] Y. Wang and J. Kato, "Reference view generating of traffic intersection," ICIC Express Letters, vol.4, no.4, pp.1083–1088, 2010.
[13] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.
[14] Y. Wang, "http://www.mv.ss.is.nagoya-u.ac.jp/~ywang/dt".
[15] M. Everingham et al., "The 2005 PASCAL visual object classes challenge," Selected Proc. First PASCAL Challenges Workshop, Springer, 2006.
[16] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient belief propagation for early vision," Int. J. Comput. Vis., vol.70, no.1, pp.41–54, 2006.
[17] D. Hoiem, A.A. Efros, and M. Hebert, "Putting objects in perspective," Int. J. Comput. Vis., vol.80, no.1, pp.3–15, 2008.

Yu Wang received the M.Sc. degree in Information Science from Nagoya University, in 2010. He is currently a Ph.D. student with the Department of Systems and Social Informatics, Nagoya University. His research interests are object recognition and video event categorization. He is a student member of IEEE.

Jien Kato received the M.E. and Ph.D. degrees in Information Engineering from Nagoya University in 1990 and 1993, respectively. She is currently an Associate Professor with the Department of Systems and Social Informatics, Nagoya University, Japan. Her research interests include computer vision, machine learning, multi-sensor perceptual computing and their applications. She is a member of IPSJ and a senior member of IEEE.