Detecting Small Objects in Natural Scene using Depth Cue

Jaesik PARK, Yekeun JEONG, Chaehoon PARK and In-So KWEON

Department of Electrical Engineering, KAIST
[jspark, ykjeong, chpark]@rcv.kaist.ac.kr
iskweon@ee.kaist.ac.kr

Abstract We present a novel object detection strategy using depth cue which is robust to small-sized objects. In practical cases, the target objects that we want to detect often appear in very small portion of the input images and most of previous object recognition and detection approaches aiming to achieve high accuracy for several object databases are not applicable. Adopting the depth cue as a prior removes the scale ambiguity and allows us to use adaptive scales for all candidate regions even when the object is seen at small size. The depth cue can be easily obtained by a Laser Range Finder (LRF) or a stereo camera which have become common for many camera-based configurations. By using our proposed method in combination with any other object detection/recognition method, we can expect a substantial improvement of success rate when an insufficient appearance of object causes trouble. We designed an experiment with a scenario in which an indoor mobile robot tries to find several target objects in general environment. The experimental result demonstrates that the proposed method is effective to improve the success rate of small object detection tasks.

1 Introduction

For intelligent robot, object detection is crucial for various tasks. However, the object detection has a lot of difficulties in practical cases since target object in natural scene is usually too small. This insufficient appearance of object often hinder extracting scale or affine invariant features that widely conducted for various object detecting tasks [1][7][8][9][10][12]. In addition, another approach - shape matching for target object detection also gives us a burden due to ignorance of accurate scale of target objects.

In this paper, we present a novel approach for the object detection with handling above difficulties. We reduce the scale ambiguity using the depth cue that comes from a LRF or a stereo camera. From a dense depth map, we can generate depth cues for every candidate region in the input image and resolve the scale ambiguity. To find a location of target object in the input image, we describe adaptive scale candidate regions into a descriptor consisting of an orientation histogram and a hue histogram [6][12]. Chang and Krumm also proposed a similar approach [11]. Chang modeled an object as a color histogram and a geometric information mixture. However, his approach needs

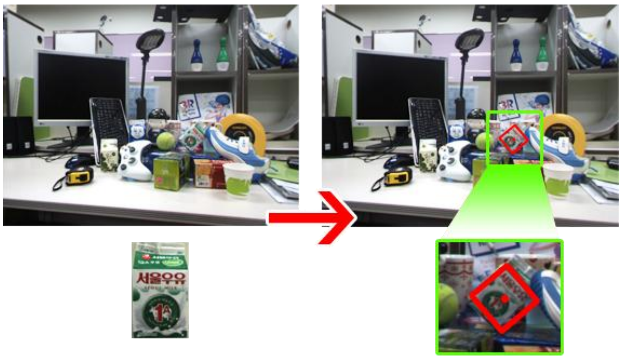


Fig. 1: Left input image is captured by our intelligent robot. When the robot asked to detect milk pack (though milk is small in input image), utilizing our approach (right image) shows appropriate object detection result - the correct location and orientation.

various viewpoints of target object images to model objects. His approach can be burdensome when registered number of object is large. On the other hand, our approach is based on adaptive patches decided by the depth cue. Therefore what we need to prepare is as simple as a small patch of target object as shown in Figure 1. In experimental results, we demonstrate our approach by testing our method on a dataset including 250 images. For each input image, we set target object with various poses: affine and similarity transformation. Also we set various depth conditions to demonstrate the advantage of our approach. The experimental result is presented in chapter 5.

On the other hand, for practical applications, we embedded our approach on our intelligent robot and tested in robot contest named 'Grand challenge 2009'. With our method the robot successfully found exact location of target object.

2 Reducing Scale Ambiguity

2.1 Acquiring Depth Cue

To get a depth cue, we can consider two configurations. One is a stereo camera and the other is a combination of a camera and a range finder (laser or sonar sensor). In the case of the stereo camera, we can estimate the depth of objects using triangulation of corresponding points. As in general approaches, depth cues are calculated from disparity. In our case we used Yoon's method [4][5] for dense matching between two stereo images. If we have a strictly rectified stereo camera like Bumblebee we can estimate depth much easier. Figure 3 (c) shows that the applied stereo method provides a high quality disparity map from a stereo image pair. In the latter configuration including a single camera and a range finder, a range finder can directly provides depth information in metric for us.

2.2 Determination of Patch Size from Depth Cue

We can apply the camera geometry to both of the above cases. Therefore, we can simply use Zhang's method [2] to calibrate the camera. Then we assume that the camera coordinates and the world coordinates are the same. From this assumption, we can get intrinsic camera matrix $\mathbf{K}$ and $\mathbf{P}=\mathbf{K}[\mathbf{I}\,|\,\mathbf{0}]$ which describes relationships between $\mathbf{X}\in\mathbb{R}^3$ and $\mathbf{x}\in\mathbb{R}^2$ as shown in Figure 2. However, we cannot obtain the exact $\mathbf{X}$ from $\mathbf{K}^{-1}\mathbf{x}$ but a ray that passes through camera center. Therefore, we use depth cue d from the object to a camera center. The following process describes how we can determine $\mathbf{X}$ using depth cue d.

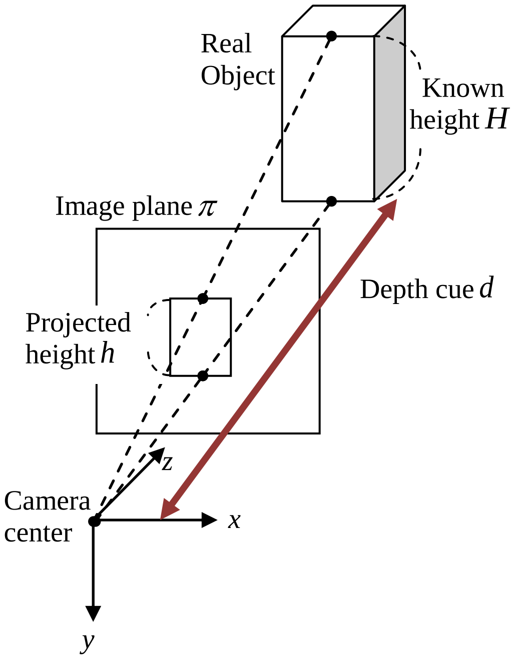
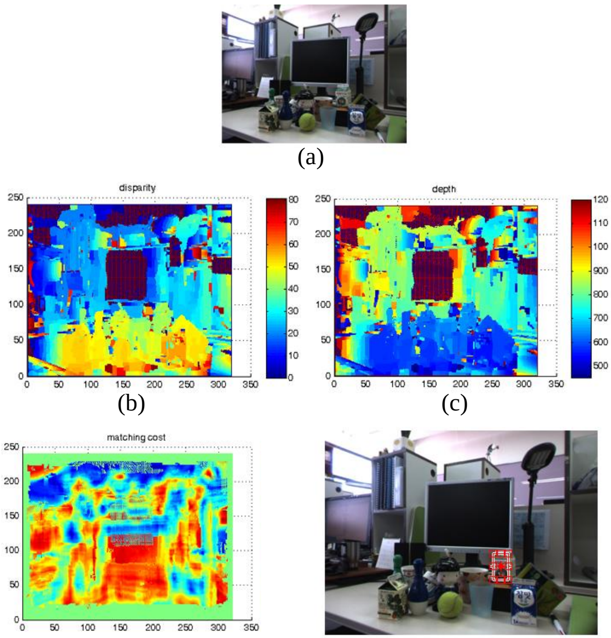


Fig. 2: Projected height (h) can be acquired by approximate depth from range finder and known real height (H).

First, we assume that we know the real height H of the object. Then, we can approximate two 3-dimesional points $\mathbf{X}_1,\mathbf{X}_2\in\mathbb{R}^3$ on top and bottom of the object respectively (Figure 2). Therefore, in homogeneous coordinates, $\mathbf{X}_1$ and $\mathbf{X}_2$ would have the following vector forms.

$$\begin{aligned}\mathbf{X}_1&=\begin{bmatrix}0 & \frac{H}{2} & d & 1\end{bmatrix}^T\\ \mathbf{X}_2&=\begin{bmatrix}0 & -\frac{H}{2} & d & 1\end{bmatrix}^T\end{aligned} \tag{1}$$

After this process, $\mathbf{X}_1$ and $\mathbf{X}_2$ are projected onto an image plane π using $\mathbf{P}=\mathbf{K}[\mathbf{I}\,|\,\mathbf{0}]$ and $\mathbf{x}=\mathbf{P}\mathbf{X}$ [3], and we get new points $\mathbf{x}_1,\mathbf{x}_2\in\mathbb{R}^2$ in the image plane π. Now we can get the projected height h of the object in the image as a value using



(a)

(b) (c)

(d) (e)

Fig. 3: Object detection using depth cue (a) original input image (b) dense disparity (c) depth map (d) matching cost with adaptive size of target object patch (e) final result; the red box located on global minimum position.

the L2-norm equation

$$h = \|\mathbf{x}_1 - \mathbf{x}_2\| \qquad (2)$$

where h is in pixel units. For the above two hardware configurations, even if inaccurate depths are present, those depths does not affect the object detection as significantly since mis-sized patch usually yields lower matching cost to the template patch.

3 Adaptive Patch for Corresponding Depth Cue

Figure 3 shows applicability of a stereo camera configuration. Since we have depth cue, we can have varying template size for each depth. In other words, we calculate patch size h for each pixel since we have depth for every pixel. Template matching is conducted between the patch of each pixel with the template patch of the target object. This procedure is shown in Figure 3 (d) and Figure 4. Due to the consideration of depth, the patch ① is smaller than patch ③. In addition, we have described a patch as a descriptor of fixed-length vector, which will be discussed more in detail in chapter 4, and this allows us to compute vector-to-vector distance (cost) for each pixel in the input image (Figure 3 (d)).

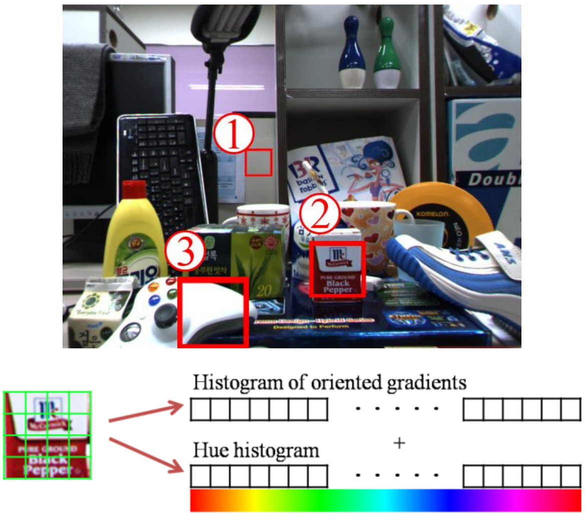


Fig. 4: (Top) top figure shows how the patch size is various in one image. The patch size is enlarged when depth cue value is small. (Bottom) the second image explains how one patch described as fixed-length descriptor vector. We used histogram of oriented gradients and hue histogram

4 Patch Description and Matching

In our approach, a patch of any arbitrary size from the input image or the template image is described as a descriptor of fixed-length vector. We have used geometric cue and chromatic cue of each patch, meaning the oriented histogram and hue histogram. For describing oriented histogram, we adopt SIFT [6] key point description method where image pyramid is skipped since we do not need to release the scale ambiguity. In addition, normalization process clips vector elements larger than 0.2 and re-normalize it as introduced in [6]. We also manipulated the hue histogram in addition to the gradient histogram for the robust object detection. Similar to the previous description method, 4×4 hue histograms have 8 hue bins respectively. Therefore, the geometric cue and chromatic cue have the equivalent length of descriptor vectors. When we are binning a θ bin, the magnitude value is

$$Hist(\theta) = \sum_{H(p)=\theta} S(p)\cdot V(p) \quad p \in X \qquad (3)$$

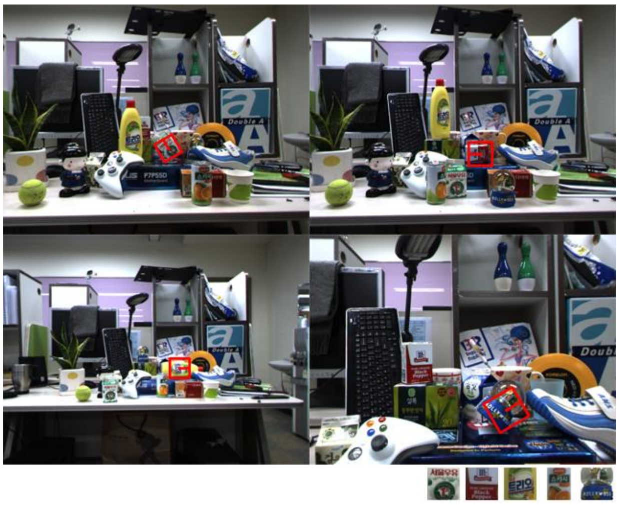
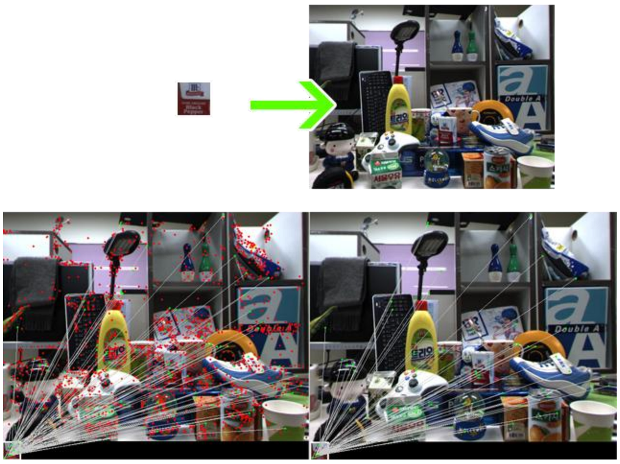


Fig. 6: Object detection results. The test images have lots of clutters and challenging color or orientation distribution. The template patch also presented.



(a) (b)

Fig. 7: Object detection results using SIFT descriptor. (a) shows initial matching result, (b) shows outlier rejection result (RANSAC with fundamental matrix). As seen above, SIFT approach shows non plausible matching result since small number of features extracted from template image.

where X is one sub-region of a patch and is a set of its pixels p. S and V are the corresponding HSV coordinate values for each pixel. Eq (3) allows us to put more weights on the pixel with larger S and V values.

After description process for every patch in one input image, we can detect object position and orientation by finding global minimum using L-2 norm. Empirical experiment concludes that L-1 norm often releases unreliable results due to image noise.

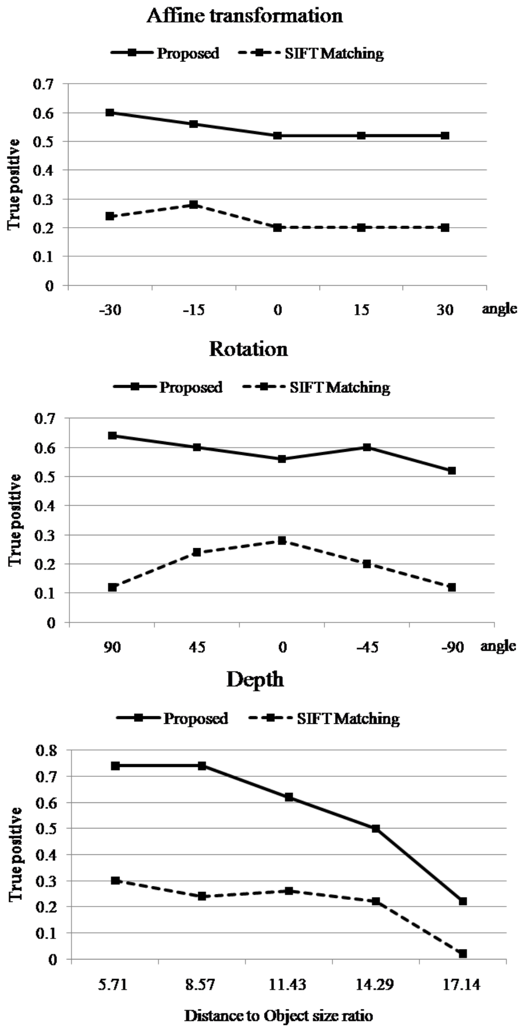


Fig. 5: Experimental result. First and second graph shows our approach works under affine and similarity transform. For various depth our approach shows true positive ratio larger than 0.6 when distance to object size ratio is 11.43.

5 Experimental Results

We evaluated object detection accuracy using our 250 test images which was captured by a calibrated camera mounted on our intelligent robot. In each test image the target object placed on various poses and depths. In addition, for general matching conditions we also placed additional objects together on the table. Thus the test set has lots of clutters. Detecting result (True positive) is presented in Figure 6. Figure 5 shows true positive ratio graphs within moderate rotate and affine transformation. For validating our approach, if the global minimum of vector-to-vector distance is located on specified ground truth region we determine this case as true positive. For SIFT matching, we determine true positive when arbitrary proportion of matching pairs are located on ground truth region. Experimental result shows that about 60% of true positive percentage when camera-to-object distance is 11 times larger than object size. To be specific, if the target object's height is 7cm, projected region's height was 40 pixel in 640×480 input image. Figure 7 shows SIFT matching result. SIFT matching with no depth cue yields too many outliers in initial matching. These outliers are not fixed even we run RANSAC using epipolar constraints (right image in Figure 7).

In addition, for a practical application we applied our method to our intelligent robot. The robot has a laser range finder and set of cameras. So we easily embedded our method and gave some tasks to the robot; finding target object on the table. Each trial showed that the robot finds correct location of object. This approach applied to robot contest named 'grand challenge 2009' held in Pohang, Korea.

6 Conclusion

In this paper, we proposed an efficient object detection approach by using depth cue. With the depth cue, we alleviated scale ambiguity and used adaptive scales for candidate region. Our approach successfully detected small target objects in the natural scene. To handle various poses of target object, all the candidate regions are described as one fixed length descriptor vector.

Experimental result shows that detection performance is reliable under the affine and similarity transformation. The result shows about 60% of true positive percentage when the camera-to-object distance is 11 times longer than the height of objects.

Our approach is applicable to practical object detection tasks for an indoor service robot on which a stereo camera or LRF is commonly equipped, and this is an additional contribution of the proposed method.

Acknowledgment This work was supported by National Strategic R&D Program for Industrial Technology, Korea.

References

[1] A. Hegerath, T. Deselaers, and H. Ney, Patch-based object recognition using discriminatively trained Gaussian mixtures, in proceeding of the BMVC, pp. 519-528, 2006

[2] Z. Zhang, A Flexible New Technique for Camera Calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, No. 11, Nov. 2000

[3] R. Hartley and A. Zisserman, Multiple view geometry; in computer vision, pp. 153-158, 2nd Edition, 2003

[4] K. Yoon and I. Kweon, Stereo Matching with Symmetric Cost Functions, in proceeding of the CVPR, June 2006

[5] K. Yoon and I. Kweon, Adeptive Support-Weight Approach for Correspondence Search, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 650-656, April 2006

[6] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, vol. 60, Issue. 2, pp. 91-110, November 2004

[7] A. Ahmadyfard and J. Kittler, Region based representation for object recognition by relaxation labeling, LNCS 1876, pp.297-307, 2000

[8] X. Wang, J. M. Keller and P. Gader, Using spatial relationship as features in object recognition, NAFIPS 1997, pp.160-165, 1997

[9] G. Bouchard, B. Triggs, Hierarchical part-based visual object categorization, in proceedings of the CVPR, pp.710-715, 2005

[10] B. Epshtein and S. Ullman, Feature Hierarchies for object classification, in proceeding of the ICCV, pp.220-207, 2005

[11] P. Chang and J. Krumm, Object Recognition with Color Cooccurrence Histograms, in proceeding of the IEEE Conference on CVPR, pp.2498, Fort Collins, CO, June 23-25, 1999

[12] S. Kim, K. Yoon, I. Kweon, Object recognition using a generalized robust invariant feature and Gestalt's law of proximity and similarity, Pattern Recognition, No 41, pp.726-741, February 2008