Image Warping for View-invariant Object Matching using Stereo Camera

Jiyoung JUNG, Yekeun JEONG, Joon-Young LEE and In-So KWEON

Department of Electrical Engineering, KAIST
{jyjung, ykjeong, jylee}@rcv.kaist.ac.kr iskweon@ee.kaist.ac.kr

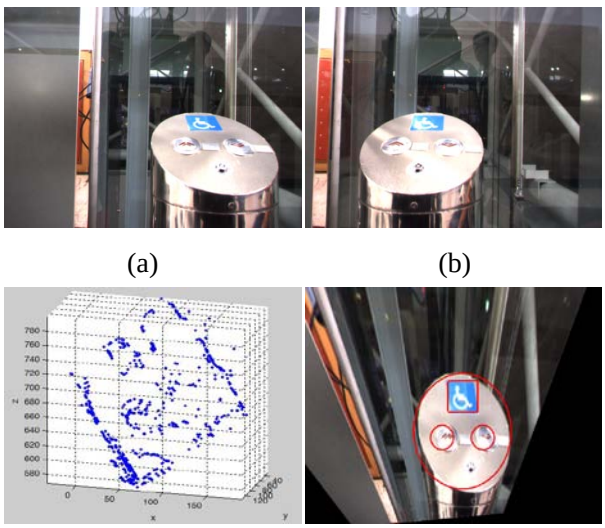
Abstract In this paper, we present a method to increase the robustness and the efficiency of template-based object matching. Instead of preparing templates of various scales and deformations, we use only one template and transform the appearance of the target object by placing a virtual camera seeing a very similar appearance to the template. This approach can solve the chronic problem of scale-dependence and distortion in the conventional template matching, and therefore be applied to the object detection and alignment tasks of general camera configuration systems. The proposed method has shown a good performance in the application of the service robot, Vistro, which attended the Robot Grand Challenge 2009 in Pohang, Korea.

1 Introduction

The field of object detection and alignment is widely explored from the perspective of robot vision application. Among many features, shape of an object can give abundant information about the object. The human visual system often recognizes an object on the basis of the shape alone [3,8,9]. One popular way for object detection using shape information is to prepare a rough silhouette of the object of interest as a template and find a match within input images, which is referred to as template matching. For the robotic application, this kind of shape-based object detection is particularly useful because the template to be prepared for matching is as simple as Fig.1(d) in red.

One of the major problems of the template matching is that it is hard to align the object of interest to the template under different scales, affine distortions and/or non-rigid deformations. Misalignments may cause the system to detect the object in the wrong position or to get no match at all.

In one case of detecting horses in various poses with a few templates, Cour and Shi [3] over-segments the image and find a combination of superpixels that is as close as possible to one of the training templates. In [8], Opelt et al. learn the boundary fragments and make a codebook with representative fragments. Both works are interestingly comparable in the sense that one segments the image and tries to puzzle them up to match the template while the other fragments the boundary and learns them to make a codebook with shattered templates.



(a) (b)

(c) (d)

Fig 1: A challenging example of localizing an elevator button for the service robot, Vistro. We extract edges from rectified stereo input image pair (a,b) and retrieve their 3D coordinates. The dominant plane is searched in 3D space (c) and warped to match the template (d). The template is indicated in red.

However, if we consider objects with rigid bodies, object detection and alignment using template matching becomes a much simpler problem. There is less necessity to break down the input image into superpixels or the template into fragmented boundaries to match them by part, which takes a lot of time and resources. What we have to consider is the difference in scale and affine distortion caused by different relative locations of cameras from the target object. The conventional template matching still cannot be free from the problems and it has to prepare many different templates over various scales and distortions.

To resolve the unpredictable differences, we suggest recovering the 3D shape of objects which is invariant to all the mentioned problems and provides a projective transformation to the known camera position. By transforming the appearance of objects to the one seen from the desired location of camera, we standardize input images instead of assuming all possible cases (preparing various templates).

Using a stereo camera configuration, we can recover 3D information of the scene by estimating depth of each correspondence from a pair of stereo images. There are many approaches for dense stereo matching including ones evaluated on the Middlebury dataset [12]. In order to reduce the false matches in ambiguous areas, Yoon and Kweon [10,11] proposed a window-based method which adjusts the support-weights of the pixels in a given support window based on color similarity and geometric proximity, and a similarity measure named the Distinctive Similarity Measure(DSM) to resolve the point ambiguity problem. Fortunately, a stereo camera which now has become a basic specification to any kind of robot system easily provides the 3D information of the scene. Therefore the proposed method is applicable to a certain service robot without any extra cost.

In this paper, we focus on detecting planar or near-planar shaped objects, because this assumption simplifies the complicated 3D transformation into the 2D projective homography. Given a pair of stereo images, we reconstruct 3D points of the matched edge pixels and find the most probable plane of interest. The plane is then warped to the standardized image patch of certain size and view point. The template matching algorithm is applied by comparing the warped image patch to the single template we have prepared.

The remainder of this paper is organized as follows: Section 2 presents our algorithm including edge-stereo, 3D reconstruction and dominant plane search in 3D space. Image warping using planar homography is explained in Section 3. Section 4 shows our results of template matching on the warped images with real-case challenging examples. A discussion and conclusion is presented in Section 5.

2 Detecting Planar Objects

The original motivation of this method is the task for a robot to locate the exact 3D coordinate of the elevator button shown in Fig. 1-(a,b). This elevator button is designed for the people on wheel chairs. The button is placed on the slanted elliptical surface on top of a 90cm pillar, which is likely to appear very differently as the view point changes. The computation time is also a factor to consider because the robot has to push the button promptly after locating the button. Therefore we have designed the system to find the planar surface in 3D space on which the object of interest (the elevator button in this case) is, and warped the image to align the object with the single template that we have prepared.

The first step is to determine the appropriate plane in 3D space to warp the input image. The selected plane should contain the object of interest. It is assumed that this plane of interest is not homogeneous, and therefore has edges to be detected. Sometimes we can take advantage of previous knowledge about the plane of interest such as the orientation of the plane normal vector.

The proposed algorithm does not reconstruct dense 3D from stereo images. We detect edges from the images and reconstruct 3D points from matched edge pixel pairs only. Since we do not plan to match the given template with the reconstructed 3D points directly, the sparse 3D points are sufficient because they are used to determine the dominant plane. Template matching is performed on 2-dimensional domain after warping the image so that the dominant plane becomes fronto-parallel.

2.1 Edge Stereo

Given a pair of rectified stereo images as input, we apply Canny edge detector [2] to find the edge pixels. Since the images are already rectified, the search range of correspondence is limited to horizontal direction. For each edge pixel in one image, its surrounding local window is compared with every local window centered on each edge pixel with the same y-coordinate in the other image. The normalized cross correlation (NCC) is used to compute the similarity between two local windows. The horizontal search range can be limited if we have previous knowledge of rough depth range for the object of interest. This kind of depth range prior can also be obtained by a laser range finder installed in the robot system.

2.2 3D Reconstruction

Using the calibration information of stereo camera, we can triangulate the 3D coordinates of matched edge pixels. Since z-coordinate of the reconstructed 3D points are initially discrete due to the integer values of pixel disparity, we further refine them using one-dimensional Kanade-Lucas-Tomasi (KLT) feature tracker.

2.3 Dominant Plane Search

A plane can be determined by three points not collinear in 3D space. Parameters for the dominant plane in 3D space are recovered using RANSAC algorithm. Among the planes parameterized by sets of three randomly selected points, we assume that the dominant plane having the most inliers includes the target object. If any weak prior knowledge is available, we use it to pre-filter the random hypotheses of planes to reject irrelevant ones.

3 Warping

Once we decide which plane to warp, we calculate the angle between the dominant plane normal vector and the principal axis of the current camera. With the computed angle and the pre-defined distance from the dominant plane, we can place a new virtual camera in 3D space so that the image plane of the camera is parallel to the selected plane and the intersection between the principal axis of the camera and the plane is identical to the center of mass of the recovered 3D points on the plane.

We then calculate the homography H that warps the input images to a new image seen from the virtual camera. Let $N = [n_1, n_2, n_3]^T$ be the dominant plane normal vector, and let R and T denote the relative rotation matrix and translation vector between the current and the virtual cameras, respectively. Let d>0 denote the pre-defined distance from the dominant plane to the optical center of the virtual camera. The planar homography H is defined as follows:

$$H = R + \frac{1}{d} T N^T \quad \in \mathbb{R}^3$$

The homography between the new image seen from the virtual camera and the input image instead of the dominant plane in 3D is required because we do not reconstruct dense 3D points which requires very wasteful computations. However, calculating additional homography to the virtual camera is much more efficient and likely to yield a better result than to warp the dense 3D points on the dominant plane in which each point is independently triangulated and the warped image naturally shows a very noisy result.

4 Experiments

In this section, we demonstrate improved performance of the proposed method by comparing it to chamfer matching [1] as a conventional template matching method. It is used to correlate the prepared model template with a distance transformation of an edge image. The matching algorithm is robust to illumination changes but weak to the background clutter. Moreover, it is difficult to apply in real images because chamfer distance is vulnerable to scale difference, rotation and translation. Therefore, multiple templates are needed to handle these geometric variations.

In the proposed method, geometric variations such as scale and rotation changes are dealt with standardization of the given input image. Therefore, we can apply chamfer matching using a single template. By providing a similar appearance to the template, it also automatically alleviates the problem with

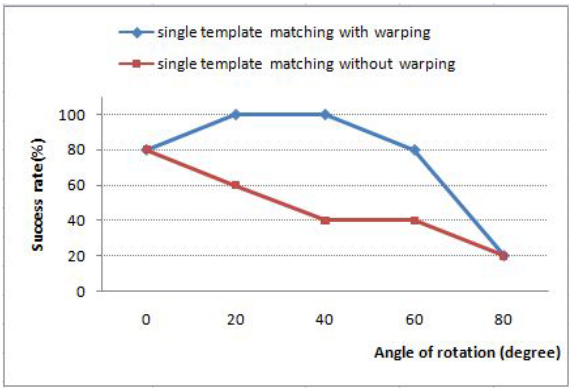


Fig 2: Template matching performance with and without image warping as the angle of camera rotation changes. The angle of rotation is the angle between the planar surface of the object and the image plane. Only a single template is used for the matching.

the presence of the background clutter.

We have tested our method to five different objects which are in planar or near-planar shapes: two kinds of elevator buttons as in Fig. 3-(b,d), two kinds of room number plates, and a clock on the wall. The images are captured from different angles of view points within a certain range of distance. The camera is rotated about either x or y axes of the image plane in 3D space. Since the proposed method is designed for mobile robot system, it is reasonable to assume that the movement of camera is mainly two- dimensional.

Figure 2 shows template matching results with a single template for each object category to the same dataset with or without image warping. The sizes of the objects of interest in the dataset are unknown except that it is controlled to be visible within the image. As the angle of camera rotation increases, the success rate drops sharply for the template matching without performing image warping beforehand. Since we use a single template, if the camera is rotated with large angle or the object of interest is of different scale with the template, the matching tends to fail. However, if we perform image warping before the matching so that the object of interest has the same size with the template with no distortion due to camera rotation, the success rate remains high even the angle of rotation is large. The extreme cases where the angle of rotation is bigger than 80 degrees, the algorithm often fails to find the right plane in 3D to warp resulting in the warped image not to contain the whole part of the object of interest.

Experimental results and the service robot Vistro, which used the proposed algorithm to locate the elevator button, performing its tasks are shown in Fig. 3.

5 Discussion and Conclusion

We have presented a robust and efficient method for shape-based object detection and alignment. In the use of stereo camera, which is now available in most robot systems, the image is warped to the virtual plane of specific scale and viewpoint. This allows us to prepare only one template to match the standardized input image. The algorithm has shown a good performance in the task of locating elevator buttons for the service robot, Vistro. We expect that the proposed method can be a simple but useful solution to the problem of scale-dependence and affine deformation in the conventional template matching.

Acknowledgment

This work was supported by the IT R&D program of MKE/IITA [2008-F-030-01, Development of Full 3D Reconstruction Technology for Broadcasting Communication Fusion]

References

[1] H.G. Barrow, J.M. Tenenbaum, R.C. Bolles, and H.C.Wolf: Parametric correspondence and chamfer matching: Two new techniques for image matching, Proc. 5th Int. Conf. Artificial Intelligence, pp. 659-663, 1977.

[2] J.F.Canny: A computational approach to edge detection. IEEE Trans Pattern Analysis and Machine Intelligence, 8(6): 679-698, 1986.

[3] T. Cour and J. Shi: Recognizing objects by piecing together the segmentation puzzle, CVPR, 2007.

[4] R. Hartely and A. Zisserman: Multiple View Geometry in Computer Vision, Cambridge University Press, 2000.

[5] H. Joo, Y. Jeong, O. Duchenne, S. Ko, and I. Kweon: Graph-Based Robust Shape Matching for Robotic Application, ICRA, 2009.

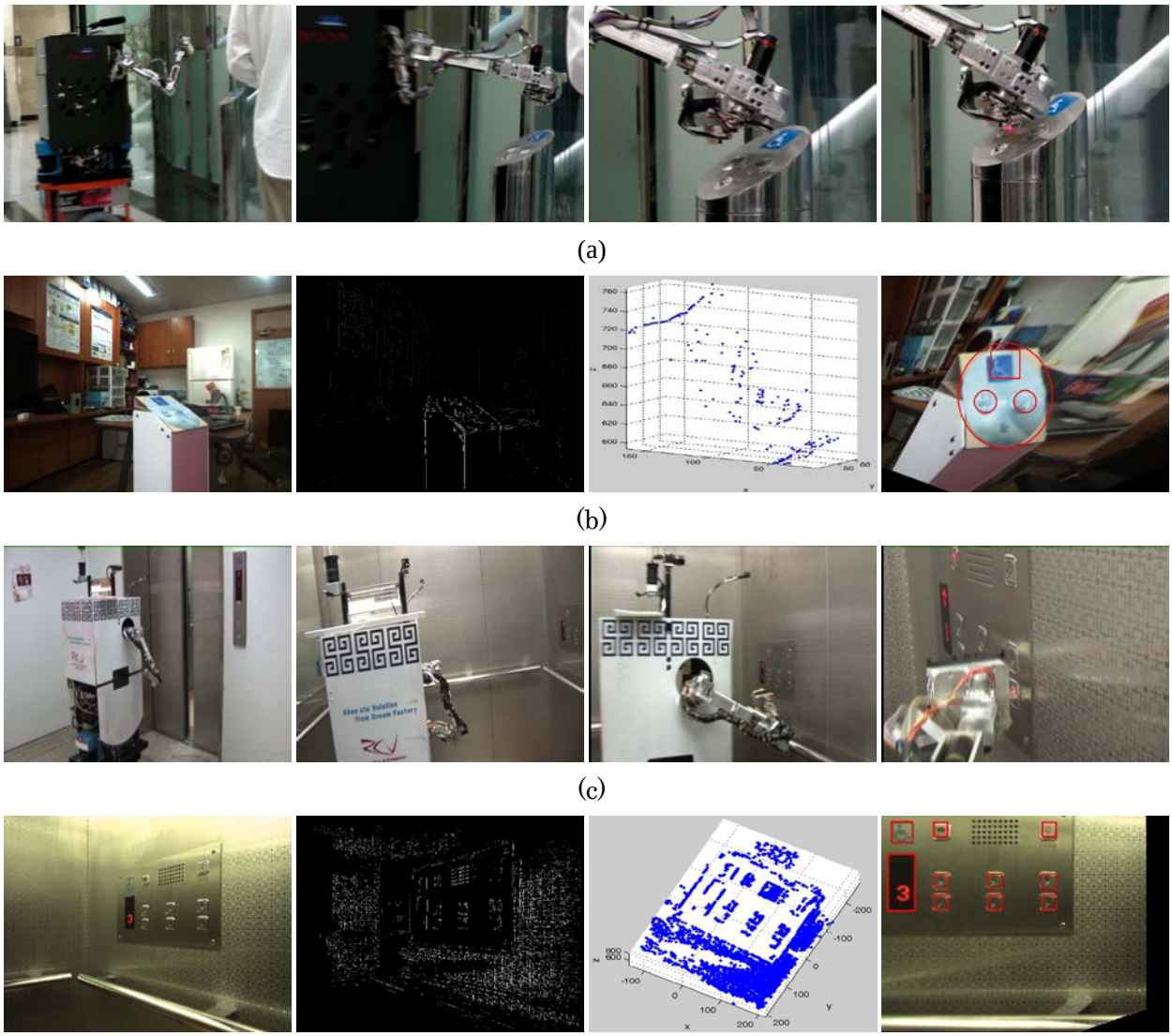


Fig 3: (a) and (c) are sequences of the service robot Vistro locating and pushing the elevator buttons on the standing pillar and inside the elevator respectively. (b) and (d) show the process of the proposed method for extreme cases in which the objects of interest in the given images contain great distortions due to the large camera angles. Figures in (b) and (d) appears in the order of left image of the stereo inputs, estimated disparity, dominant plane in 3D and template matching result.

[6] S. Lao, Y. Sumi, M. Kawade, and F. Tomita: 3D Template Matching for Pose Invariant Face Recognition Using 3D Facial Model Built with Isoluminance Line Based Stereo Vision, ICPR, 2000.

[7] Y. Ma, S. Soatto, J. Kosecka, S. S. Sastry: An Invitation to 3-D Vision, Springer, 2004.

[8] A. Opelt, A. Pinz, and A. Zisserman: A boundary fragment model for object detection, ECCV, 2006.

[9] J. Shotton, A. Blake, and R. Cipolla: Contour-based learning for object detection, ICCV, 2005.

[10] K.Yoon and I. Kweon: Adaptive Support-Weight Approach for Correspondence Search, TPAMI, vol. 28, no. 4, pp. 650-656, 2006.

[11] K.Yoon and I. Kweon: Distinctive Similarity Measure for Stereo Matching Under Point Ambiguity, CVIU, vol. 112(2), pp 173-183, 2008.

[12] http://vision.middlebury.edu/stereo/