Visual Recognition of a Door and Its Knob for a Humanoid Robot

Nosan Kwak, Hitoshi Arisumi, and Kazuhito Yokoi, Member, IEEE

Abstract—This paper deals with the practical problem of how a humanoid robot recognizes a door and its knob in an office environment. This is the initial work where a humanoid robot recognizes, approaches, and opens a door. We propose an integrated solution for visual recognition of a door and its knob with minor constraints. In our approach, the door recognition is transformed into the classification problem of feature points, which can be done quickly enough to conduct it on-line. The knob is extracted by a segmentation method and a few thresholds: blob size, blob ratio, and the distance from the floor. We show in the experiment that our humanoid robot can recognize a door and its knob reliably and quickly.

I. INTRODUCTION

Humanoid robots have several degrees of freedom, high computation power, and a variety of sensors. However, the time has not yet come that any of them serves practical services for humans like the ones as shown in Fig. 1. One practical application to use the high-performance humanoid robot is opening a door autonomously. On one hand, it is a very complex task, which makes it difficult to realize the task with a real robot. On the other hand, it is a highly integrated task of perception, walking, and manipulation, which can show all the functionalities of a humanoid robot. To open a door, a robot should be able to conduct several sub-tasks: recognition of a door, path planning, obstacle avoidance, walking, and manipulation. Among these tasks, in this work as the initial work, we focus mainly on the visual recognition of a door and its knob. We have a very nice humanoid platform, HRP-2 [7], which has been shown in several literatures. For the HRP-2, walking and path planning software is so powerful that on-line walking and manipulation can be easily realized on-line. However, it is hard for HRP-2 and for other humanoid robots to precisely recognize the location of a door and its knob. Moreover, stereo cameras and force-torque sensors are generally the only devices for a humanoid robot to percept the environment, which is the case of HRP-2. The scenario that we plan is that a humanoid robot in an office environment recognizes a door, approaches it, and extracts the knob. Since the HRP-2 has the powerful tool for walking, the remaining issues are firstly, how to recognize a door from far away in an office environment? and secondly how to extract the knob after approaching the door? In addition to the two issues, a robot needs to perform the tasks on-line for our long-term goal, that is, a humanoid robot provides people with practical services. Two main approaches to visual recognition of a door are model-based and appearance-based methods. In the model-based method, a robot has the model of a door and its knob, meaning that a robot knows the shape and dimension of a door and its knob. In this case, a monocular camera is sufficient to detect a door and to compute the distance to a door. However, it is a burden to provide the precise model of a door and knob every time to open a new door. In the appearance-based method, a robot uses natural points and line features as well as other appearance characteristics to recognize a door, since in general a door has distinct edges and several corner points. The disadvantage of this method is that the detection results are severely influenced by light condition and the viewpoint. For both the detection and grasping as well as opening of a door, the work [8][7] of Willow Garage that uses a laser range sensor and stereo cameras on a wheeled robot is very impressive since the robot opens both the pushing and pulling doors. It can autonomously detect a door and its knob, open, and pass through it. It assumes that a robot has prior knowledge of a door such that a door is distinguished from a wall and a knob is located in limited heights, which are general information of a door. The robot senses a door several times to make it sure that the robot detects the knob. In addition, it adopts a double-check approach that derives consensus of results from a laser range sensor and stereo vision. Due to this double-check process, it takes a few tens of seconds to detect the knob in front of a door. It also has the dynamic stability problem to interact with a door, but its effect is much less than the case of a bipedal robot. Recently, Prats et al. [7] proposed a visual servoing method for the door opening that combines vision, tactile, and force-feedback data. In their work, they showed the better performance comparing with vision-force and force-alone cases. However, it has the same burden as a model-based approach. Mansard et al. [7]
showed an impressive experiment on grasping a color ball while the HRP-2 is walking. In their work, however, the visual recognition problem is simplified by using a distinct color ball. We propose an integrated solution for a humanoid robot to recognize a door and its knob. For recognition of a door, we use not only the feature points of a door, but also points around the door since a door itself has few feature points. Furthermore, in general doors do not have a very discriminative appearance, and the appearances of doors vary dramatically as viewpoint changes. Our approach to the fast recognition of a door is to classify an input image using reference images on a robot. Once an input image is considered as a scene of a door, then the robot computes the distance to the door using stereo cameras on the robot. Then the robot walks toward the door and stops in front of the door. The robot searches the door knob using a segmentation technique and general constraints for a door knob: size, ratio, and height bounds. Finally, it computes the position of the knob using stereo vision, and moves to the place where the robot can grasp the knob. This paper is organized as follows: The system structure of hardware and software is described in Section 2 as well as the background techniques. In Section 3 and 4, the proposed visual recognition of a door and a knob are explained in details. After showing the experiment on the HRP-2 robot in Section 5, conclusion and brief description on grasping the knob are given in Section 6.

II. SYSTEM STRUCTURE

In this section, the hardware specifications of our humanoid platform are briefly described, and, the software structure for the visual recognition is described in the sense of data communication between three computers. In addition, the background techniques for the visual recognition is briefly introduced.

A. Hardware Structure

Our humanoid platform is HRP-2 No. 10 as shown in Fig. 2, which is shown in many papers on humanoid robot. Recently, we have added two more joints to the wrists. Now the robot has 32 degrees of freedom (DOF). HRP-2 is 1.54m in the fully stretched-leg posture, and total weight is about 58kg. HRP-2 has four cameras, the three of which are narrow-angle and the other is wide-angle. For the moment the two narrow-angle cameras attached in the ears as stereo vision is mainly used for visual recognition. The wide-angle camera that has view-angle of 120° is used only when it is necessary. The camera of the narrow-angle stereo vision system has view-angle of 25° and 33° in horizontal and vertical directions, respectively. The specifications in detail are presented in [?].

B. Software Structure

Since our humanoid platform, HRP-2 has two computers inside; one for controlling joints and the other for vision processing, it requires a communication method between them. Moreover, we use a terminal computer to run OpenHRP3 [?] that manipulates the control computer in the robot. For the communication among the three computers, we use CORBA that provides robust and efficient communication. The vision CORBA server that is implemented using CORBA with C++ [?] provides several functions such as image capturing and visual data processing. For example, the vision server periodically captures images from the stereo cameras, and when the client (the terminal computer) asks the distance to the door, then the vision server conducts the matching process and returns the result to the client. After the client receives the feasible distance, it sends the distance to the controller computer in the robot where several control functions commercially available exist as CORBA servers. Here the client is implemented in JAVA to communicate with the OpenHRP3. As being seen here, CORBA is a universal communication architecture that allows different types of programming languages and devices to exchange data.

C. Background Techniques

The key techniques for recognizing a door and its knob can be found in open source communities such as OpenCV [?] and ROS [?]. Of the techniques, the important ones are listed below:

- STAR Detector
- Randomized Tree Classifier
- MSER

To recognize of a door real-time, the STAR detector [?] is used as a feature extractor in our approach. It is the improved version of the Center Surround Filter (CenSurE) [?]. The randomized tree classifier is used to quickly match feature points, which was proposed by Calonder et al. [?]. The maximally stable extremal region (MSER) is used for blob segmentation that extracts distinguishable regions such as door knob on a door plate. It was proposed by Matas et al. [?] to find correspondences between image elements from two images with different viewpoints, but now it is commonly used for blob segmentation thanks to its robustness against the viewpoint and the illumination changes.
III. VISUAL RECOGNITION OF A DOOR

Our approach to recognize a door is to match an input image with a reference image that has visual data around the door. In other words, it is required that the robot should extract the best matched reference image whenever it searches a new door. In this case, not only features on the door but also features around the door such as features of light switches are used for the matching. This matching is done on-line using the randomized tree classifier that is learned off-line with the similar image to the reference image. Note that the term 'reference image' is used to match with the input image whenever a new input image comes, and the term 'base image' is used to train the classifier off-line. If the input and the reference images have a certain number of matched features, then it is considered that the input image is the scene including the door. Note that the matching process tries to match the patches of the reference image with those of the input image, because according to our test results, this way gives the slightly better performance than the reverse way. As a result of stereo geometry, the 3D points of the matched features are computed, and the mean of the 3D points excluding outliers as a rough distance to the door is given to the robot so that the robot walks toward the door. The robot stops walking when it is close enough to detect the knob of the door. Until it detects the knob from both cameras, the robot turned the head horizontally. The algorithm mentioned above is shown as a flowchart of recognition of door and its knob in Fig. 3. To match a scene with the reference image, the features of the input image should be extracted quickly for an on-line application. The randomized tree classifier is learned using the image shown in the left of Fig. 4. The learning process took 46 minutes for 80 trees, 9 depths, and 5000 views (for the parameter setting please refer to the paper, [2].) Figure 4 shows an example of matching process between a patch (32x32 pixels) of the input image and the reference image. The signature of the image patch is quickly computed by the learned tree classifier. Then the matching distance between the two patches is computed using the Euclidean distance between two signatures. The matching process continues until it finds the best patch that has the smallest matching distance. The success of the door recognition is decided by two ratios: matching ratio, $R_m$ and duplicated ratio, $R_d$, which are defined as follows:

$$R_m = \frac{N_m}{N_k},$$

$$R_d = \frac{N_m}{N_{rm}},$$

where $N_m$ is the number of matches, $N_k$ is the number of keypoints in the input image. And the $N_{rm}$ is the real number of matches after removing the duplicated matches, since several patches in the input image may be matched with one in the reference image due to the brute-force matching scheme. If the input image is successfully matched with the reference image, then 3D points of the matched features in the input image are computed using the stereo geometry.
IV. VISUAL RECOGNITION OF A KNOB

In general, a door knob belongs to a material that has less feature points and is reflective. To get the position of a knob, we can not use the disparities of point features on a knob because the stereo matching fails. Here we explain how to extract the knob in an image and compute the position of the knob from the robot.

A. Knob Extraction

According to our test, the maximum difference of feature points on a knob in the depth (forward direction) was 1.5 meter in the case that the robot was about 1 meter away from the knob. Instead of feature points, we can use the edges of a knob, which can be easily detected from the door plate since there are usually no other materials on the door plate than the knob. However, it is not robust since the threshold to extract edges should be changed to the conditions such as illumination. In this work, instead of edge detection, we propose an approach that uses a blob segmentation method, MSER with minor constraints: blob size, blob ratio, and blob height from the floor. Through several tests, we found that a knob can be repeatedly extracted when the knob is located around the center of an image. In other words, to robustly detect the knob, a humanoid robot lowers the view to make the knob placed in the center of the image plane. The amount of lowering the view can be decided by the height bounds of a general door. Figure ?? shows how to extract the knob from the original image with the three constraints. First, we run the MSER on the original image, and the result is shown in Fig. ??(b). Second, we apply the threshold for blob size, $T_s$, and the result is shown in Fig. ??(c). Third, we apply the threshold for the width/height ratio of a blob, $T_r$, and the result is shown in Fig. ??(d). In this example, we could extract the blob of the knob with two thresholds. For more robust recognition, we can apply the threshold of the height of a knob from the floor, $T_h$. In general the knob should be located in the reachable height not only for the normal people but also for the handicapped. Since the robot knows the transformation matrix from the ground coordinate, $\Sigma_G$ that is attached under the left foot, to the coordinate of the left camera, $\Sigma_C$, it can compute the height of the blob by using the simple geometry of the view-angle in vertical, $\theta$ and the distance to the blob, $d$ as shown in Fig. ??(d). Note that the camera of the robot should be parallel to the floor plane to compute the height of a knob in the following method. The $d$ can be obtained by subtracting the walked distance from the door distance at the initial location. To obtain the ratio $\kappa$ of length per pixel, $h$ in the figure is computed, followed by the height of the knob $h_{\text{knob}}$ from the floor by adding the height of the camera, $h_{\text{cam}}$ as follows:

$$h = d \times \tan(\theta/2), \tag{1}$$

$$h = \frac{T_h - P_{cy}}{T_r - P_{cy}}, \tag{2}$$

$$h_{\text{knob}} = \kappa(P_{cy} - P_y) + h_{\text{cam}} \tag{3}$$

where the $P_{cy}$ is the y pixel value of the principal point of the camera and $P_y$ is that of the centroid of the knob based on the image coordinate, $\Sigma_I$. Through the three thresholds, a knob can be extracted robustly and repeatedly. Note that the $h_{\text{knob}}$ is roughly computed since the $d$ is not accurate. Thus, the height is checked with the stereo geometry once more for ones that satisfy the three constraints.

B. Computation of Knob Position

Once the knob is detected from the left and the right cameras on the robot, the knob position can be easily obtained using stereo geometry. Before applying the stereo geometry, the left and right images should be rectified so the rows of the left and the right images are aligned. When using feature points on the knob, there are many wrong disparities...
was set to 1000 R the upper limit for the duplicated ratio.

For the knob extraction, the constraint for the blob size, knob. In the figure, the robot recognized the door at the initial position to the position where to grasp the detected knob. Finally, the robot moves to the place to grasp the knob.

V. EXPERIMENTAL RESULTS

The experiment for the visual recognition using the HRP-2 humanoid robot was conducted in an office of AIST as shown in Fig. 2.

A. Experiment Setup

The robot heading to the door was standing at a place about 5.3m away from the door. The scene of the door has been slightly changed since the reference image was taken. We used OpenHRP3 and the on-line pattern generator developed by Kajita [?] to make the walking motion. For the feature matching, the size of the sliding window was set to (160,120) for horizontal and vertical directions, respectively. The lower limit for the matching ratio, $R_m$ was set to 0.5 and the upper limit for the duplicated ratio, $R_d$ was set to 3.0. For the knob extraction, the constraint for the blob size, $T_s$ was set to 1000 and the lower and upper bounds of $T_r$ were set to 3.0 and 6.0, respectively. The range of the allowed height of a knob, $T_h$ was set to between 0.8m and 1.2m.

B. Visual Recognition Result

Figure ?? shows some still images of the experiment from the initial position to the position where to grasp the detected knob. In the figure, the robot recognized the door at the initial position (a) and walked to the door after computing the distance to the door (b). After approaching the door (c), it lowered its waist to search the knob (d) and computed the position of the knob. Finally, it stood up (e) and moved to the final position to grasp the knob (f). Figure ?? shows the result of recognition of the door, which took about 120ms. In this case, the $R_m$ was 0.68 and the $R_d$ was 1.46, which are acceptable as a door scene. In the figure, there are some false matches since some features in the reference image do not exist in the input image. The true distance from the robot to the door was 5.3m and the resulting distance of the matching process was 5.4m. More specifically, the mean and the variance of all the 3D points excluding outliers, based on the $\Sigma_C$ were $(-0.1562m, -0.0236m, 5.4038m)$ and $(0.0239, 0.0867, 1.1001)$, respectively. We made the robot stop at the 1.0m away from the door, but the true stop-distance from the door was 1.1m, which influenced the initial constraints for the height of the knob. To better detect the door knob, we made the robot lower the height of the cameras to 1.09m from the floor after it approached the door. Until it detects the knob from the both cameras, the robot turned the head horizontally. After detecting the knob, the robot computed the position of it and moved to the place to grasp the knob. The computation time of each task is shown in Table ??.

Regarding the repeatability of the visual recognition, the proposed solution could mostly recognize a door with a different viewangle and position as well as a different position of the poster on the door. However, in the case of the sunlit office, the failure of the recognition was observed from time to time. The knob extraction was also robust during the several experiments with different types of knobs as long as the robot could see the whole knob in the both images, especially robust when the knob is located in the center of images. The real size of the knob was 0.12m in width and 0.025m in height, and the resulting size from the vision was 0.08m in width and 0.015m. The resulting position of the

![Fig. 7. Disparity of the centroids to get the 3D position of the knob.](Image)

![Fig. 9. Resulting matches between the reference image and the input image.](Image)

<table>
<thead>
<tr>
<th>Task</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rectification</td>
<td>25</td>
</tr>
<tr>
<td>Door Matching</td>
<td>120</td>
</tr>
<tr>
<td>3D points for Door</td>
<td>0.69</td>
</tr>
<tr>
<td>Knob Extraction</td>
<td>786</td>
</tr>
<tr>
<td>3D point for Knob</td>
<td>0.114</td>
</tr>
</tbody>
</table>
knob based on the $\Sigma_G$ was $(0.825m, -0.074m, 0.98m)$ in the case that the true position was $(0.88m, -0.02m, 1.0m)$. According to the result of the knob recognition, the robot moved to $0.37m$ along $x$-axis and $-0.34m$ along the $y$-axis to grasp the knob. Due to the few centimeter errors, the collision between the gripper and the knob might happen when the robot tries to grasp the knob.

VI. CONCLUSION

This work is the initial work of the door opening for a humanoid robot from visual recognition to manipulation. The originality of this work is that we propose an integrated solution which visually recognizes a door, approaches it, and detects its knob for the door opening. More specifically, for the visual recognition of a door, we solved the matching problem as a classification one, so that the robot could quickly match two images: one from the database and the other from the camera. Moreover, a knob could be robustly extracted by combining with the motion of lowering the camera. As seen in the experiment, the size and the position of the knob was not so accurate that the collision-free grasping era. As seen in the experiment, the size and the position of the knob could be robustly detected by using combined with the motion of lowering the camera. Moreover, a knob could be robustly extracted by combining with the motion of lowering the camera.

VII. ACKNOWLEDGMENTS

This work was supported by Grant-in-for Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Foreign Researchers (20-08610). The authors would like to thank the members of Humanoid Research Group and CNRS-AIST JRL UMI3218/CRT at AIST.

REFERENCES