ステレオカメラを用いたランダムに積み重なった状態の物体の認識技術

Stereo Vision Based Piled Object Recognition System for Robot

王　晓霞* 熊　怀欣* 李　磊*
Xiaoxia WANG Huaixin XIONG Lei LI

要　旨

ステレオカメラの画像情報を基にしたロボット技術が，通常の2次元カメラによるものに比べ注目を集めるようになっている．このような中，リコーでも産業用ロボット分野で使われることを目指したステレオカメラの開発が行われている．現在，3次元画像を使うロボットの研究や応用は，その多くが予め決められた形状の物体を扱うことを想定しており，形状が決まっていない自然の物や手作りの食品などについては扱いが難しい．

本研究は，形状や大きさが一定ではない物体がランダムに積み重なった状態での物体認識技術を扱うものである．このシステムは，積み重なった物体の集まりから1つずつ物を取り出すといった，製造工程内で人が実際に行わなければならないような動作をロボットに行わせることを狙う．

我々が開発した認識システムは，リコー製のステレオカメラから得られる3次元情報と2次元情報を使う．3次元情報は，物体の全体像の解析，全体的な積み重なりの情報の取得，物を掴むためのロボットの位置の算出などに用いられる．一方，2次元情報は，3次元情報を検証したり，物体同士の細部の重なりを見たりするのに用いられる．現時点では，1秒以下の計算時間で約90%の精度を実現している．

ABSTRACT

3D vision based robotic systems have recently attracted more attention than ordinary 2D vision ones. As such, Ricoh is now developing a special model of stereo camera aimed at industrial robots. Current research and applications of 3D vision guided robots are mostly targeting machine-made objects, which are characterized by strictly defined and unified shapes and sizes. The approaches could not work well for objects such as natural produces or hand-made food.

This paper focuses on how to recognize randomly piled up objects that are of irregular shape and size. The system aims to guide the robot to pick up objects one by one from a batch of piled junk, which is now done by humans in factories.

Our recognition method uses both 3D and 2D vision using a Ricoh stereo camera. The 3D information is used for overall target analysis, global segmentation, pose/position, and graspability estimation; the 2D information is for 3D information verification and clear local segmentation. At present, the method has achieved an accuracy of about 90% while the processing time is less than 1 second.

* リコーソフトウェア研究所（北京）有限公司
Ricoh Software Research Center (Beijing) Co., Ltd.

1. Introduction

As vision-aided robots have been greatly developed in recent years, people are continuously trying to extend the scope of robot capabilities. One of the most central topics is using 3D vision to guide robots to pick and place under more complex and unstructured environments, which is almost impossible with a 2D vision system. A simple example is when the objects are piled together and the robot needs to pick and place them one by one.

By targeting the FA (factory automation) domain, Ricoh is now developing a binocular camera that offers real time distance and grayscale video streams. With a working range of 800 mm to 1200 mm, which is a typical operational distance for a lightweight robot aiming to replace human labor in a factory, its precision is as high as 1 mm, which is enough for most tasks involving picking up objects by hand.

Fig. 1 Left: system composition, right: chickens.

While we know there exist many utilities or applications that have been developed targeting piled objects, most of them focus on working parts that are shaped strictly by design, such as mechanical parts or objects that have relatively smooth surfaces, such as boxes or cups. In a food factory, e.g. a bento making plant, various irregular shaped foods including fried chickens, fish pieces, sushi, and chopped carrots are to be packed into bento boxes. They are randomly piled in some containers and need to be picked up one after another. This is quite a challenging task for a robot since the food surfaces are quite rugged and the food shapes are irregular.

To best utilize the advantages of both 2D and 3D vision provided by the stereo camera, our approach is that we use depth information to do global analysis of the food pile so as to get to the top layer of the pile, on which we can use a gray image based approach to segment each object more robustly. Then, after segmentation, again, we use the 3D information of the segmented object to evaluate the position, pose, and graspability of the target. Then, we output the real world bounding box and gripper positions for each graspable object identified to the robot controller.

To make our system applicable for various types of targets, it has a registration module that can learn and store shape, size, and gray level related features of each target. With this, on one hand the system can recognize the target type when facing a new pile of objects; on the other hand, the physical features are used to support object segmentation and position evaluation.

The system settings and our target objects are shown in Fig. 1.

2. Related Work

As Microsoft Kinect becomes popular, lots of work have been published on depth image based cluttered object segmentation. Among them one of the typical approaches is LINEMOD[1] based. LINEMOD relies on two different features: color gradients computed from the color image and surface normals computed from the object 3D model. For example, Stefan Hinterstoisser[2] has proposed model based training, detection, and estimation in a heavily cluttered situation. The prerequisite of their methods is a definite known 3D model whether predefined or got from training. Other works use, for example, the gradient of depth feature for planar segmentation[3], which relies on the smooth surface features of the objects.

On the other hand, some work has been done on graspability evaluation for robot grippers in a bin picking task[4]. The method basically uses grasping template matching, which is comparatively time consuming and only suitable for flat and non-tightly cluttered objects.

There is also related work in range image based intelligent surveillance, where typically depth slicing and region growing methods are combined for segmentation[5].

Although the above related works have successfully solved problems in their particular cases, none of their assumptions fit our target, which is irregular and has non-fixed shapes, highly rugged surface, and closely cluttered positioning.

3. Technology Solution

3-1 Overview

As Fig. 1 shows, our system is composed of four parts: Ricoh stereo camera, patterned light, robot, and a work plane. The stereo camera generates 1280 x 960 pixels with a gray/depth image at 30 fps with a distance precision of 1 mm. The patterned light is designed to facilitate the stereo camera to produce enough dense depth data. The work plane or support surface is a black colored planar field vertical to the camera axis. Although the robot and its controller characteristics are not specified here as they are beyond the concern of this paper, we assume the robot has two grippers with given sizes (length, width, and thickness). They are used to judge whether the target has enough marginal space for the robot to move its hand around for picking up objects.

Fried chicken picking is one of the most challenging automation task, because every piece of chicken is of different shape and its surface is quite uneven, especially when they're piled together. So we set our target here as a pile of fried chicken bumps to be picked up one by one. Each chicken has different and irregular shape, and its surface is rugged with depth discontinuity here and there, while each one has same color and texture, as shown in Fig. 2.

3-2 2D and 3D combined segmentation

Given the special characteristics of fried chickens, it is nearly impossible to do segmentation either with simply a 2D image or a purely 3D image. So far, most binocular 3D cameras, constrained by their depth generation algorithms, have noise and imprecision in their output images, especially at the verge and in concave areas, while gray image has no such problem. On the other hand, depth information is critical for occlusion and pose estimation related computation. Therefore, our approach aims to make the best use of both gray and depth data for chicken segmentation and graspability evaluation.

The policy for the pick-up sequence is from the top to the bottom of the pile. The objects on the top layer of the pile will be analyzed for graspability evaluation; if no object is graspable, the “shake” operation is indicated.

The overall process of our method is described in Fig. 2. First, the depth image is analyzed so as to find the best layer segmentation depth and output the top layer depth image. It is then used to get the top layer gray image. The distance map based watershed method is then used to segment each object in one layer; finally, each segmented object position/pose and graspablity is evaluated based on its depth information.

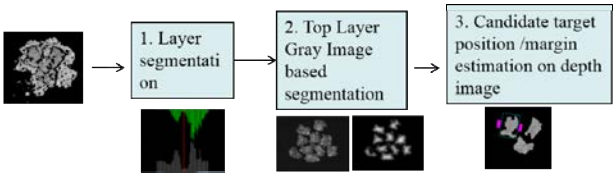

Fig. 2 Process overview of the method.

The details of each step will be described in the following sections.

3-3 Top layer segmentation

For a pile of objects, it is not necessary to make full analysis for each of the objects. Our policy is that we first get the information about those on the top level of the pile and mask off others for further analysis. In this paper, we assume that chickens have a length/width ratio smaller than a specified value (e.g. < 2.5) and height/width ratio greater than a specified value (e.g. > 0.3), and the chickens are piled in a stable way, i.e. they are not squeezed together.

Figure 3 shows the pile depth image, histogram, and top layer image. From the depth distribution histogram, we can see that there is a "valley" between the layers in the pile. From that we can draw a line (red line) so as to separate chickens of different layers.

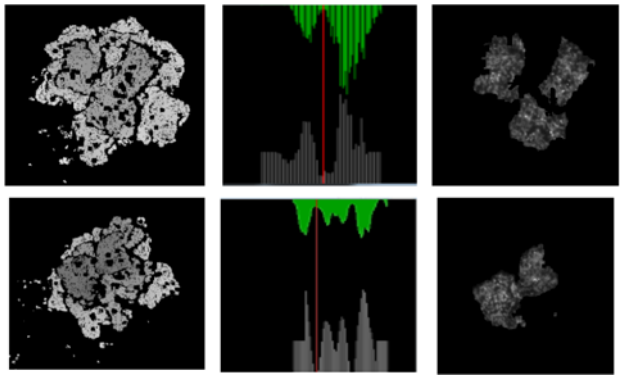

Fig. 3 From top to bottom: 2 layer and 3 layer case; From left to right: piled depth image, histogram of depth distribution in green color (X-axis is relative depth in millimeter, Y-axis is normalized total of that depth in whole image), contrast of histogram in gray, and layer separation line (red), and top layer objects in gray image.

To get the valleys, the following equation is used to calculate the contrast of the histogram, where n is the parameter for adding up the contrast of the valley and its neighbor, i is the inspecting position, and h_k is the value of position k in the depth histogram. The minimum value is where the layers are separated within a given range.

$$h^* = \min\left\{ Contrast(i,n) = \sum_{k=i-n}^{i+n} h_k - \sum_{k=i-2n}^{i-n-1} h_k - \sum_{k=i+n+1}^{i+2n} h_k \right\} \forall n \quad (1)$$

The number n is used for filtering the unexpected local peak. In our experiment we set it to 3. After getting h^*, we perform a threshold operation on the gray image as the following equation shows, by which we can remove all others except the top layer objects.

$$\begin{cases} gray'(x,y) = gray(x,y) & depth(x,y) < h^* \\ gray'(x,y) = 0 & depth(x,y) > h^* \end{cases} \quad (2)$$

In the equation, $gray'(x,y)$ and $gray(x,y)$ is the value at (x,y) of the new and original gray image, respectively. $depth(x,y)$ is the depth value at (x,y). With the one layer gray image of the pile, we can do further object segmentation, which is described in the next section.

3-4 Tiled object segmentation

Through the above segmenting from the depth point of view, only the top layer chickens are retained on the image with some interspaces between them. Those interspaces will be used to segment each other. Here, we adopt an improved region growth method based on the watershed drawdown idea. The segmentation process can be interpreted as shown in Fig. 4. Firstly, we separate the foreground from the background to get the top layer chicken binary silhouette image (b), then we execute distance transform on a binary image to get the distance map image (c); here, the distance transform of a binary image is the distance from every pixel to the nearest non-zero-valued pixel according to Felzenszwalb[6]. The distance then is normalized to 0 to 255 to prepare for watershed drawdown. Before region growing, we mark some initial cluster regions with thresholding operation on the distance image. Each cluster region must be connected by blobs of pixels inside each of the foreground objects, which represent the segmentation seed and the base of the region growth as shown in image (d). Then, we simulate the water level dropping and start with the initial cluster

region to grow in each water level region alternately until the water level reaches the bottom level. To prevent over segmentation, we only grow those existing regions and do not process new region occurring, and finally we can get each chicken region labeled indifferent colors as shown in (d).

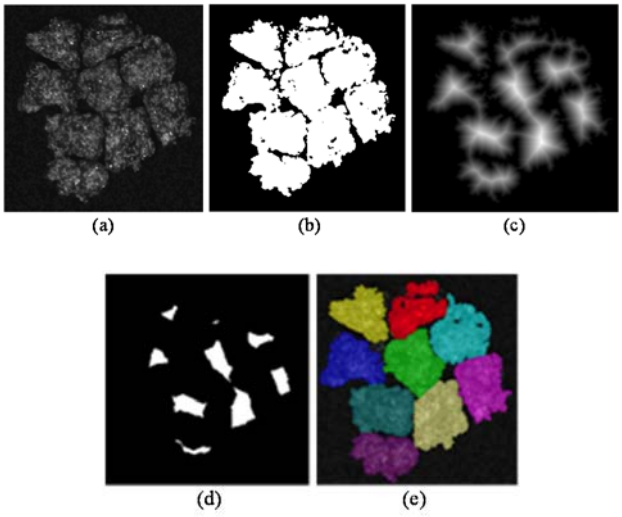

Fig. 4 Process of tiled object segmentation (a) top-layer image, (b) binary silhouette image, (c) distance transform image, (d) initial cluster region, and (e) chicken segmentation result, different color indicating each segmented chicken.

If the initial cluster region is not properly marked, some touching chickens will be segmented as one. In this case, we can pick out all suspected touching regions with the knowledge of the chicken size and shape, and send them to the iterative segmentation process until each individual region meets the constraints of size and shape. From the result shown in Fig. 4 (e), very closely bordering chickens are correctively segmented.

3-5 Object pose/position and graspability estimation

After we get the boundary of each object from the last step, we return back onto the depth image. We get the 3D information of each individual object, and then for each one we compute its bounding cubic information and gripping availability for the robot to pick up, as Fig. 5 shows.

We first perform PCA (principal component analysis) on the depth data or 3D point cloud of the target object and get three Eigen vectors from big to small according to their Eigen values, X', Y' and Z'. Then, the 3D point cloud is rotated onto the new coordinate system formed by X', Y', and Z'. From the new 3D point cloud, we calculate the bounding cubic, and along the primary and secondary axis, the margin space is searched according to the size of the gripper. If there is room for the gripper to pick an object, the bounding box and gripper position will be outputted.

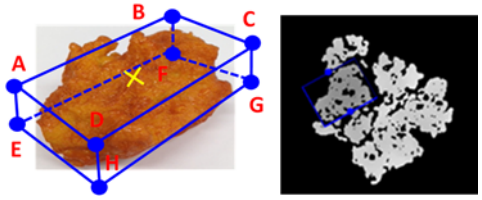

Fig. 5 Sketch of bounding box and sample result on depth map.

3-6 Discussion

The above presented segmentation method is not only applicable to fried chickens but to other kinds of objects of regular or irregular shape, so long as they are 1) bulky (no strong concavity along their border and no hollow space in them), 2) piled not in an extremely disorderly fashion, 3) of a known size range with a tolerant level of variance. Among these, we suggest No. 3) should be learned through a registration step, which is implemented in our system. Because it is not relevant, it is not discussed in this paper.

As to graspability, because our computation is based on 2.5D information that ignores the unseen sides of the object, it should be used with some adjustment in real robot picking according to both the robot gripper characteristics and the known mass distribution features of the object.

However, one important assumption of our method is that the objects are piled neatly, not hastily and messily. We assume that often this condition could be satisfied when dealing with food.

4. Results

We have conducted experiments with 25 different-shaped fried chicken models in different piling situations and evaluated the accuracy of the system. Some of the 25 chicken shapes can be found in Fig. 1 to Fig. 4.

The following is the testing result. The bounding box accuracy criteria values are within 5 mm in world coordinates. As the first trial of piled non-fixed object recognition, our result is promising, though as the pile becomes higher and more complex, the accuracy decreases to some extent.

Table 1 Chicken segmentation evaluation.

Testing Scenarios	Segmentation Accuracy
1 layer	93%
2 - 3 layers	90%
4 and above	86%

The known issues come from the following scenarios: 1) when two chickens have hardly any gap between them or they tightly border each other, 2) when the chicken has scattered small masses around its verge and they are connected with neighbors, and 3) when the pile becomes high and complex. They are the issues to be solved in future.

5. Conclusion and Future Work

In this paper, we have shown a 2D and 3D combined method for piled object segmentation. From the experiment it proves that this approach is effective to aid the robot bin picking for fried chickens. This approach is also easily extensible to other objects. Nevertheless, to make the system robust for any objects in any cluttered status or for any piling style, the method needs to be refined or even reframed, which is our next step in this domain of work.

References

1) Hinterstoisser, S. et al.: Multimodal Templates for Real-Time Detection of Texture-Less Objects in Heavily Cluttered Scenes, *Computer Vision (ICCV), 2011 IEEE International Conference*, pp. 858-865 (2011).
2) Hinterstoisser, S. et al.: Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes, *Computer Vision–ACCV 2012*, Springer Berlin Heidelberg, pp. 548-562 (2013).
3) Enjarini, B., Gräser, A.: Planar segmentation from depth images using gradient of depth feature, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2012).
4) Domae, Y.: Fast Graspability Evaluation on Single Depth Maps for Bin Picking with General Grippers, *Proc. of 2014 IEEE Int. Conf. on Robotics and Automation*, pp. 1197-2004 (2014).
5) Chen LC et al.: Object segmentation method using depth slicing and region growing algorithms, *International Conference on 3D Systems and Applications, Tokyo, Japan* (2010).
6) Felzenszwalb, Pedro F., Huttenlocher, Daniel P.: Distance Transforms of Sampled Functions, *Technical Report*, TR2004-1963, Cornell Computing and Information Science (2004).