The Visual Internet of Things System Based on Depth Camera

Xucong Zhang[1] , Xiaoyun Wang and Yingmin Jia

Abstract The Visual Internet of Things is an important part of information technology. It is proposed to strength the system with atomic visual label by taking visual camera as the sensor. Unfortunately, the traditional color camera is greatly influenced by the condition of illumination, and suffers from the low detection accuracy. To solve that problem, we build a new Visual Internet of Things with depth camera. The new system takes advantage of the illumination invariant of depth information and rich texture of color information to label the objects in the scene. We use Kinect as the sensor to get the color and depth information of the scene, modify the traditional computer vision technology for the combinatorial information to label target object, and return the result to user interface. We set up the hardware platform and the real application validates the robust and high precision of the system.

Keywords Vision Internet of Things • Kinect • depth camera • detection

1 Introduction

The Internet of Things (IoT) leads the trend of the next information technology and creates the communication between "things". Most of the Internet of things systems utilize the RFID or other non-contact wireless technology as their sensor and achieved successes in the past. However, the RFID label has to be attached on every object for recognition, which cannot be implemented in some situations. Be-

[1] Xucong Zhang(✉)

the Seventh Research Division and the Department of Systems and Control, Beijing University of Aeronautics and Astronautics

F907, Xinzhu Building, NO 37 Xueyuan Road, Haidian District, Beijing, China

e-mail:yongshengsiling@163.com

sides, the cost of RFID labels should under consideration when there are huge amounts of objects. The Visual Internet of Things (VIoT) is proposed to provide a visual method to access the object labels. With the help of visual cameras, the VIoT can get the object location via image information of the scene, and attach the visual label to the object, then return the label to the information network.

The color camera used by VIoT based on the passive light source, can be greatly influenced by the change of illumination condition, and may result in the serious performance decline. The average precision of the best object detection [1] based on the color image is still in a low stage according to the Pascal VOC challenge. So the object detection based on the color image can not match the requests of the real application in the VIoT.

To overcome the shortcoming of color image provided by current visual camera, we proposed to take depth camera as the sensor of VIoT. The depth camera can generate the depth image of the scene, in which every pixel indicates the distance between the point and camera. The depth camera offers a new dimension of the scene and will change the detection strategy profoundly.

The depth camera we used is the Microsoft Kinect sensor [2]. The Kinect obtain the depth information of the scene by the light code technology with active light source. Besides, the Kinect equipped with another color camera, so we can obtain the depth image and color image of the scene at the same time. In the process of building the new VIoT, we restrict our attention to the monocular case, where a serial image processing is to analyze the image and detect the target from the depth and color image. Once we get the location of the target, the visual labels will be attached on it and return the information to the system, thus we achieve the function of integrate VIoT.

2 System architecture of VIoT

According to the mainstream VIT, our VIoT include three parts: perception, information processing and application. The Fig.2.1 shows the whole architecture of our VIoT, and the following is detailed introduction.

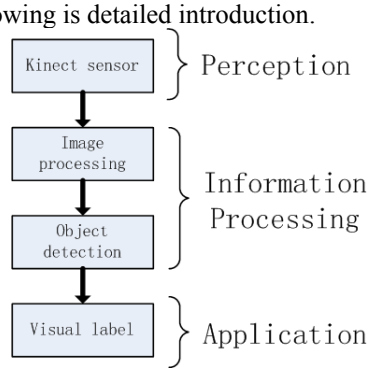

Fig. 2.1. System architecture of VIoT

3 Perception

The perception part of traditional VIT consists of RFID or other wireless sensor, while we use the Kinect as the sensor of our novel VIoT. The Kinect is illustrated in Fig.3.1.

Fig. 3.1. Kinect

The Kinect sensor can real-time get the color image and the depth image of the scene at the same time. Each image is VGA (480*640), and the color image is three channels while the depth image is only one channel. Every pixel of the depth image indicates the distance between the point of scene and camera with millimeter precision.

First, we make the calibration between color and depth image, because the initial data provided by Kinect is not calibrated. Second, we find that the initial depth data provided by Kinect is very noisy and incomplete. Noisy refers to variations between 2 to 4 different discrete depth levels. And the incomplete data which means the 0 value points of the depth data, will appears on the specular surface, edge of object, black region etc. We smooth the depth data by taking the mean over 9 depth frames for the noisy region. For incomplete regions we use a modified median filter: for every zero value point, we collect the depth value with a 5x5 pixel window and calculate the median of non-zero value as the new value of this point. After the two steps, we get the depth image ready for the next processing. On the other hand, the color image just smoothed by a single general Gaussian filtration. The Fig.2.3 shows the color picture and depth picture after the steps above.

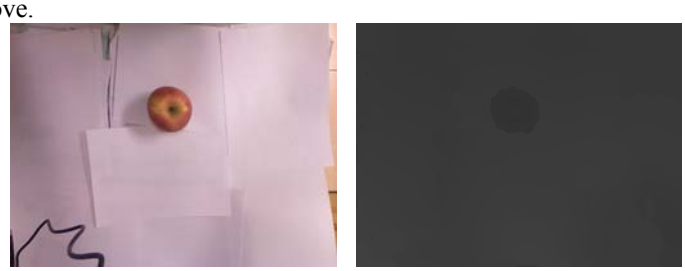

Fig. 3.2. Color image and depth image

4 Information Processing

The procedure of information processing is illustrated in Fig.4.1.

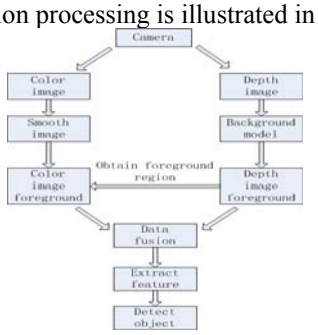

Fig. 4.1. Procedure of information processing

We separate the data from Kinect to be color image and depth image. The color image is just smoothed by the methods introduced before. For the depth image, the depth information is obtained by the active light source, so it is independent of illumination condition and shadow. We take full advantage of it to use the depth image to build the background subtraction model. For the color image and depth image is calibrated, the foreground region of them have the same location in the image, so we can get the foreground region of the color image by contrasting with depth image. We fuse the color and depth information carefully to achieve both the rich feature and illumination variant. Then the modified HOG feature is extracted on the fused data. On the last step of information processing, we will implement the object detection with a trained model, which learned from the labeled samples with support vector machine (SVM) [3]. In the following, we will introduce the important parts of the information processing.

4.1 Background subtraction model

The background subtraction model has been researched for decades, the recent stat-of-art methods are the Gaussian mixture model Maddalena [4], Zivkovic[5] and Barnich [6] according to the [7]. But all the methods applied for the color image will be great influenced by the illumination condition and shadow. The flash and other external interference will also be fatal for those methods. For the illumination invariant of the depth image, it will be the perfect source data for the background subtraction model. Although there are still much noise in the depth image from Kinect sensor, but that kind of noise will be eliminated easily. For the depth information of the image will changes only if there is really something moving in the scene, so we utilized the simply Gaussian model for the background, which can be described as a Gaussian function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (4.1)$$

where the x means the time sequence of pixel.

We compare the top background subtraction models with ours, and the performance is same but our model is faster and resource is economized.

When object moving in or out the scene, the background subtraction model will give the depth foreground region of the object, and we compare the region with color image to get the color foreground of the object. The procedure is illustrated in Fig.4.2.

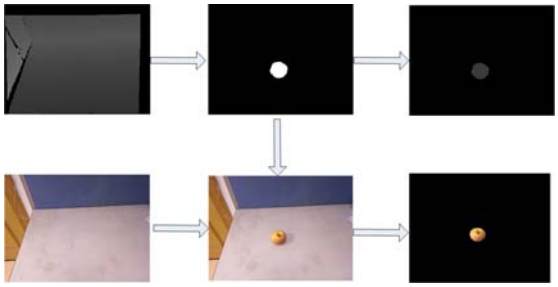

Fig. 4.2. Background subtraction

4.2 Data fusion and feature extraction

Since the pioneer of Kinect application, how to take advantage of the depth data had to be a hot topic again. The Microsoft researchers proposed "Depth image features" [8] for the Xbox application, which archived great success in the human pose recognition, but the method need large train dataset. The [9] evaluated the application HOG feature on depth images for recognition, and then extracted the better 3D point cloud descriptor viewpoint feature histogram (VFH) in the point cloud, further combined it with the DPM [10] based on color image. Another attempt is [11], which proposed a simple "Average" descriptor for the depth data.

In the above works, the descriptors for the depth data are based on the single depth channel or weak connection with color information. The traditional color descriptors can't explore all the potential of depth, independent depth feature is not the perfect way either.

In this paper, we combine the color information and depth information to be a fused image, which include four channels: three for the RGB of color image and one for the depth image. We choose the HOG [12] as the feature of the fusion image, for its geometrical and optical transformation invariance. The feature is ex-

tracted for each channel of every point and cascaded to a vector. Thus the feature consists of color feature and depth feature. For the scale invariance, we resize the fusion image for HOG extracting and cascade all features to be a long vector as the description of the point.

4.3 Object detection

We train the object model on the positive and negative samples, which can be collect off-line. Besides, our VIoT allow the on-line training, which means the user can select the object samples and random negative samples in the scene for training the model. We use the LibSVM [13] to train the model.

In detection, although the background subtraction model gives the foreground, we still should give the window of image to the model and get a confidence score when more than one objects in the scene. The traditional way to get the windows is sliding a fixed scale window in pyramid image. It is time-consuming. Inspired by [14], we utilize the image segmentation to get the independent parts of the foreground based on the fusion image. We build the frequency histogram for every channel of every part in foreground. If the difference between neighbor histograms is larger than threshold, the neighbor parts will merge to be one along with their frequency histogram. The segmentation result is illustrated in Fig.4.3.

Fig. 4.3. Image segmentation based on fusion image

With the help of image segmentation, we take the independent parts to object model and get confident scores, which determine whether the part is targeted.

5 Application

After obtaining the object location, the information communication and feedback become an easy task. We program the user interface with C++ to achieve function of VIoT, including operation of Kinect sensor, information processing, collecting samples in the scene, parameter setting and object detection. The main interface of the system is illustrated in Fig.5.1.

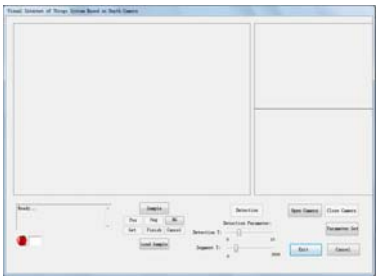

Fig. 5.1. Main interface of VIoT

The system will get the color image and depth image of the scene, and we can select the object we want to train, as show in Fig.5.2, where the left picture is the color image, the down-right picture is the depth image, and the top-right picture is the color foreground.

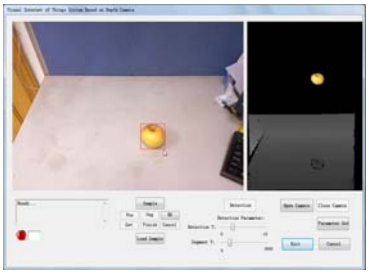

Fig. 5.2. Function of VIoT

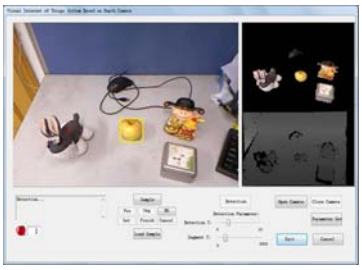

Fig. 5.3. Object detection

In the detection, the system will detect the target object in the scene and label it with yellow rectangle, as shown in the Fig.5.3. The number of target object is on the down-left of the interface.

6 Conclusion

In this paper, the depth camera is first applied in the Visual Internet of Things. The proposed VIoT can get the color and depth information of the scene, and take full advantage of them to get the target object location in the scene. Besides, the fusion of color and depth has been discussed and the feature extraction method has been modified for the novel fusion image. The real application validates the robust and high precision of the system.

7 Reference

1. Van de Sande, K.E.A. and Uijlings, J.R.R. and Gevers, T. and Smeulders, A.W.M.(2011) Segmentation as Selective Search for Object Recognition, Computer Vision (ICCV), 2011 IEEE International Conference on
2. Microsoft Corp. http://www.xbox.com/kinect.
3. J. A. K. Suykens(2001) Support vector machines: A nonlinear modelling and control perspective, Eur. J. Control 2001, 7, 311-327.
4. L. Maddalena and A. Petrosino.(2008) A self-organizing approach to background subtraction for visual surveillance applications. IEEE Transactions on Image Processing, 17(7):1168–1177
5. Z. Zivkovic and F. van der Heijden.(2006) Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters, 27:773–780
6. O. Barnich and M. Van Droogenbroeck.(2009) Vibe: A powerful random technique to estimate the background in video sequences. In IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pages 945 –948
7. Brutzer, S. , Hoferlin, B. and Heidemann, G.(2011) Evaluation of background subtraction techniques for video surveillance, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on 2011
8. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake.(2011) Real-time human pose recognition in parts from single depth images. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1297–1304. IEEE, 2011.
9. W. Susanto, M. Rohrbach, and B. Schiele. 3D object detection with multiple kinects. In Computer Vision–ECCV 2012. Workshops and Demonstrations, pages93–102. Springer, 2012.
10. P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(9):1627–1645, 2010.
11. L. Jourdheuil, Allezard.N, Chateau.T, and Chesnais.T. Heterogeneous adaboost with real-time constraints - application to the detection of pedestrians by stereovision. In VISAPP (1), pages 539–546. SciTePress, 2012.
12. Dalal, N. and Triggs, B.(2005) Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 1, 886-893
13. C.-C. Chang and C.-J. Lin. LIBSVM(2011) : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27
14. Felzenszwalb, P.F. and Huttenlocher, D.P.(2004) Efficient graph-based image segmentation, International Journal of Computer Vision, 2(59), 167-181