ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

RGB-D 모델을 이용한 강건한 객체 탐지 및 추적 방법☆

A Robust Object Detection and Tracking Method using RGB-D Model

박 서 희1 전 준 철1*
Seohee Park Junchul Chun

요 약

최근 지능형 CCTV는 빅 데이터, 인공지능 및 영상 분석과 같은 분야와 결합하여 다양한 이상 행위들을 탐지하고 보행자와 같은 객체의 전반적인 상황을 분석할 수 있으며, 이러한 지능형 영상 감시 기능에 대한 영상 분석 연구가 활발히 진행되고 있는 추세이다. 그러나 일반적으로 2차원 정보를 이용하는 CCTV 영상은 위상학적 정보 부족으로 인해 객체 오 인식과 같은 한계가 존재한다. 이러한 문제는 두 대의 카메라를 사용하여 생성된 객체의 깊이 정보를 영상에 추가함으로써 해결 할 수 있다. 본 논문에서는 가우시안 혼합 기법을 사용하여 배경 모델링을 수행하고, 모델링 된 배경에서 전경을 분할하여 움직이는 객체의 존재 여부를 탐지한다. RGB 정보 기반 분할 결과를 이용하여 깊이 정보 기반 분할을 수행하기 위해 두 대의 카메라를 사용하여 스테레오 기반 깊이 지도를 생성한다. RGB 기반으로 분할된 영역을 깊이 정보를 추출하기 위한 도메인으로 설정하고, 도메인 내부에서 깊이 기반 분할을 수행한다. 강건하게 분할된 객체의 중심점을 탐지하고 방향을 추적하기 위해 가장 기본적인 객체 추적 방법인 CAMShift 기법을 적용하여 객체의 움직임을 추적한다. 실험을 통하여 제안된 RGB-D 모델을 이용한 객체 탐지 및 추적 방법의 우수성을 입증하였다.

☞ 주제어 : 3차원 CCTV, 영상 감시, 객체 탐지, 객체 추적, 깊이지도, 이미지 분할

ABSTRACT

Recently, CCTV has been combined with areas such as big data, artificial intelligence, and image analysis to detect various abnormal behaviors and to detect and analyze the overall situation of objects such as people. Image analysis research for this intelligent video surveillance function is progressing actively. However, CCTV images using 2D information generally have limitations such as object misrecognition due to lack of topological information. This problem can be solved by adding the depth information of the object created by using two cameras to the image. In this paper, we perform background modeling using Mixture of Gaussian technique and detect whether there are moving objects by segmenting the foreground from the modeled background. In order to perform the depth information-based segmentation using the RGB information-based segmentation results, stereo-based depth maps are generated using two cameras. Next, the RGB-based segmented region is set as a domain for extracting depth information, and depth-based segmentation is performed within the domain. In order to detect the center point of a robustly segmented object and to track the direction, the movement of the object is tracked by applying the CAMShift technique, which is the most basic object tracking method. From the experiments, we prove the efficiency of the proposed object detection and tracking method using the RGB-D model.

☞ keyword : 3D CCTV, Video Surveillance, Object Detection, Object Tracking, Depth Map, Image Segmentation

1. Introduction

Recently, the basic CCTV system has been combined with various fields such as big data, image analysis, and artificial intelligence. It has evolved into an intelligent CCTV that can detect and analyze the overall situation of objects such as pedestrians[1]. A variety of image analysis researches has been conducted to recognize such situations as crime, fire, and anomalies using such intelligent CCTV. However, CCTV images using two-dimensional information generally lack the topological information. In order to solve the object misrecognition problem caused by the loss of information caused by projecting the 3D real world into 2D image, the object can be correctly recognized by combining the depth information. Therefore, this paper proposes an method to

1 Department of Computer Science, Kyonggi University, Gyeonggi-do, 443-760, Korea.
* Corresponding author (jcchun@kyonggi.ac.kr)
[Received 1 June 2017, Reviewed 12 June 2017(R2 10 July 10), Accepted 18 July 2017]
☆ This work was supported by kyonggi University's Graduate Research Assistantship 2017.
☆ A preliminary version of this paper was presented at ICONI 2016 and was selected as an outstanding paper.

detect and track moving objects by combining depth information with RGB information in two CCTV environments.

This paper is composed as follows. In Section 2, we introduce various image analysis researches related to video surveillance systems such as intelligent CCTV. In Section 3 introduces the object detection and tracking algorithm proposed in this paper. In Section 4, we introduce the methods for robust detection of objects by segmenting them into RGB-based segmentation and depth-based segmentation. In Section 5, we describe how to detect the center point of a robustly segmented object and track it using CAMShift. In Section 6, we show that the results of object detection and tracking using RGB-D information proposed in this paper are better than those of object detection and tracking using only RGB information. Finally, In Section 7 concludes.

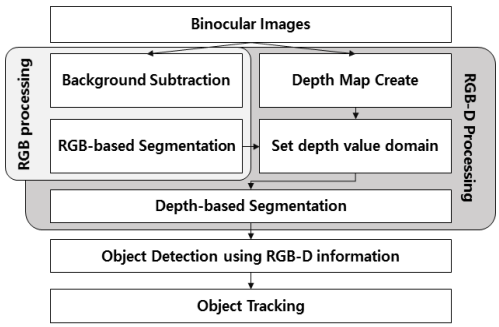
2. Related works

Due to the recent development of artificial intelligence technology, image analysis using intelligent CCTV is getting attention of the security industry. Monitoring of large crowds has increased in facilities such as sports stadiums and airports, and analysis of human activity has become important as the desire to process and analyze crowd scenes grows. In analyzing and processing these human activities, objects must be detected, extracted, recognized and tracked in order to obtain accurate motion information (position, speed, etc.)[2]. Also, the image must be analyzed by solving problems such as foreground extraction, object occlusion, crowd accurate aggregation, complex motion, and illumination change. For several years, various object detection methods have been proposed[3-5]. However, since it processes images based on 2D images, it has limitations in solving the problem of object occlusion in detecting and tracking objects. Therefore, it is necessary to add robust depth information to the object occlusion phenomenon and obtain motion information by performing object detection and tracking based on RGB-D model.

A similar study is a human detection method based on RGB-D sensors using Microsoft's Kinect. It performs preprocessing based on the input depth image, detects the head part of the person, creates a 3D model, and extracts the person from the background[6]. However, since Kinect's skeleton model generation is limited to a maximum of 6 persons, and the depth detection range is limited to 4.5m, it can be detected only at the near distance. (Based on Kinect version 2) To overcome this limitation of the Kinect, two cameras can be used to detect objects by creating depth information. Therefore, this paper proposes object detection and tracking method using robust RGB-D model for object occlusion. By using the RGB-D model, it is possible to solve the occlusion problem of the object robustly, and to provide precise data for future human motion estimation and crowd behavior analysis.

3. Proposed Object Detection and Tracking Algorithm

In the intelligent video surveillance system, robust object detection and object tracking processed must be preceded in order to recognize the activity of the object. In this paper, we use two CCTV environments to obtain depth information. The background is modeled using the Mixture of Gaussians technique and the foreground is extracted from the modeled background. The depth map can be obtained by matching two CCTV images. The RGB-based segmented region is set as a domain for depth-based segmentation, and segmentation is performed within the domain. After a moving object is robust and precisely detected, the object must be tracked by calculating the center point of the object of interest. We use the CAMShift algorithm, which is the most basic method for



(Figure 1) A Robust Object Detection and Tracking Method using RGB-D Model

tracking objects, to calculate the center point of the object and track the moving object. The proposed algorithm is shown in Figure 1.

4. Object Detection

4.1 RGB-based Segmentation

In an intelligent video surveillance system, moving objects must be robustly detected in order to track the behavior of objects. Moving objects can be detected by extracting the changing area from the background image. In this paper, background image is extracted using Mixture of Gaussians technique which shows the color distribution of each pixel point. Since it includes gaussian density, objects can be extracted by adaptation to various image changes due to illumination change. It shows good performance in outdoor environment where there is illumination, light, noise, etc., and accurately shows intensity change of pixel. The following equation is the probability of a pixel with intensity I_t at time t.

$$P(I_t) = \sum_{i=1}^{K} \omega_{i,t} \cdot \eta(I_t, \mu_{i,t}, \sum_{i,t}) \quad (1)$$

The parameters are K is the number of distributions, $\omega_{i,t}$ is the weight of the I^{th} distributions at time t, and $\mu_{i,t}$ is the mean distribution. The standard deviation $\sum_{i,t}$ is a parameter defining the probability density function, which is the covariance matrix of the gaussian distribution. The gaussian probability density function shown in equation 2.

$$\eta(I_t, \mu_{i,t}, \sum_{i,t}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sum_{i,t}|^{\frac{1}{2}}} W \quad (2)$$

$$where,\ W = e^{\frac{1}{2}(I_t - \mu_{i,t})^T \sum_{i,t}^{-1} (I_t - \mu_{i,t})}$$

The weight ω/σ of each gaussian distribution is calculated to obtain a background pixel model. Then, each gaussian distribution is sorted in descending order by the weight, and if B gaussian distributions in K gaussian distributions are called background models, the following equation 3 is satisfied.

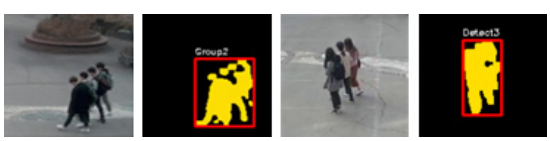
$$B = \arg min_b (\sum_{K=1}^{b} \omega_k > T) \quad (3)$$

In this paper, background modeling is performed by calculating the initial background model and updating it continuously. Then, the object is separated by performing a difference operation on the moving object in the modeled background. In order to extract only the object of interest, the residual image and white noise are removed using morphology operations and binarization through thresholds. Figure 2 shows the results of RGB-based segmentation using two-dimensional image information.

(Figure 2) This is the RGB-based segmentation result. (a) Input image (Left). (b) Mixture of Gaussians. (c) Morphology. (d) Binarization.

4.2 Depth-based Segmentation

The 3D real world is transformed into 2D image information through the camera, resulting in loss of information. Due to the loss of topological information, there is a problem of object misrecognition in which objects are recognized as one object.



(Figure 3) A problem of object misrecognition

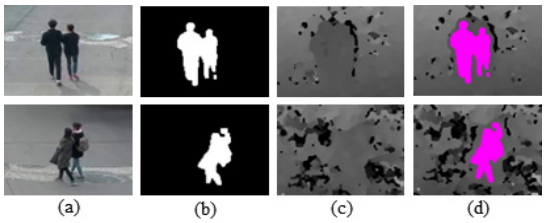
In this paper, depth information of object can be used to overcome the limitation of object recognition. In order to extract the depth information, depth map should be generated using CCTV two images. The depth map is a map showing the difference in 3D distance between the object in the image and the camera. Each pixel is represented by a one-dimensional value between 0 and 255. In order to obtain

depth information, depth map is generated using two connected CCTV images. The camera input image should be photographed horizontally to reduce the error in matching between images, which can be improved by adding a camera calibration step in the future.

The difference between the images input from the left and right cameras is called Disparity, which plays an important role in generating the depth map. In order to calculate the disparity for generating the depth map, the similarity of the left and right images is measured. The features are matched on a block-by-block basis using the measured similarity, and then the minimum matching cost is calculated[7]. Figure 4 shows the result of generating depth map by performing stereo-based matching.

(Figure 4) This is the result of creating a depth map. (a) Left image. (b) Right image. (c) Depth map.

In order to perform the segmentation based on the generated depth map, the result segmented by RGB is set as the domain region to extract the depth value.



(Figure 5) This is to set the domain for searching depth values. (a) Input image. (b) RGB-based segmentation. (c) Depth map. (d) Domain setting.

(Figure 6) This is the object detection result using RGB-D information. (a) Frame. (b) RGB-based segmentation. (c) Depth-based segmentation. (d) Object Detection.

The depth value is sequentially searched in the set domain and the previous pixel value is compared with the current pixel value. Even if the object is the same object, since the depth value of the moving object may be slightly different, object segmentation is performed according to a certain depth value range. In the following equation 4, the case of being included in a certain range value is regarded as the same region, and the case of not including the range is regarded as another region and the segmentation is performed.

$$-\alpha < Depth\ Value < +\alpha \tag{4}$$

However, even if it is a different object, if it is placed on a straight line having the same depth information value, the problem is segmented into the same object. This problem is solved by obtaining the coordinate information of an object and separating it into other labels when it goes out of a certain range. In this method, since the depth value is searched only in the RGB-based segmented region, the

amount of calculation can be reduced compared with the method of searching the depth value in the entire region. Also, since the disparity is clear when there is motion of the object, only the appropriate depth value that does not include noise can be used. The depth-based segmentation results are shown in the figure 6.

5. Object Tracking

Object tracking is an essential process of an intelligent video surveillance system that obtains information such as the moving direction of objects detected in consecutive frames, movement path, and whether objects are continuously appearing in the image. Therefore, the direction information of the object should be calculated based on the center point of the moving object in the continuous image. In this paper, motion of moving objects is tracked by applying CAMShift(Continuously Adaptive Mean Shift) algorithm that can be applied to CCTV environment in real time[8].

The CAMShift tracking method moves along the average value of the data and finds the peak or center of gravity of the distribution. The color histogram of the object to be tracked is compared with the histogram of the current input image to find the most similar window region. Setting the search window area is based on the location of the object that was previously tracked. In such a search window, the region and center point of the object to be tracked are calculated. The CAMShift algorithm uses a color-based probability model and updates the probability distribution for each frame adaptively as the state of the object changes. Therefore, even if the size of an object changes, it can be adapted and tracked continuously. The following is the result of applying the CAMShift method to the object detection results based on the RGB-D model proposed in this paper.

(Figure 7) The result of tracking with CAMShift using RGB-D model

6. Experimental Results

The following figure shows the comparison between object detection method using only RGB information and object detection method using RGB-D information proposed in this paper.

(Figure 8) This is the result of comparing object detection methods. (Top) Input image. (Middle) Object detection using only RGB information. (Bottom) Object detection using RGB-D information.

The following figure shows the result of comparing the object tracking method detected using only RGB information and the object tracking method detected using RGB-D information proposed in this paper.

(Figure 9) This is the result of comparing object tracking methods. (Top) Input image. (Middle) Object tracking using only RGB information. (Bottom) Object tracking using RGB-D information.

7. Conclusions

In this paper, robust object detection is performed by combining RGB-based segmentation results and depth-based segmentation using two CCTV cameras. We can obtain information about the region and center of the object by performing object tracking based on detected object. As a result, the problem of object misrecognition when recognizing an object using only RGB information is improved by adding depth information.

In future research, it is possible to classify objects by adding a process to identify whether objects are human or not, using a classifier based on the method presented in this paper[9]. In addition, the identified persons can be classified by body parts and a skeleton model can be created based on the classified information[10]. Point tracking is performed by setting a point of interest in the generated skeleton model, and the activity of the object can be predicted by learning based on the obtained data[11, 12]. The object detection and tracking method based on the RGB-D model proposed in this paper can be applied to expand the scope of the research into the field of human activity recognition.

Reference

[1] Grant, Jason M., and Patrick J. Flynn., "Crowd Scene Understanding from Video: A Survey," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Vol 13, No. 2, pp. 19, 2017. https://doi.org/10.1145/3052930

[2] Dixit, Astha, Manoj Verma, and Kailash Patidar., "Survey on Video Object Detection & Tracking," International Journal of Current Trends in Engineering & Technology, Vol 2, No. 2, pp. 264-268, 2016. http://ijctet.org/issuedetail.php?id=317

[3] Zivkovic, Zoran, "Improved adaptive Gaussian mixture model for background subtraction," Proceedings of the 17th International Conference on Pattern Recognition, pp. 28-31, 2004. https://doi.org/10.1109/icpr.2004.1333992

[4] Viola, Paul, and Michael Jones, "Rapid object detection using a boosted cascade of simple features," Computer Vision and Pattern Recognition (CVPR), Vol 1, pp. 511-518, 2001. https://doi.org/10.1109/cvpr.2001.990517

[5] Dalal, Navneet, and Bill Triggs, "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition (CVPR), Vol 1, pp. 886-893, 2005. https://doi.org/10.1109/cvpr.2005.177

[6] Xia, Lu, Chia-Chih Chen, and Jake K. Aggarwal, "Human detection using depth information by kinect," Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 15-22, 2011. https://doi.org/10.1109/cvpr.2005.177

[7] Hirschmuller, Heiko, "Stereo processing by semiglobal matching and mutual information," IEEE Transactions on pattern analysis and machine intelligence, Vol 30, No. 2, pp. 328-341, 2008. https://doi.org/10.1109/tpami.2007.1166

[8] Bradski, Gary R, "Computer vision face tracking for use in a perceptual user interface," Intel Technology Journal, pp. 214-219, 1998. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.7673

[9] Nguyen, Duc Thanh, Wanqing Li, and Philip O. Ogunbona, "Human detection from images and videos: A survey," Pattern Recognition, pp. 148-175, 2016. https://doi.org/10.1016/j.patcog.2015.08.027

[10] Shotton, Jamie, et al., "Real-time human pose recognition in parts from single depth images," Communications of the ACM, Vol 56, No. 1, pp. 116-124, 2013. https://doi.org/10.1145/2398356.2398381

[11] Wei, Shih-En, et al., "Convolutional pose machines," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724-4732, 2016. https://doi.org/10.1109/cvpr.2016.511

[12] Cao, Zhe, et al., "Realtime multi-person 2d pose estimation using part affinity fields," Computer Vision and Pattern Recognition (CVPR), 2017. https://arxiv.org/abs/1611.08050

◐ 저 자 소 개 ◑

박 서 희(Seohee Park)

2017 B.S. in Computer Science, Kyonggi University, Suwon, Korea
2017～Present : M.S. Student in Computer Science, Kyonggi University, Suwon, Korea
Research Interests : Computer Vision, Human Activity Recognition
E-mail : eehoeskrap@kgu.ac.kr

전 준 철(Junchul Chun)

1984 B.S. in Computer Science, Chung-Ang University, Seoul, Korea
1986 M.S. in Computer Science(Software Engineering), Chung-Ang University, Seoul, Korea
1992 M.S. in Computer Science and Engineering (Computer Graphics), The Univ. of Connecticut, USA
1995 Ph.D. in Computer Science and Engineering (Computer Graphics), The Univ. of Connecticut, USA
2001.02～2002.02 Visiting Scholar, Michigan State Univ. Pattern Recognition and Image Processing Lab.
2009.02～2010.02 Visiting Scholar, Univ. of Colorado, Wellness Innovation and Interaction Lab.
Research Interests : Augmented Reality, Computer Vision, Human Computer Interaction
E-mail : jcchun@kgu.ac.kr