An Automatic Camera Tracking System for Video Surveillance

[1]Sang Hwa Lee [2]Siddharth Sharma [3]Sanglin Lin [4]Jong-Il Park

[1,2]서울대학교 [3,4]한양대학교

[1]lsh529@snu.ac.kr

Abstract

This paper proposes an intelligent video surveillance system for human object tracking. The proposed system integrates the object extraction, human object recognition, face detection, and camera control. First, the object in the video signals is extracted using the background subtraction. Then, the object region is examined whether it is human or not. For this recognition, the region-based shape descriptor, angular radial transform (ART) in MPEG-7, is used to learn and train the shapes of human bodies. When it is decided that the object is human or something to be investigated, the face region is detected. Finally, the face or object region is tracked in the video, and the pan/tilt/zoom (PTZ) controllable camera tracks the moving object with the motion information of the object. This paper performs the simulation with the real CCTV cameras and their communication protocol. According to the experiments, the proposed system is able to track the moving object(human) automatically not only in the image domain but also in the real 3-D space. The proposed system reduces the human supervisors and improves the surveillance efficiency with the computer vision techniques.

1. Introduction

Video surveillance and security systems have been studied for a long time and become the vital parts of our everyday lives. As the society becomes more complex and unstable, the need of security systems increases. The surveillance and security systems usually consist of CCTV cameras and some IR sensors to detect or investigate the intruders. Multiple cameras are required to reduce the blind spots and occluded regions. Recently, intelligent surveillance and security systems are developed using pattern recognition and computer vision techniques, such as face detection and recognition, fingerprint and iris recognition, object tracking, gesture recognition, and so on [1, 2].

This paper deals with an intelligent video surveillance system using computer vision techniques. The goal of the proposed system is that the CCTV cameras track the human object automatically without supervised inspection. When an object in the video scene is detected, the proposed system first examines whether it is a human intruder or not. This reduces unnecessary alarms and supervised investigation. If the object is human or something suspicious, then motion estimation and object tracking algorithms are performed to track the moving object in the video scene. And the motion information in the image domain is converted to the camera parameters which define three camera movements, panning, tilting, and zooming in real 3-D space. Finally, the parameters are transferred to the camera via communication protocol, and the camera controls its movement according to the parameters. When the object moves into a blind region of the camera, the path of object is relayed to another camera in the neighborhood to track the object continuously. For this intelligent video surveillance system, we introduce computer vision techniques such as background subtraction, shape recognition, face detection, and motion estimation and tracking.

The rest of the paper is organized as follows. The overview of the proposed system is described in Section 2. The computer vision techniques used in the proposed system are explained in Section 3. The system integration and experimental results are shown in Section 4. Finally, this paper is concluded in Section 5.

2. System Overview

The goal of the proposed system is to track the moving object by controlling the camera's movement. Figure 1 shows the processing flow of proposed system. The proposed system integrates some computer vision techniques and CCTV camera hardware.

First, the object region is extracted using background subtraction techniques [3]. The object region is expressed as a binary image. Then, it is examined that the object is human or any other error since this paper focuses on the human object. There are many erroneous

alarms in video security systems because of abrupt illumination changes, wind, and animals, which increases the costs and unreliability of security systems. Thus, it is important to make sure that the object in the image is a human intruder before some investigation processes or alarms operate. This paper exploits the angular radial transform (ART), which is the region-based shape descriptor in MPEG-7 [4], to identify the human objects from the extracted object regions. The shapes of human bodies are learned using the ART coefficients in advance, and the object regions extracted by background subtraction are examined by the learning results.

After the object is determined as a human or something to be investigated, the proposed system tries to detect human faces in the image. Since human face is the crucial information in the surveillance systems, the human face region is zoomed up to a desirable scale. And, the motion estimation is performed to track the moving object in the video. The object or face region is tracked by the pan/tilt/zoom (PTZ) controlled camera and motion information. The motion information of object in the video is converted to the corresponding amount of PTZ parameters, which operate real panning, tilting, and zooming of camera. The parameters are packetized and transferred to the PTZ control module embedded on the CCTV camera via a communication protocol of commercial CCTV camera system.

The proposed system tracks the moving objects not only in the image domain but also real 3-D camera space. The object detection and motion estimation are the tracking processes in the video domain, and the automatic PTZ camera movement is another tracking process in the real 3-D space. Thus, a CCTV camera in the proposed system can investigate much wider area, and the surveillance efficiency is much improved. When the moving object disappears to the occlusion area where the camera can't watch the objects, the object tracking is relayed by a proper neighboring camera. The topological relation among the cameras is obtained by off-line camera calibration.

3. Automatic Camera Tracking System

The proposed system integrates several processes, object extraction, shape recognition, face detection, and motion tracking.

3.1 Object extraction

The object in the video scene is extracted by background subtraction (BS) technique. The BS techniques are popular and work well to extract the foreground objects in a single image [3].

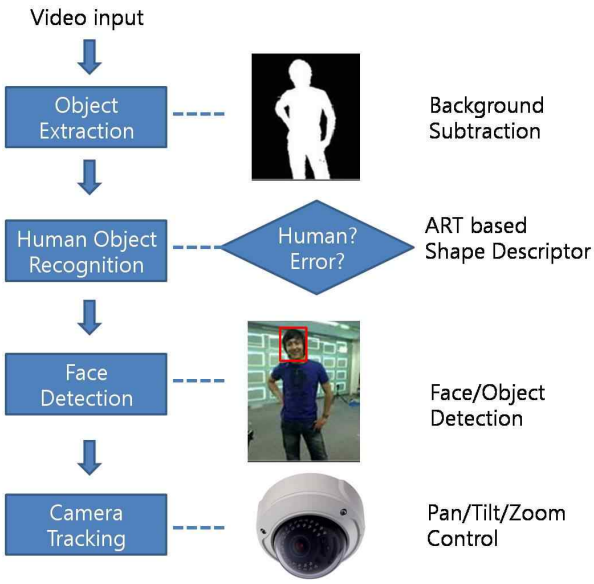

Fig. 1 Processing flow of proposed system.

In the paper, the colors of image are defined as RGB space. The pixel colors of background are modeled from multiple images without any objects. Using the multiple images reduces the noise and enables us to model the background statistically. For each pixel i, the mean $m_K(i)$ and variance $\sigma_K^2(i)\$$ of color components $(K = R, G, B)$ are calculated from multiple images. We also define the gradient values, $g_x(i)$ and $g_y(i)$, in x and y directions on the mean luminance values. Thus, the background for each pixel $B(i)$ is modeled as 8 parameters, 3 means and 3 variances of chromatic components, and 2 gradients in x and y directions on the mean luminance values.

To extract object regions, we need error models between background image and observed one. The object pixels are examined by the chromatic distance between the observed image and background one,

$$CD(i) = \left(\frac{I_R(i) - \lambda_i m_R(i)}{\sigma_R(i)}\right)^2 + \left(\frac{I_G(i) - \lambda_i m_G(i)}{\sigma_G(i)}\right)^2 + \left(\frac{I_B(i) - \lambda_i m_B(i)}{\sigma_B(i)}\right)^2. \quad (1)$$

In (1), I_R, I_G, I_B are the observed color components for object extraction. And the parameter λ_i is defined as below,

$$\lambda_i = \frac{\frac{I_R(i)m_R(i)}{\sigma_R(i)^2} + \frac{I_G(i)m_G(i)}{\sigma_G(i)^2} + \frac{I_B(i)m_B(i)}{\sigma_B(i)^2}}{\left(\frac{m_R(i)}{\sigma_R(i)}\right)^2 + \left(\frac{m_G(i)}{\sigma_G(i)}\right)^2 + \left(\frac{m_B(i)}{\sigma_B(i)}\right)^2}. \quad (2)$$

We have another distance measure using the gradient of luminance component,

$$GD(i) = (I_x(i) - g_x(i))^2 + (I_y(i) - g_y(i))^2 \quad (3)$$

where $I_x(i), I_y(i)$ are the gradients of luminance in the observed image. The total distance is defined with two distance models in (1) and (3),

$$D(i) = \gamma CD(i) + (1-\gamma) GD(i) \quad (4)$$

where γ is to adjust the relative weight between chromatic and gradient distances. The pixel is decided to be object pixel when its distance $D(i)$ is greater than a threshold. We finally apply morphological filters, dilation and erosion, to reduce some erroneous small regions and to fill the holes in the object.

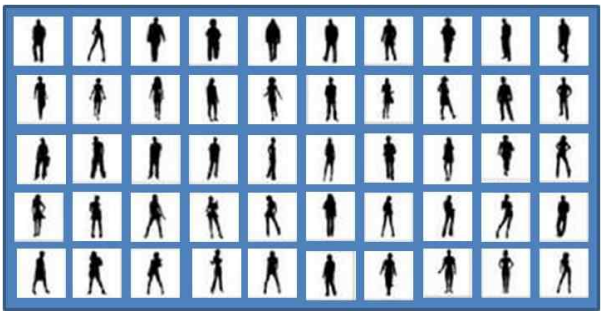

Fig. 2 Some database images of human shapes.

3.2 Human object recognition

The proposed system inspects that the detected object is human from the extracted region. Human object recognition is performed using the region-based shape descriptor, angular radial transform (ART). The extracted object is expressed as a binary image or region, so the region-based shape descriptor is suitable for recognizing the human shapes from the extracted regions. The ART is a complex orthogonal unitary transform defined in the polar coordinates system [4].

For recognition of human shapes, we construct database images of human bodies. Figure 2 shows some of database images of human shapes. The human shapes are collected on the standing or walking attitudes because the proposed system assumes the normally moving humans. Each shape image is normalized to 32x32 binary image. We approximately model the distribution of ART vectors using the mean vector and Euclidean distances. We estimate the mean of ART vectors of database images, and calculate the distribution of distances between the estimated mean vector of database A_m and ART vector of database images A_{DB}. Then, we calculate the mean and variance of the distances which define a threshold to determine the extracted object is human or not,

$$T_{human} = m_{ART} + \alpha \sigma_{ART}, \quad (5)$$

where m_{ART} and σ_{ART} are the mean and standard deviation of the distances. A parameter α reflects the faithfulness of database images. In other words, when the shapes of human bodies in database are widely distributed, α and T_{human} are increased since the distances between ART vectors in database become large. In this case, the false-positive recognition errors may be increased. On the other hand, α and T_{human} are decreased when the shapes of human bodies in database are locally distributed. In this case, the false-negative recognition errors may be increased. We set the threshold such that the recognition rate for database images is higher than 90%.

3.3 Face detection

Faces are crucial information in the video surveillance and security systems. Thus, it is important to obtain the clear face image of intruder in the system. When the extracted object is determined to be human, the proposed system tries to find the face in the video, and zooms up the detected face. If no face is detected in the human object region, the alarm and more investigation are performed by the supervisor. Face detection is the most classical pattern recognition problem in computer vision [5]. The face detection algorithm in the paper is based on the 2-D Haar patterns and Adaboosting method [6].

3.4 Camera tracking

The motion of face or object region is estimation, and converted to the parameters which control the amount of panning, tilting, and zooming in the CCTV camera. We calibrate the parameters and motion information from empirical observation in advance. The parameters are packetized into a protocol and transferred to a CCTV camera that is investigating the object. The protocol for PTZ camera is 7-byte format [8]. We can control the PTZ movements of CCTV camera through the protocol. The amount of panning, tilting, and zooming is determined by the motion information of the object region. The moving object is tracked not only in the image domain (motion tracking) but also in the 3-D space (camera tracking). Since the camera moves automatically as the object moves, a single camera can inspect wider area and the blind spots of camera are eliminated. When the object moves where the camera can't track, the neighboring camera is selected to relay the object tracking. For the relay of tracking, the topological relation of cameras is calibrated in advance.

4. Experimental Results

We had experiments with real CCTV cameras to test the proposed system. We used PC for computing the algorithms, and commercial analog CCTV cameras for camera tracking. The PTZ control signals are generated through PC's serial port (RS-232) and converted into CCTV protocol through RS-485 port. Then, the signals of RS-485 port are transferred to the CCTV camera. We evaluated each algorithm in the proposed system, and tested the integrated system.

Figure 3 shows some experimental results of object extraction and tracking. The objects in Fig. 3 are decided to be human, then, the object or face region is tracked by the PTZ control of CCTV camera. The object region is located at the center in the image when the object is moving, since the CCTV camera tracks the moving object. In Fig. 3, three left (or right) pictures are the contiguous frames where the object moves in the same direction. The amount of PTZ movements are adjusted by the motion information and empirical calibration. The calibration for motion information and PTZ movements should be performed whenever the camera is setup because the human motion information in the image domain changes with respect to 3-D environment. Consequently, the proposed system tracks the moving object in the real 3-D space, and improves the efficiency of surveillance by reducing malfunctions.

Further works are required in the future. First, the multi-mode modeling of human body's shapes are necessary to reduce the recognition errors. Second, the background subtraction technique should be improved for illumination changes and shadows. Finally, the proposed system integrates multiple cameras and networks.

5. Conclusion

This paper has proposed an intelligent video surveillance system using computer vision techniques. The proposed system enables the CCTV cameras to track the moving object automatically by integrating the object extraction, human object recognition, face detection, and pan/tilt/zoom (PTZ) camera control. This paper has set the experimental environment with the real CCTV cameras and their communication protocol. The experiments have shown that the proposed camera system tracks the moving human object automatically, and improves the surveillance efficiency. It is expected that the proposed system becomes a major system of video surveillance and security systems in the future.

Fig. 3 Tracking results. The moving object is located at the center in the camera image, since the camera tracks the moving object.

6. References

[1] M. Valera and S. Velastin, "Intelligent distributed surveillance systems: A review," *IEEE Proc. Vis. Image, Signal Proc.*, vol. 152, no. 2, pp. 192-204, April 2005.

[2] W. Hu, T. Tan, L. Wang, and S. Maybank. "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 334-352, Aug. 2004.

[3] Massimo Piccardi, "Background subtraction techniques: a review," *IEEE Int'l Conf. on Systems, Man and Cybernetics*, pp. 3099-3104, 2004.

[4] M. Bober, "MPEG-7 visual shape descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 6, June 2001.

[5] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. on PAMI*, vol. 24, no. 1, pp. 34-58, Jan. 2002.

[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. of IEEE CVPR*, pp. 511-518, 2001.

[7] R. E. Schapire, "The boosting approach to machine learning: an overview," *MSRI workshop on Nonlinear Estimation and Classification*, 2002.

[8] PELCO-D Protocol manual.