광범위한 지역 감시시스템에서의 행동기반 키프레임 검출 및 비디오 요약

Activity-based key-frame detection and video summarization in a wide-area surveillance system

권 혜 영* 이 경 미**
HyeYoung Kwon Kyoung-Mi Lee

요 약

본 논문에서는 광범위한 지역을 감시하기 위해 설치된 여러 대의 카메라로부터 획득된 비디오에 대해 행동을 기반으로 한 비디오 요약 시스템을 제안한다. 제안된 시스템은 시야가 겹쳐지지 않은 다수의 CCTV 카메라를 통해서 촬영한 비디오들을 30분 단위로 나누어 비디오 데이터베이스를 구축하여 시간별, 카메라별 비디오 검색이 가능하다. 또한 30분 단위로 나눈 비디오에서 키프레임을 추출하여 카메라별, 행동별로 비디오를 요약할 수 있도록 하였다. 행동 검출과 관련된 11가지(in, out, stay, left, right, forward, backward, left_forward, left_backward, right_forward, right_backward)에 대한 요약된 정보를 가지고 현재 사람의 행동이 어떤 영역에서 어떤 방향으로 움직이고 있는 지에 대한 정보를 행동별 비디오 요약을 통해 보여줌으로써 더 자세히 행동 추적된 결과를 볼 수 있다. 또한 카메라 3대에 대한 전체적인 키프레임에 대한 행동별 통계를 통해서 감시지역의 행동기반 이벤트들을 한 눈에 간단히 확인해 볼 수 있다.

Abstract

In this paper, we propose a video summarization system which is based on activity in video acquired by multiple non-overlapping cameras for wide-area surveillance. The proposed system separates persons by time-independent background removal and detects activities of the segmented persons by their motions. In this paper, we extract eleven activities based on whose direction the persons move to and consider a key-frame as a frame which contains a meaningful activity. The proposed system summarizes based on activity-based key-frames and controls an amount of summarization according to an amount of activities. Thus the system can summarize videos by camera, time, and activity.

☞ keyword : Surveillance system, summarization, retrieval, key-frame, tracking

1. Introduction

As interest in safety and protection of people and property as well as national and social safety is increasing in modern society, the importance and applications of security systems are increasing. However, most video surveillance systems require high costs and manpower because someone must visually monitor the videos coming from CCTV cameras. Therefore, we need a system to automatically summarize and search surveillance data for a given monitoring time without the involvement of any personnel.

Video retrieval and summarization systems are already being actively researched, and an recent interest is concentrating on the video retrieval and summarization in surveillance systems. Silva *et al.* proposed a system to summarize video frames

* 준 회 원 : 덕성여자대학교 일반대학원 전산정보통신학과
khy8355@duksung.ac.kr
** 종신회원 : 덕성여자대학교 컴퓨터공학부 교수
kmlee@duksung.ac.kr
[2008/01/25 투고 - 2008/02/04 심사 - 2008/04/25 심사완료]

acquired from multiple cameras installed in a ubiquitous environment [1]. Siebel and Maybank proposed a real-time system which analyzes persons' activities by person tracking and crowd monitoring [2]. Their system tracked the movement of people by installing detection sensors in the ground and extracted key-frames to retrieve and summarize video frames by time and location. Yamazawa *et al.* developed the OVISS system which detects "in," "out," "stay," and "move" activities from sensors installed in the ground in four zones, and displays them by time [3]. Gorodnichy developed the ACE system which extracts seven types of activities from video frames taken with one camera at night [4].

In this paper, we propose a video summarization system that extracts key-frames based on 11 types of activities using time-independent background removal, person detection, and activity detection. While the ACE system[4] extracted seven activities (left, right, closer, further away, on/off) during the night, the proposed system extracts 11 types of activities from day to night by three CCTVs with no overlapping views. With activity-based key-frames, we can confirm to which direction a person is moving, so we can detect key-frames according to the activities of people. Also, we can summarize videos using the detected activity-based key-frames. The proposed video summarization system can efficiently summarize and retrieve bulky video frames for the monitored area by setting rough or detailed retrieval according to a threshold values on the level of summarization.

2. Wide area surveillance system

The wide-area surveillance system is a surveillance system that consists of multiple CCTVs scattering a wide-area [5]. Each camera with no overlapping FOV(field of view) [8] can track multiple persons and exchange the tracked person's information with other cameras through a server. The person can be modeled to an object, a color, a moving direction, a central point and a bounding box, etc in frames of camera. This person model can deliver the information of the tracked person between the server and each camera. The server can register and store the person model in a person list. The person stored model can be updated whenever the person makes movements. Thus every camera in the surveillance system can access the person list to check whether the person was previously tracked by other cameras. Such using the person list make possible exchange of information from each camera.

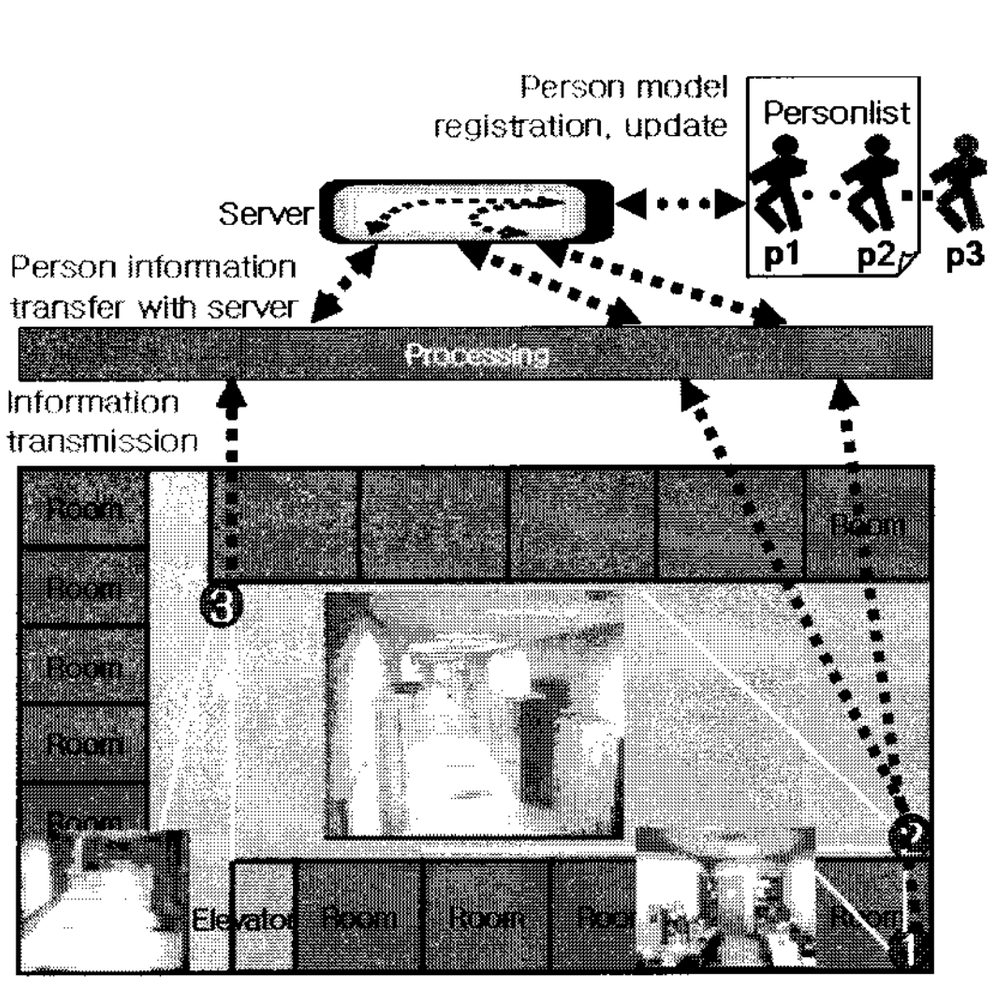


Fig. 1 An experimental surveillance area with three non-overlapping networked CCTV cameras

Fig. 1 shows a surveillance area tested in this paper. Three CCTV cameras are placed at the cordial in a building. They are viewed different areas. A person extracted by the first camera is

stored in a personlist on the server by taking account of personal data. When the person enters the view of the second camera, the camera checks if the person is registered in the personlist. If she is the one on the list, the second camera keeps on tracking the person. If not, it determines that she is a new person entering the network of cameras and registers a new human model on the personlist. This way each camera can check if a person entering its field of vision, or video frame, is one registered on the list or if she is a new person, or if the one who has been here and away is coming back. With these continued checks and tracking, the cameras that are distributed over a wide area can be used to make an end-to-end tracking of a person.

3. Activity-based key-frame detection

Traditionally, video is represented as a sequence of consecutive frames [6]. Video summarization is to extract a small but meaningful number of continuous characteristic frames, which is useful for various video applications [7]. A key-frame is the frame that contains important events, and thus is an important element in a video summarization system. In this paper, we define a key-frame as the frames which include a person with a large activity or movement.

3.1 Background removal

To detect frames having a person, the proposed system should remove background and segment foreground (or a person). Then the system selects key-frames if the segmented person has a large motion. Videos acquired at multiple video cameras scattering a wide area are constantly changed by light and illuminations according to time. Such illumination noises are easy to distort and lose information of a person and thus make difficult an exact segmentation of a person [4]. In this paper, we use time-independent illumination compensation before background removal. We define an intrinsic background image R which has no influence by light. An incoming current frame acquired from a fixed camera (Fig. 2(a)) can be predicted from previous frames, therefore, the illuminations existed in the current frame can be predicted from the previous frames. The calculated illuminations at previous frame compensates the current frame. All incoming frames are compensated as follows:

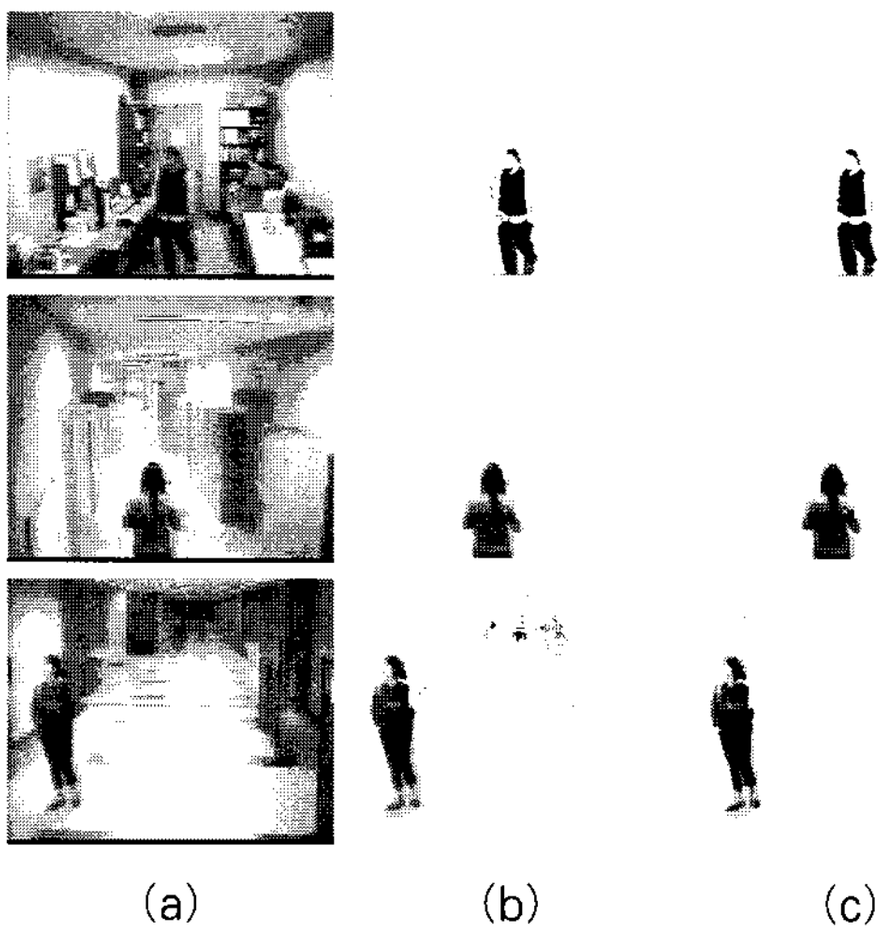
$$N^0 = average(R - C^0),$$
$$C^{1*} = C^1 + N^0 \ ,$$
$$N^{(t-1)} = average(R - C^{(t-1)}),$$
$$N^{(t-1)*} = \frac{N^{(t-1)} - (t-2)N^{(t-2)*}}{t-1},$$
$$C^{t*} = C^t + N^{(t-1)*},$$

where C^t is the frame at time t and C^{t*} is the frame compensated by using an intrinsic background image R. $N^{(t-1)}$ is the average of an entire difference between R and $C^{(t-1)}$ and $N^{(t-1)*}$ is the illumination computed from $C^{(t-1)}$.

After illumination correction, the proposed system conducts background subtraction. In this paper, we build the adaptive background model using the mean and standard deviation of the background [4]. Whenever a new frame arrives, a change in pixel intensity is computed using the Mahalanobis distance to classify background or foreground (moving persons). The computed distance is compared to a predefined threshold. If the distance is less than the threshold, the pixel is classified to

background and the adaptive background model is updated with the pixel.

Fig. 2 shows frames captured by camera 1, 2 and 3. Since the captured frames (Fig. 2(a)) include different illuminations depending on position of cameras, they should be corrected adaptively based on the illumination condition. Fig. 2(b) presents frames after illumination correction and background removal.



(a) (b) (c)

Fig. 2 Key-frame detection : (a) an original captured frame at camera 1(top row), 2(middle row) and 3(bottom row), (b) a foreground frame after compensating illuminations and removing background, and (c) a key-frame after removing unnecessary areas.

3.2 Person detection

Still the extracted key-frames can have noises or materials unlike a person. So we group pixels based on color using a connected-component algorithm in order to group segmented foreground pixels into a blob and locate the blob on a person. Then small blobs are merged into neighboring blobs that share similar colors to overcome over-segmentation generated by initial grouping.

Each blob contains information such as an area, average colors, locations of two vertices of the bounding box, the center point of the person. The grouped blobs are filtered out depending on the bounding box of the blob by removing noise-like blobs and non-person-like blobs using the following three methods: First, noises are removed when their area are equal to or smaller than a predefined threshold value. Second, noises are removed when the sides of their bounding boxes are small. The last, noises are removed when their values of elongatedness are small. Then, we can get clear noises in key-frames. Fig.2 (c) shows the persons detected from Fig. 2 (b) by removing unnecessary parts that are too small, too long, or not dense.

3.3 Activity detection

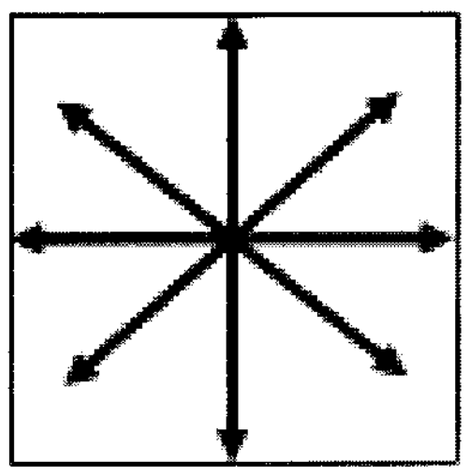


Fig.3 An activity of eight directions

To detect persons' activities or movements, this paper computes person's motions between consecutive frames to extract activities from key-frames. When we observe frames taken by CCTV, person's motions do not escape from eight directions to top, bottom, left, right, left forward, left backward, right forward and right backward as shown in Fig. 3. By adding entering into , exiting out, and staying in frames, we could define 11 types of activities.

Table I 11 types of activities

Activity	Explanation
In	A new person exists in a current frame, but the person does not exist in a previous frame.
Out	A person exists in a previous frame, but the person does not exist in a current frame.
Stay	A person exists in both a previous and a current frame and has a motion.
LEFT move	The person moves to left.
RIGHT move	The person moves to right.
FORWARD move	The person moves to forward.
BACKWARD move	The person moves to backward.
LEFT_FORWARD move	The person moves to left and forward.
LEFT_BACKWARD move	The person moves to left and backward.
RIGHT_FORWARD move	The person moves to right and forward.
RIGHT_BACKWARD move	The person moves to right and backward.

Table I describes 11 types of activities defined in this paper. The activity-based key-frames can be detected by computing a person's movement, that is, a difference between a previous and current frame after background removal. The activity "In" occurs when there is a new person in the current key-frame while there was none in the previous frame. The activity "Out" occurs when there was a person in the previous frame but he/she is not found in the current key-frame.

When a person exists both a current and previous frame and his/her motion is smaller than a given threshold value, it is called the activity "Stay". After background removal, the system detects all frames having persons even who stay in one place for a time. Therefore, to remove such duplicate frames with small motions, we need to detect only frames with considerable motions as key-frames. The motion of a person can be determined by differences with a previous frame. Among activities, we exclude the activity "Stay" for detecting activity-based key-frames if the motion is less than the threshold value.

When a motion is greater than the threshold value, it is called the activity "Move", which can be divided into eight directions as shown in Fig. 3. According to the moving direction of the person between the current and previous frame, we can obtain eight types of the activity "move". Fig. 4 shows the key-frames which contain activities.

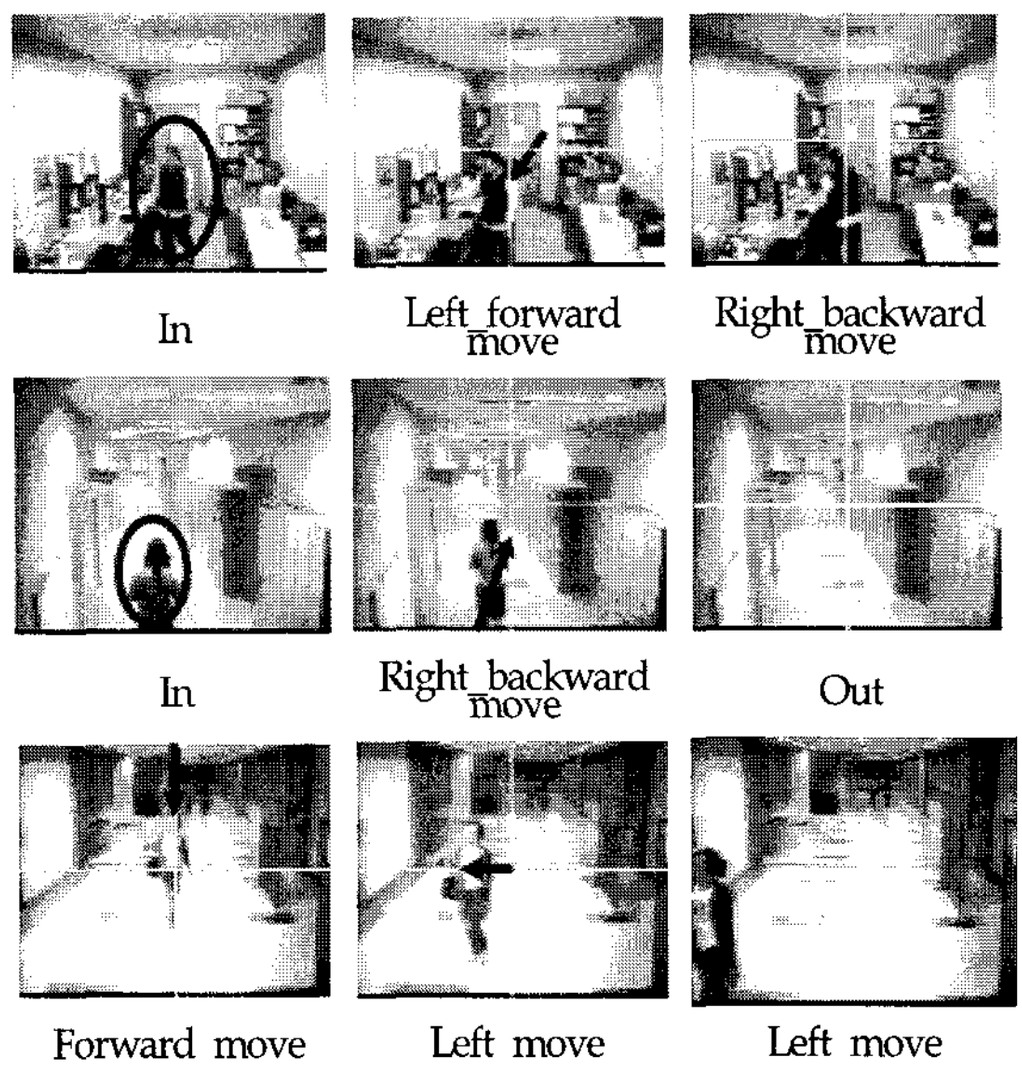


Fig. 4 Activity-based key-frame detection at camera 1(top row), 2(middle row) and 3(bottom row)

4. A key-frame-based video summarization system

The proposed activity-based key-frame detection and video summarization system for wide-area surveillance is implemented with JAVA(JMF) on the Microsoft Windows 2000 XP platform. The experiment was carried out on a computer using the

video frames (320×240) acquired from three UNIMO CCN-541 security cameras.

4.1 Database

The surveillance system recorded a video taken for 8 hours with three CCTVs and stored 30-minute units into the video database, so that the video units can be searched by date and time and by camera. The cameraDB stores the information on the three cameras used in this paper, and the key-frameDB stores the information of the activity-based key-frames which consist of camera no., date-time, key-frame no., the size of foreground after background removal, the size of movement obtained by the difference between a previous and current frame, a number of detected persons, areas, a bounding box, a center of person, and an activity.

4.2 Experimental results

The GUI of the proposed key-frame detection video search and summarization system is shown in Fig. 5. The videos can be searched by date-time (hour and minutes) and camera. The videos can be searched in a 30 minute unit. Also, we can check the information on the monitored area where three cameras are installed. We can see the installed location and views of the cameras as shown in Fig. 1.

Furthermore, the proposed system also provides a video summarization as shown in Fig. 6 by listing the key frames of the searched surveillance video in the order of time. The video summary in the proposed system is displayed on the screen by camera, by person, and by activity. When more than one camera is selected, the camera summary sequentially shows the key frame summary of the photographed video images from each camera. For

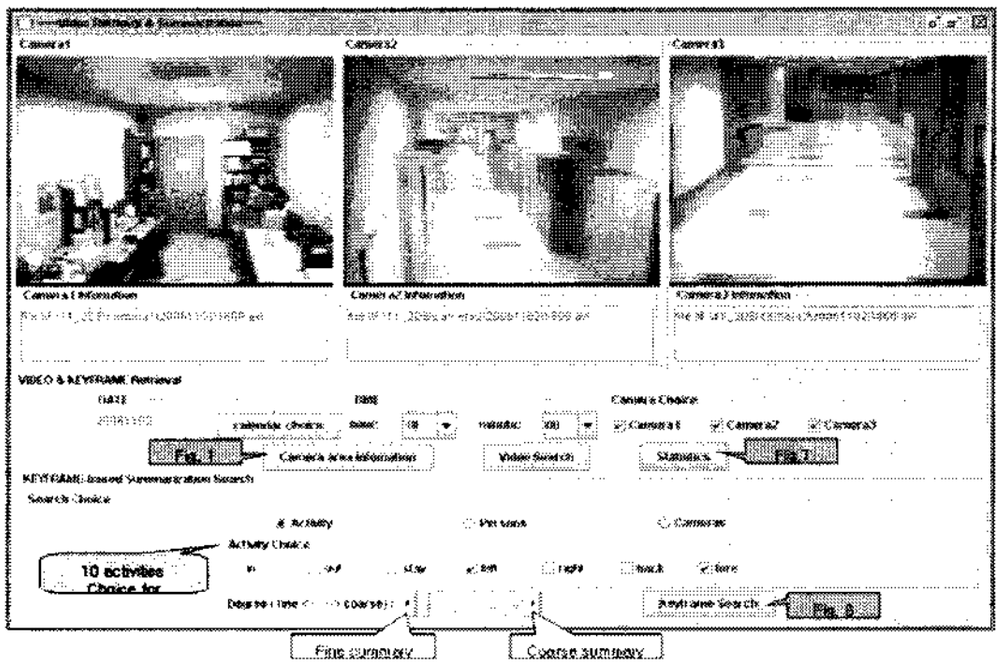


Fig. 5 GUI of the proposed video summarization system

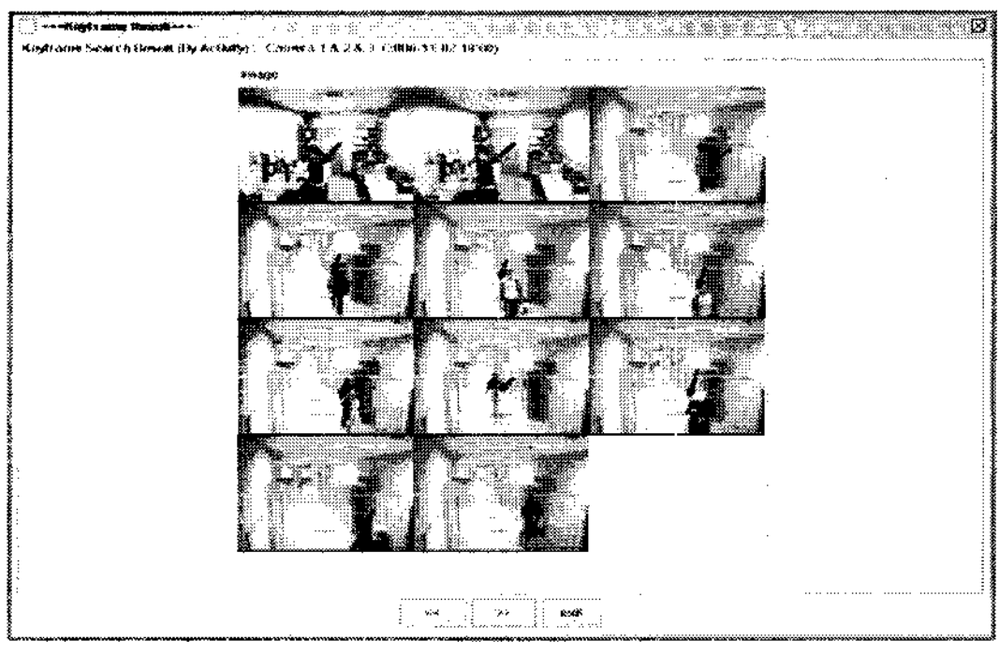


Fig. 6 Results of activity-based key-frame detection and summarization: left, forward

example, when you select cameras 1, 2 and 3, the video summaries of the camera 1, 2 and 3 are displayed in this order. Further, search by activity shows the key frame summary results for 10 types of activity. When only one of in, out, stay, left, right, forward, and backward conditions has been selected, the seven types of activities from all cameras are displayed regardless of the camera. When you select left and then forward or backward, left forward or left backward activities are displayed. When you select right and then forward or backward, the right forward or right backward activities are displayed.

The size of a video summary from a surveillance

system that has many cameras is still very large. Therefore, this system allows user to autonomously adjust the summary content provided by the system. By adjusting the scroll bar at the bottom of Fig. 5, we can set the threshold value by camera, by person and by activity for the size of movement determined by the size of foreground image from which the background image has been removed and the difference between previous and current frame.

Furthermore, this proposed system provides the overall statistics by activity of the people appearing in the camera so as to allow user to view the content of a 30 unit video at one glance. Fig. 7 shows the overall statistics by activity-based key-frame for three cameras. Camera 1 has a smaller number of activities than other two cameras, and Camera 2 shows the greatest number of activities among the three cameras. We can see that Camera 3 has the largest number of activities after 20 minutes. The key frames from which activities have been detected by three cameras during 30 minutes are displayed.

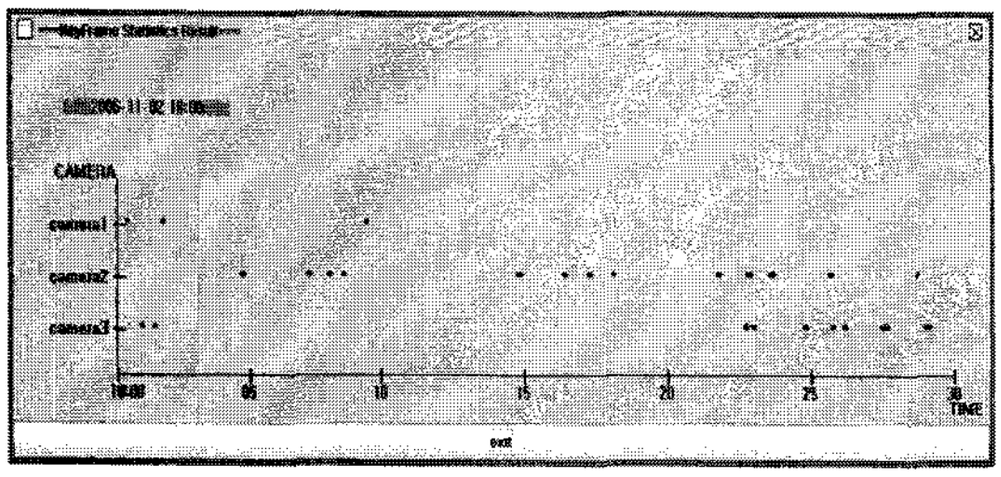


Fig. 7 Video statistics of three cameras

To evaluate the proposed system, we can compute the accuracy of the activity detection from the extracted key frames as followings:

$$\frac{\textit{Sum of correctly detected activities}}{\textit{Sum of detected activities}} \quad (3)$$

Using Eq. (3), we achieved the accuracy of 92.86%. Table II is the confusion matrix for the activity extracted from three cameras by the proposed system. Because the background removal also removed the foreground (person) in image processing, the activities "left_forward" and "left_backward" were detected as the activities "in" and "out". Furthermore, the activity "forward" was detected the activity "backward" or vice versa, because a person moves forward or backward in real but the center of the person moves to backward or forward. For examples, when the person walks by shaking or moving hands, arms and legs to the front and back of the upper body, the computed center could not represent the person and the moving direction of the center could not be considered as the moving direction of the person. Fig. 8 shows the cases of detection of forward condition as backward condition, and vice versa.

The right part of Fig. 9 lists the information on the activities detected by the proposed system, in the order of the key frame number, the size of the foreground image from which the background has been removed, the size of activity determined by the difference between previous and current frames, the number of subjects, the size of person, labeling count, the maximum and minimum points of bounding box, the center point of human body, and the detected activities. These extracted key frames comprise the information of the video summary system.

Table III shows comparisons between the proposed system and other systems presented by other papers. Compared with the OVISS system [2] and the ACE system [3], the proposed system extracts activities from various places using non-overlapping multiple cameras to effectively

acquired from multiple cameras installed in a ubiquitous environment [1]. Siebel and Maybank proposed a real-time system which analyzes persons' activities by person tracking and crowd monitoring [2]. Their system tracked the movement of people by installing detection sensors in the ground and extracted key-frames to retrieve and summarize video frames by time and location. Yamazawa *et al.* developed the OVISS system which detects "in," "out," "stay," and "move" activities from sensors installed in the ground in four zones, and displays them by time [3]. Gorodnichy developed the ACE system which extracts seven types of activities from video frames taken with one camera at night [4].

In this paper, we propose a video summarization system that extracts key-frames based on 11 types of activities using time-independent background removal, person detection, and activity detection. While the ACE system[4] extracted seven activities (left, right, closer, further away, on/off) during the night, the proposed system extracts 11 types of activities from day to night by three CCTVs with no overlapping views. With activity-based key-frames, we can confirm to which direction a person is moving, so we can detect key-frames according to the activities of people. Also, we can summarize videos using the detected activity-based key-frames. The proposed video summarization system can efficiently summarize and retrieve bulky video frames for the monitored area by setting rough or detailed retrieval according to a threshold values on the level of summarization.

2. Wide area surveillance system

The wide-area surveillance system is a surveillance system that consists of multiple CCTVs scattering a wide-area [5]. Each camera with no overlapping FOV(field of view) [8] can track multiple persons and exchange the tracked person's information with other cameras through a server. The person can be modeled to an object, a color, a moving direction, a central point and a bounding box, etc in frames of camera. This person model can deliver the information of the tracked person between the server and each camera. The server can register and store the person model in a person list. The person stored model can be updated whenever the person makes movements. Thus every camera in the surveillance system can access the person list to check whether the person was previously tracked by other cameras. Such using the person list make possible exchange of information from each camera.

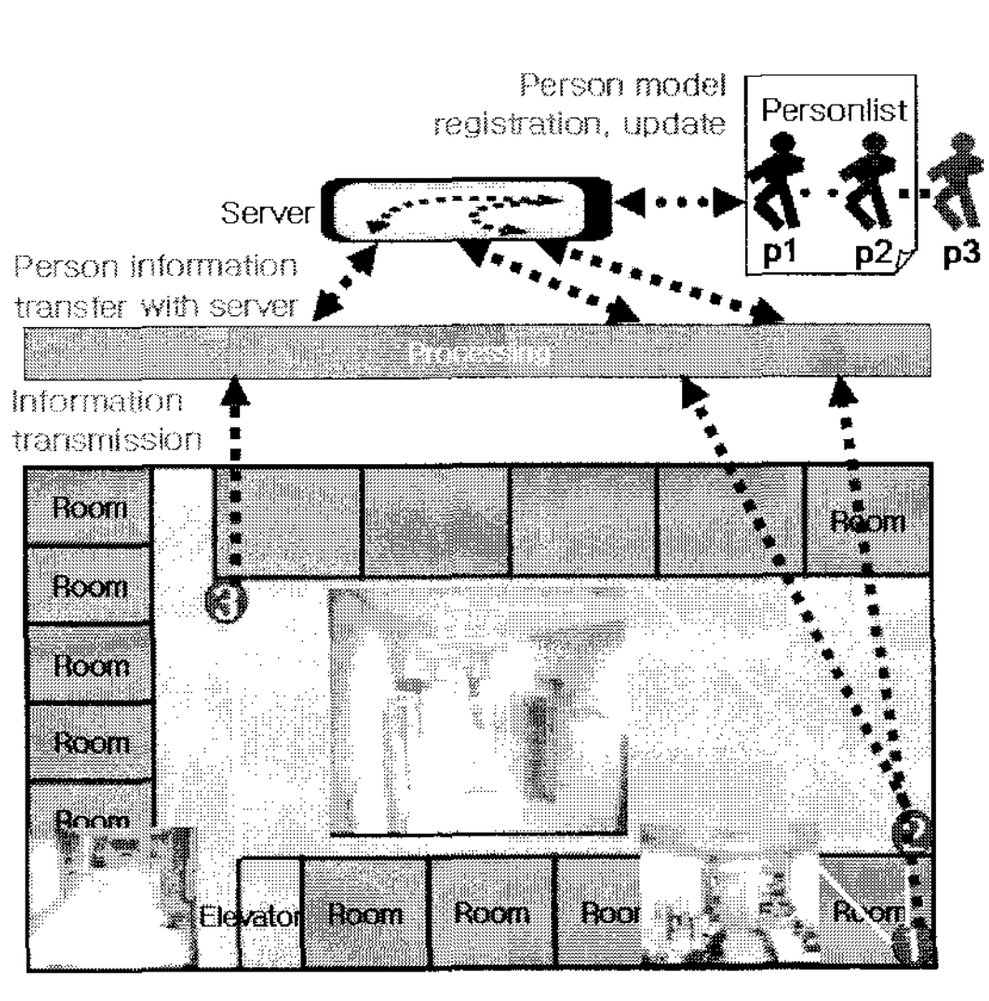


Fig. 1 An experimental surveillance area with three non-overlapping networked CCTV cameras

Fig. 1 shows a surveillance area tested in this paper. Three CCTV cameras are placed at the cordial in a building. They are viewed different areas. A person extracted by the first camera is

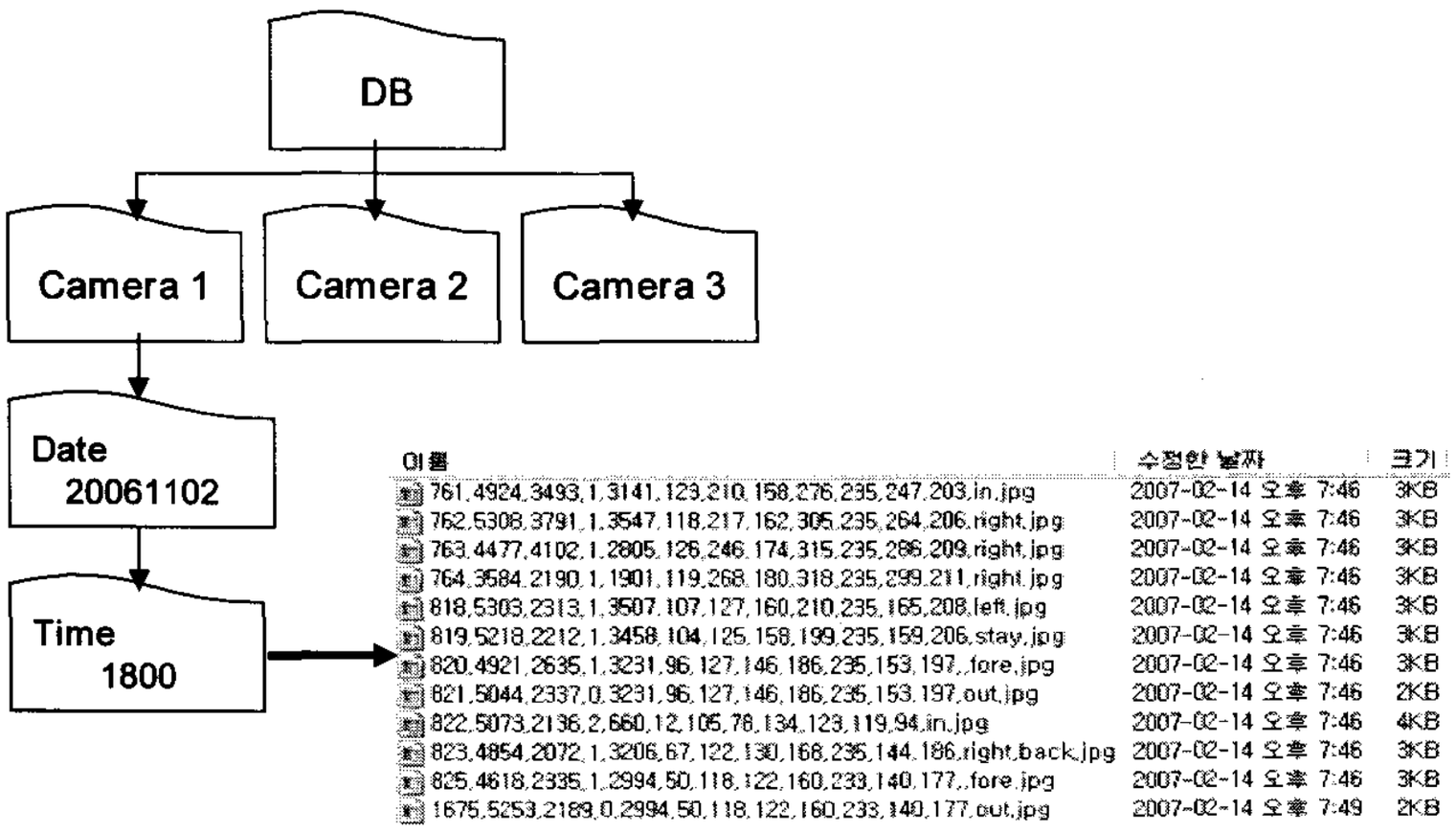


Fig. 9 Text-based annotations, encoded in file names. In the file name key-frame, BGcount,prevcount,regs_num,area, label,minX,minY,maxX,maxY,center_x,center_y,total_activity.jpg, key-frame stands for a key-frame number in a 30-minute video unit, BGcount the total number of pixels whose difference with the background image is large, prevcount the total number of pixels whose difference with a previous frame is large, regs_num the number of regions, area the number of pixels of a person, and label a region number in a frame. Also, minX, minY, maxX and maxY mean locations of two vertices of the bounding box and center_x and center_y the center position of the person. Total_activity stands for the activity in the key-frame.

5. Conclusions

The proposed key-frame detection and video summarization in a wide-area surveillance system used an activity-based approach in order to solve the difficulties of monitoring with multiple cameras and to effectively summarize a video. Furthermore, the efficiency of search improved through key frame search by camera, time, and activity using threshold values which enabled autonomous summarization by the system.

A future study will be to exactly detect persons' activities using a person model which consists of a head, a torso, and legs, and to track the movement paths of people in surveillance areas. Because a human body can move freely, the results in Table II were caused by the movements of hands, arms, and legs. These problems could be improved by designing a person model of tracking human activity with the elements of a head, a torso and legs. A further development is to track people using the characteristic colors of each person to track the movement paths of multiple people.

References

[1] G.C. de Silva, T. Yamasaki and K. Aizawa, "Evaluation of video summarization for a large number of camera in ubiquitous home", Proceedings of ACM multimedia, pp.820-828, 2005.

[2] N.T. Siebel and S.J. Maybank, "The ADVISOR visual surveillance system", Proceedings of ICCV, 2004.

[3] N. Babaguchi, Y. Fujimoto, K. Yamazawa and N. Yokoya, "A system for visualization and summarization of omnidirectional surveillance video," Proceedings of the 8th international workshop on multimedia information system, pp.18-27, 2002.

[4] O. Gorodnichy, "ACE surveillance : the next generation surveillance for long-term monitoring and activity summarization", Proceedings of NRC-CNRC, NRC Publication Number: NRC 48493, 2006.

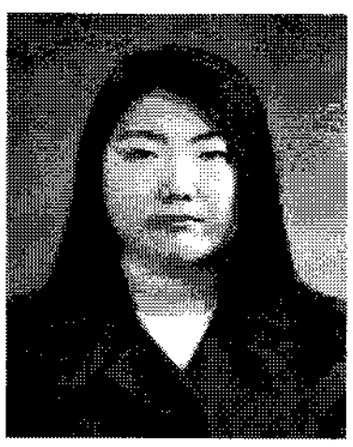
[5] 이윤미, 이경미, "모델기반 다중 사람추적과 다수의 비겹침 카메라를 결합한 감시시스템", 한국정보과학회: 컴퓨팅 실제, 제 12권, 제 4호, pp.241-253 2006.

[6] D.I. Kosmopoulos, A. Doulamis and N. Doulamis, "Gesture-based video summarization", Proceedings of ICIP, pp.11-14, 2005.

[7] N. Doulamis, A. Doulamis and K. Ntalianis, "An optimal interpolation-based scheme for video summarization," Proceedings of ICME, Vol.1, pp. 297-300, 2002.

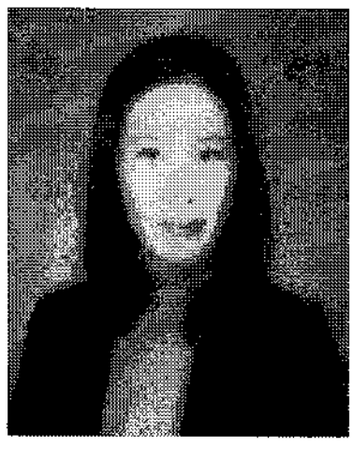
[8] N.T. Nguyen, S. Venkatesh, G. West and H.H. Bui, "Multiple camera coordination in a surveillance system", Acta Automatica Sinica, vol. 29, no. 3, pp. 408 - 422, 2003.

◐ 저 자 소 개 ◑



권 혜 영

2006년 2월 : 덕성여자대학교 인터넷정보학과 (공학사)
2006년 3월 ~ 현재 : 덕성여자대학교 일반대학원 전산정보통신학과
관심분야 : 영상처리, 감시시스템, 멀티미디어 데이터베이스
E-mail : khy8355@duksung.ac.kr



이 경 미

1993년 2월 : 덕성여자대학교 전산학과(이학사)
1996년 2월 : 연세대학교 전산과학과(이학석사)
2001년 12월 : 아이오와 주립대학교 전산학과(이학박사)
2003년 3월 ~ 현재 : 덕성여자대학교 컴퓨터공학부 교수
관심분야 : 내용기반 영상검색, 멀티미디어
E-mail : kmlee@duksung.ac.kr