Optimal View Selection and Event Retrieval in Multi-Camera Office Environment

Han-Saem Park, Soojung Lim, Jun-Ki Min and Sung-Bae Cho

Abstract—Recently, diverse sensor technologies have been advanced dramatically, so that people can use those sensors in many areas. Camera to capture the video data is one of the most useful sensors among them, and the use of camera with other sensors or the use of several cameras has been done to obtain more information. This paper deals with the multi-camera system, which uses the several cameras as sensors. Previous multi-camera systems have been used to track a moving object in a wide area. In this paper, we have set cameras to focus on the same place in an office so that system can provide diverse views on a single event. We have modeled office events, and modeled events can be recognized from annotated features. Finally, we have conducted the event recognition, view selection and event retrieval experiments based on a scenario in an office to show the usefulness of the proposed system.

I. INTRODUCTION

THERSE days, most people can obtain and use the multimedia data easily as the sensor technologies to obtain those data have been advanced dramatically [1]. In particular, camera comes to a representative and useful sensor device since video data is much more informative than other data types such as text document, sound, and still image in that it contains specific and realistic information even though its analysis is very challenging [2].

In a stage for public performance or in a field for sports broadcasting, a number of cameras are being used to cover a wide area and to display scenes from various angles. Accordingly, the cameras show different image and information, so it is required to analyze those images and select a specific one whose information is the most useful.

This paper deals with a problem that selects the camera of which view is the most informative in multi-camera system set in the office environment. In order to analyze information from input video, we have exploited Bayesian network, which is a model of a joint probability distribution over a set of random variables. Bayesian network is represented as a directed acyclic graph where nodes correspond to variables and arcs correspond to probabilistic dependencies between connected nodes [3]. These models are used to recognize and retrieve office events.

In previous studies, researchers have used the multi-camera system to track a certain object or a person in a wide area [4-6]. Multi-camera systems, however, have another advantage. That is, we can obtain diverse views using multi-camera system. In this paper, we mainly focus on this possibility.

II. BACKGROUNDS

A. Multi-Camera Systems

Previously, the multi-camera systems have been used to track a certain object in a wide area. Black and Ellis exploited multi-camera systems to track and detect moving object in outdoor environment [4]. These days, a few research groups have used multi-camera system in indoor environment. Sumi et al used multiple cameras with other ubiquitous sensors to capture simple interactions between humans in a conference room [5]. Silva et al presented a system for retrieval and summarization of multimedia data using multi-camera system and other sensors in a home-like ubiquitous environment [6].

All these researches used multi-camera system to cover wide areas, that is, they used only one camera at one place. As mentioned before, we focused on other possibility of the multi-camera system. We have set multi-camera system in office, so most of the cameras focused on the same place. Diverse views obtained from this system have two advantages: recognition of hidden object and higher recognition accuracy with ensemble. We can obtain information we need even if some cameras are hidden accidentally. Besides, higher recognition accuracy is expected due to accurate features extracted from multiple inputs.

B. Event Retrieval in Videos

Event retrieval in video data has been researched a lot according as the sensor technologies including cameras have been advanced [7-9].

In particular, one in sport videos is a very popular problem. Li et al. applied the event detection and modeling algorithms to different types of sports videos, and they provided retrieval and summarization of sport events [7]. Ekin et al. proposed a fully automatic and computationally efficient framework for soccer video summarization and analysis using some novel low-level processing algorithms and high level detection algorithms [8]. On the other way, Ersoy et al. presented a framework for the event retrieval in general video. They exploited domain-independent event primitives to provide
adaptability of the system [9]. All these works dealt with the video data captured by single camera.

III. OPTIMAL VIEW SELECTION AND EVENT RETRIEVAL USING BAYESIAN NETWORK BASED EVENT MODELING

Fig. 1 illustrates an overall process of event recognition, view selection, and event retrieval in the proposed multi-camera system. The whole process is divided into three parts: event recognition, view selection, and event retrieval. Low-level features have been annotated by human expert manually based on predefined domain knowledge. In event recognition part, the designed Bayesian network (BN) model recognizes office events with all features from all cameras. This model also recognizes event with features from each camera. View selection part selects the camera that provides the optimal view considering the recognition probability of given event and event priority.

A. Office Event Modeling

To model office events, we have defined eight events, and each event is related to proper objects or poses based on event definition. Events, objects and poses used in this paper are as follows.
- **Event**: Calling, cleaning, conversation, meeting, presentation, sleeping and work
- **Object**: Computer, phone, note, vacuum, users
- **Pose**: Stand, sit, rest

Basic features such as object, person's pose, position in office and person's direction have been annotated by expert based on predefined event, and these annotated features and events have been used to make Bayesian network event model learn.

We have modified this learned model. First, we removed the dependencies among evidences because they are not significant if evidences were set. If we did not find the evidence, it was checked as ‘no’, meaning there is no evidence. Subsequently, parameters were also modified because they were just based on the learning data so that they can recognize office events generally. Fig. 2 shows the designed Bayesian network model.

![Bayesian network structure for event modeling](image)

We have used three different models by person. They share the same structure, but the parameters are different. As modeling like this, we can recognize several events happening at the same time. Performance of models by each person is
shown in Table 1.

B. Optimal View Selection

Once an event at a certain time point is recognized by Bayesian network model, the system selects an optimal camera view at that time considering the probability and priority of a recognized event. Given an event $e_k$ at the segment $S_i$, an optimal view, $V_j$ is decided as following equation:

$$V_j = \arg \max_j f^{ij}(e_k) = V_j | \text{Max}(\text{prob}(e_k)) \wedge \text{Max}(\text{priority}(e_k))$$

where $\text{prob}(e)$ is the recognition probability of event $e$ and the $\text{priority}(e)$ is defined by an expert based on the domain knowledge. A view with the highest probability is selected as an optimal one if only one event is captured at one segment. If two or more events are captured at the same segment, views of events with the highest priorities are selected as candidates first, and then a view with the highest probability is selected among the candidates.

C. Event Retrieval

Based on the selected view information and the annotated event information in database, event retrieval is conducted as simply the system retrieves the events, which satisfies the user query.

IV. EXPERIMENTS

A. Experimental Environment

For the validation of the proposed system, we have made the experimental environment using 8 cameras in an office. Fig. 3 shows the location and coverage of cameras, and Fig. 4 shows an example captured images. We have used Sony network camera (SNC-P5), and video has been saved with the resolution of 320x240 and frame rate of 15fps using MPEG video format.

For learning of Bayesian network model, we have collected the learning data of three persons in the preset experimental environment. For the event recognition and view selection experiments, we have collected video data based on the designed scenario shown in Fig. 5. The scenario is based on the events happening during office hour—from 9 AM to 6 PM—on one day, and all events have the possibility that can be occurred in the office.

B. Event Recognition

We have conducted the event recognition experiment to

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 5. A scenario in an office

<table>
<thead>
<tr>
<th>Units for Magnetic Properties</th>
<th>Person A</th>
<th>Person B</th>
<th>Person C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>80.0</td>
<td>90.9</td>
<td>86.4</td>
</tr>
</tbody>
</table>

show the performance of Bayesian network event model.
Table 1 summarizes the accuracies by person. It shows high accuracy for each person, and the total accuracy is also high.

There are four incorrectly recognized events: three ‘presentation’ and one ‘meeting’. All ‘presentation’ events are recognized as ‘conversation’ because call cameras does not satisfy the definition of ‘presentation’. ‘Presentation’ requires three persons, computer and note, but most cameras do not detect all three persons and note is very difficult to be recognized correctly when three persons are close together.

C. View Selection

Fig. 6 demonstrates an example of view selection result. In this frame, a person is performing event ‘cleaning’. Our system selected a view in camera 7, and it is reasonable because it shows an event ‘cleaning’ clearly.

D. Application Implemented and Event Retrieval

We have implemented the application to provide events retrieval in office video. The screen shots are shown in Fig. 7 and Fig. 8. In Fig. 7, all views of eight cameras are displayed in normal mode. Fig. 8 shows the retrieved scene with the keywords Calling, Conversation, and Meeting. Event ‘Meeting’ is shown in this shot.
V. CONCLUSION AND FUTURE WORKS

This paper presented the novel multi-camera system that provided diverse views of each single office event and recognized those events using BN model. Our system also presents the retrieval of events with the application. As shown in experiments section, the proposed system performs acceptable results, and BN model recognizes office events with good accuracies.

There are some limitations that should be solved. Current system considers the recognition probability at a single time point or for a short time, but it may cause a significant problem because the system selects image sequence. Therefore, view selection process should consider the entire scenario. Also, feature annotation part should be replaced by fully automatic annotation.

Future work will focus on applying the proposed system to the video in other domains such as sport videos and comparing the performance of the proposed method with conventional video retrieval methods. The semantic analysis and summarization of videos in diverse domains also can be interesting topics.

REFERENCES