A personalized summarization of video life-logs from an indoor multi-camera system using a fuzzy rule-based system with domain knowledge

Han-Saem Park, Sung-Bae Cho*

Department of Computer Science, Yonsei University, 262 Seongsanno, Seodaemun-Gu, Seoul 120-749, Republic of Korea

Article info
Article history:
Received 6 August 2008
Received in revised form
14 January 2010
Accepted 28 April 2011
Recommended by: R. Baeza-Yates
Available online 10 May 2011

Keywords:
Personalized summarization
Multi-camera system
Life-logging
Fuzzy rule-based system
Domain knowledge

Abstract
Information summarization and retrieval are significant research topics associated with recent advancements in sensor devices, data compression and storage techniques, and high-speed internet. As a result of these advances, it is possible for people to collect huge life-logs. Video is one of the most important life information sources. This paper describes a method of summarizing video life-logs in an office environment with a multi-camera system. Previously, multi-camera systems have been used to track moving objects or to cover a wide area. This paper focuses on capturing diverse views of each office event using a multi-camera system with several cameras observing the same area. The summarization process includes camera view selection and event sequence summarization. View selection produces a single event sequence from multiple event sequences by selecting an optimal view at each time, for which domain knowledge based on the elements of the office environment and rules from questionnaire surveys have been used. Summarization creates a summary sequence from whole sequences by using a fuzzy rule-based system to approximate human decision making. The user-entered degrees of interest in objects, persons, and events are used for a personalized summarization. We confirmed experimentally that the proposed method provides promising results.

1. Introduction
The popularization of digital cameras and camcorders has led to significant amounts of information being created as advancement in data compression and storage techniques permit storage and sharing of video [1]. As a result, information summarization/retrieval has become a significant research topic [2]. Among the various types of information, video is potentially more useful and realistic than others such as text documents, audio data and still images in many situations though it depends on the problem domains. Therefore, video summarization/retrieval has been investigated for many areas [2-4].

Popularization of mobile devices and wireless internet has permitted the creation of life-logs [5]. Microsoft Research conducted the project MyLifeBits, which is a system for storing everyone’s digital information [6]. These can be used for diverse and useful services such as summarization, retrieval, and visualization. Cho et al. [7] presented an automatic summarization system, AniDiary, for personal mobile life-logs. AniDiary provided a cartoon-style summary using collected mobile life-logs. Aizawa [8] collected life-logs using a wearable life-logging system and provided retrieval service, and Silva et al. [9] captured life-logs with several cameras and microphones. Video capability is essential to most life-logging systems [6,8,9].
This paper deals with multi-camera systems in which the user is proactively deciding to generate personal content for a life-logging system. We collected video life-logs in an office environment with a multi-camera system, and attempted to create a video summary. The multi-camera system is normally used to track moving objects or to cover wide areas [10,11]. However, we exploited the multi-camera system to obtain diverse views of office events by focusing all cameras on the same area. The proposed method divides the target videos into event sequences, selects an optimal event sequence with view selection, and summarizes the event sequence by considering the user-entered degrees of interests. For view selection, we used the domain knowledge of the elements in the office and rules from questionnaires. For the summary, we used a fuzzy rule-based system to approximate the human decision making process [12]. The users input their degrees of interest in each event, person, and object so that the system could retrieve the target video.

The rest of this paper is organized as follows. Section 2 presents related works of life-logging and video summarization, the multi-camera system and the fuzzy rule-based system. Section 3 describes the multi-camera office environment. Section 4 describes the proposed summarization method. Section 5 describes the experiments and provides an evaluation of the results. Section 6 discusses application scenarios and limitations of the proposed system. Section 7 concludes the paper.

2. Background

2.1. Life-logging and video summarization

Recently, life-logs have been collected by various kinds of sensors [6–9]. Research on life-logging can be classified into collection, analysis, and service provision. Life-log information collected from diverse sensors and environments is analyzed by mining, modeling, or learning techniques. These collected and analyzed logs may be exploited to provide services such as summarization and retrieval.

Collecting information is an essential part of life-logging studies, and various sensors have been used. Microsoft Research’s MyLifeBits project collected all possible information including images, files, sound, applications used, call logs, and TV/radio programs [6]. Aizawa [8] utilized gyro and acceleration sensors, wearable cameras, GPS, microphones, and a laptop. In addition to the sensors described above, smart phones and application programs are important [7]. Analyses of collected logs can be conducted in many ways. MIT Media Lab analyzed mobile life-logs to construct social networks [13]. Cho et al. [7] modeled the inference modules to detect important events based on mobile life-logs using Bayesian networks. Service provisioning with life-logs includes summarization and retrieval. Many works targeted these problems [6–8]. Cho et al. [7] presented a cartoon-style summary as mentioned previously.

Summarization is essential to video life-logging systems, and video is very useful log information in this type of system. Here, introduction of studies on general video summarization follows, and summarization of video collected in multi-camera systems, which are our targets, will be presented in the next section.

There are various types of videos for summarization, and one possible criterion for classification is background characteristics. Some videos have dynamic backgrounds and others have static backgrounds. Examples of the former include broadcast videos such as sport, news, and movies and examples of the latter include footage from security/surveillance cameras. The targets of this paper correspond to the latter. Many research groups have investigated how to summarize videos with dynamic backgrounds. Li et al. [4] provided a movie skimming method using audiovisual tempo analysis and specific cinematic rules, Albanese et al. [14] introduced the priority curve algorithm to summarize sports videos, and Zhu et al. [2] proposed a hierarchical video summarization method with different granularities using semi-automatic content description ontology applied to medical and news videos. The summarization of videos with static backgrounds has been investigated only minimally, though some works deal with static videos as an example of general videos [15]. As mentioned before, we targeted multi-camera life-logging systems where the user proactively generates content.

Video with dynamic backgrounds has clear scenes and shots divided by background. Generally, the summarization process includes content analysis, structure parsing, and summarization [3]. The content analysis step maps low-level features to high-level semantic concepts. This can be conducted automatically using image processing and recognition techniques [16,17]. The structure parsing step divides the video into scenes and shots based on the result of content analysis, and the summarization step provides significant parts of videos using analyzed information from previous steps. Some works do not consider these steps and instead attempt to summarize videos using direct modeling of highlights (excitement) based on low-level audio and visual signals in sport videos [18].

In contrast, our target videos have static backgrounds and are usually created indoor. Many people try to remember special days such as birthdays and anniversaries by making videos. These videos do not have clear scenes because the backgrounds are static indoor environments, so they are divided by either the locations of the users [9] or activities occurring in that domain [19]. Locations and activities are detected in various ways. Silva et al. [9] set pressure sensors in the floor to detect people’s correct location. Sumi et al. [19] used fixed IR trackers and LED tags worn by people to capture events such as stay, coexist, gaze, attention, and facing based on predefined event definitions. Some works annotated events manually and provided a summary of that [20,21]. The AMI project provides an AMI meeting corpus, which consists of 100 h of meeting recordings. It also provides diverse manual annotation of the meeting including dialog acts, summaries, head/hand gesture, movement, and gaze direction [22]. Based on this corpus and annotation, many studies have been conducted [21].

Redundancy elimination is an option for the summarization of videos. In particular, it is important to detect redundancies in the rushes, which are raw material collections and to include manual editing effects and re-takes of
similar scenes [23]. The well-known international benchmarking activity, TRECVid [24], held a workshop for summarizing BBC rushes. Many research groups presented summarization and evaluation methods [25]. For example, Ren et al. [23] conducted histogram analysis to detect rainbow screens, used an SVM classifier to detect clipboard shots, and used a graph matching algorithm to estimate the sequential similarities among re-take instances.

2.2. Multi-camera system

Most conventional multi-camera systems utilize multiple cameras to cover wide areas and to track moving persons and objects. Black and Ellis [10] developed a multi-camera system that tracked and extracted moving objects in an outdoor environment, and they attempted to apply the system to an indoor environment. The POLYMNIA project [26] used multiple cameras to detect, track, and record a person and provided the edited videos of the person in indoor and outdoor environments. The videos can be used as souvenirs for visitors. Another main purpose of using the multi-camera system is security. Porikli and Divakaran [11] set several cameras in a building to track moving objects and summarized that information. Recently, studies of summarization/retrieval or studies of human activity analyses have been conducted. Sumi et al. [19] analyzed subjects using sensors including multiple cameras and provided a simple summarization, and Silva et al. [9] exploited the multi-camera system and created summary videos of persons living in a home-like environment using adaptive spatio-temporal key frame extraction.

Multi-camera systems can provide diverse views. Because a single view of a certain activity or event might not provide correct information, we can obtain correct information by using multiple views. In a previous work, we extracted event information using IR tags and four cameras, and provided a retrieval system that searched for the event of interest with a simple query [27]. We emphasized the advantage of diverse views from multi-camera systems for the first time. This paper also focuses on this possibility, and on top of that, we proposed the summarization method of video-type life-logs in indoor environment. It used the users’ degrees of interests to events, persons, and objects to provide the personalized summary of event sequences and exploited the domain knowledge for summarization and view selection in indoor multi-camera environment.

2.3. TSK fuzzy rule-based system

We used a fuzzy rule-based system to evaluate the event shots. Fuzzy system, a method to represent human knowledge including uncertainty and ambiguity, is a significant and representative application of the fuzzy set theory proposed by Zadeh [28]. The fuzzy system has been used to model the human decision making process [29]. Dorado et al. [12] used a fuzzy rule-based system to approximate human-like reasoning in video annotation problems, and Yamashita [30] proposed an effective support system for students making career choices using fuzzy reasoning and fuzzy modeling. Chang and Liu [31] applied the TSK fuzzy rule-based system to model the stock price prediction process.

There are several modeling methods for the implementation of a fuzzy system, and we used the TSK fuzzy model. The TSK fuzzy model was proposed by Takagi, Sugeno, and Kang, and it can avoid the time-consuming process of defuzzification in other models because it is based on rules where the consequent is not a linguistic variable but a function of the input variables [32]. A typical fuzzy rule in this model has the following form:

$$\text{IF } x_1 \text{ is } A_1 \text{ and } \ldots \text{ and } x_n \text{ is } A_n \text{ THEN } z = f(x_1, \ldots, x_n)$$

(1)

here, $A_i$ represents a fuzzy set and $x_i$ represents an input variable.

3. Multi-camera office environment

3.1. Setting up a multi-camera system in an office environment

To collect the office event sequence, eight cameras were set in an office, as shown in the left part of Fig. 1. We set all cameras to focus on the same area so that the system can provide diverse views of a single event. The right figure illustrates an example of the different views.

3.2. Annotation of events, persons and object

This paper considers the collected video log as an event sequence and provides a summary of that sequence. All basic information including events, persons, objects, and their positions have been annotated manually, and these works have been performed based on the following event definitions:

- Entry $(A)$, If stand $(A, \text{entrance-area})$ and face $(A, \text{in})$
- Leaving $(A)$, If stand $(A, \text{entrance-area})$ and face $(A, \text{out})$
- Calling $(A)$, If hold $(A, \text{phone})$ and speak $(A)$
- Vacuuming $(A)$, If hold $(A, \text{vacuum cleaner})$ and stand $(A, \text{center-area})$
- Eating $(A)$, If hold $(A, \text{food})$
- Resting $(A)$, If rest $(A, \text{corner-area})$ or stretch $(A, \text{corner-area})$
- Work $(A)$, If sit $(A, \text{corner-area})$ and $(\text{use } (A, \text{computer}) \text{ or hold } (A, \text{document}))$
- Printing $(A)$, If exist $(A, \text{printer-area})$ and hold $(A, \text{printout})$
- Conversation $(A, B)$, If exist $(A, x)$ and exist $(B, y)$ and close $(x, y)$ and $(\text{use } (A) \text{ or speak } (B))$
- Meeting $(A, B, C)$, If exist $(A, x)$ and exist $(B, y)$ and exist $(C, z)$ and close $(x, y, z)$ and $(\text{use } (A) \text{ or speak } (B) \text{ or speak } (C))$ and hold $(A, \text{document})$ or hold $(B, \text{document})$ or hold $(C, \text{document})$
- Seminar $(A, B)$, If stand $(A, \text{screen-area})$ and sit $(B, \text{center-area})$ and speak $(A)$

These 11 events are all possible in an office environment. We have identified these events based on related works dealing with human activities or events. Some works classify events based on objects such as a phone, table, chair, book, or keyboard [33,34], and some other works classify events based on persons and their activities.
Several previous video summary studies have performed annotation automatically or semi-automatically using image processing and pattern recognition techniques [2,16]. The current version of the system relies on manual annotation, and it will be gradually replaced by automatic annotation in future works.

4. The proposed method

Fig. 2 provides an overview of the proposed summary system. As mentioned before, the summary process is divided into a view selection module and a summarization module. View selection, which selects an optimal view for a single point in time, is performed using rules based on domain knowledge. Summarization is performed using the fuzzy rule-based system. During summarization, the fuzzy system evaluates each event in the event sequence, selecting as the important events with a high evaluation score.

4.1. User input

To summarize an office event sequence based on users’ personal interests, we attempted to use the degree of interest (DOI) input for each event, person, and object.
User interest is an important factor to interact with users. User inputs were integers between 0 (low interest) and 3 (high interest), and they can be used for view selection and the personalized summary.

We have segmented the event sequence into shots. The shot, a basic unit of video data, is defined as a set of consecutive frames with the same event, person, and object. The DOI value is calculated by shot, and the DOI value in shot $S_i$ is defined as follows:

$$DOI_{E_j,S_i} = \frac{\sum_{frame \leq S_i} DOI_{E_j}(f)}{SL(S_i)}$$  \hspace{1cm} (2)$$

$$DOI_{P_k,S_i} = \frac{\sum_{frame \leq S_i} DOI_{P_k}(f)}{SL(S_i)}$$  \hspace{1cm} (3)$$

$$DOI_{O_l,S_i} = \frac{\sum_{frame \leq S_i} DOI_{O_l}(f)}{SL(S_i)}$$  \hspace{1cm} (4)$$

Eqs. (2)–(4) show the DOI value for one shot ($S_i$), event ($E_j$), person ($P_k$), and object ($O_l$). After adding the DOI values for all frames within a shot, the values are divided by the shot length, $SL(S_i)$, to obtain the final DOI value.

4.2. Use of domain knowledge

Domain knowledge contains domain elements and domain rules for view selection. The former describes each element in the domain environment and their relationships, and the latter describes information required to design rules for view selection. In the current setting, the domain is in the office environment. Fig. 3 shows an office element description for “Place” and “Meeting”. The domain rule for the office includes information for camera view selection and view transition by event and user position, and information for determining shot duration is based on the event. Most of this information is based on the questionnaire surveys and their analyzed results. Ten university students were surveyed by questionnaire. We let them select the camera view, which best showed each event from a total of eight camera views for each event. After that, we generated the rules for view selection by event and user position based on their statistical frequency. Fig. 4 shows a rule description example for view selection, which means “If person A is in area 2 and event Working happens, view #2 should be selected.” The format is based on Owl Rule Language (ORL) [36].

4.3. View selection and event sequence summarization

View selection generates a single event sequence from multi-camera event sequences. It includes a selection of an optimal event at the same point in time and a selection of an optimal view among views showing the same event. Previous works exploited simple rules to select one sensor among many. Sumi et al. [19] selected a sensor that had a high priority based on predefined priority. To select an optimal camera view by considering all variables including user input, the proposed system used the domain knowledge from the questionnaire survey. Fig. 5 shows a pseudo-code explaining the view selection process.

Next, the system evaluates each event in a single event sequence generated by view selection in order to produce...
The TSK fuzzy system was utilized in this study. Fuzzy rules designed using domain knowledge, which reflect the relative significance of events, persons, and objects, are as follows. Using user DOI values as inputs, the system calculates the final score for each event.

Before fuzzy inference with these rules, the fuzzy membership function shown in Fig. 7 is used to fuzzify the user inputs and the DOI values so that they can be used as input variables to the fuzzy system. Fig. 7 depicts a fuzzy membership function for DOI_E, and functions for DOI_P and DOI_O are also trapezoidal. This type of membership function is very simple, but is widely used in practice [37].

Assuming that each score of eight rules in Fig. 6 is score\_i (i = 1 – 8), we can calculate the final score, FS, as the
weighted average of each output for eight rules, shown as follows:

$$FS = \frac{\sum_{i=1}^{8} h_i \cdot \text{score}_i}{\sum_{i=1}^{8} \text{score}_i}$$  \hspace{1cm} (5)

Here, $h_i$ is the match between the antecedent part of the $i$th rule and the current system inputs with a $t$-norm. That is, it weights the inputs to combine the results of the eight rules in Fig. 6. Each score is proportionate to the DOI values input by the user. The fuzzified value of the fuzzy membership function is shown in Fig. 7.

In the summarization step, important events are selected by rank based on the evaluated score. Here, shots of a single long event, which was split into several events due to events in-between, cannot be selected more than twice, and events with a low evaluated score are excluded from the summary. The duration of each event is determined based on the domain knowledge, and central frames are selected for the summary.

5. Experiments

5.1. Data collection

Experimental life-log data were collected in the office environment as presented in Section 3. We designed a realistic three-person scenario that could happen in an office in one day (from 9:00 a.m. to 6:00 p.m.). Fig. 8 illustrates this scenario. Here, EN, PR, CONVERS, CA, and LE stand for entry, printing, conversation, calling, and leaving, respectively.

For data collection, Sony network cameras (SNC-P5) were used, and video was saved at a resolution of $320 \times 240$ and frame rate of 15 fps using MPEG video format.

5.2. Result of view selection

Based on the scenario in Fig. 8, we made a single event sequence with view selection, and user inputs were assumed as in Tables 1 and 2. The DOI values of objects were assumed to be 0.

Table 3 summarizes the selected event sequence after view selection. Here, C# refers to the camera view number, and F# refers to the frame number. As shown in the table, mostly events related to person C and Vacuuming, Printing, Meeting, and Seminar were selected because they had high DOI values.

5.3. Result of summarization

Experiments were conducted using the event sequence in Table 3. This sequence contains 29 shots, and user
inputs are assumed as in Section 5.2. Figs. 9 and 10 show changes in the DOI values for events and persons. Fig. 11 provides a final score using the fuzzy rules. In Fig. 11, the shots with the highest scores are shot #6, shot #20, and shots #22–26. The common characteristics of these shots are as follows. They are related to events with high DOI_E values and person C, who has a high DOI_P value. Shot #22 is the only exception, as it does not have a high DOI_E value; however, it is still related to person C. In addition to the shots above, the shots, which are related to either events with high DOI_E value or those related to person C generally had high scores. Table 4 shows a summarized event sequence. S# refers to the shot number, and more than two views were selected for C# for events including movement. Score describes a fully evaluated score via the fuzzy rules. View selection in view transition was performed using domain knowledge as described in Section 4.2.

### 5.4. Evaluation of summarization

First, we compared the summarized result using the proposed method with the summaries of the three users. After inputting the DOI values, users selected 12 different shots from all 29 shots. We calculated the precision and recall values for each user with Eqs. (6) and (7)

\[
\text{Precision} = \frac{n(\text{shots summarized automatically}) \cap n(\text{shots selected by user})}{n(\text{shots summarized automatically})} \tag{6}
\]

\[
\text{Recall} = \frac{n(\text{shots summarized automatically}) \cap n(\text{shots selected by user})}{n(\text{shots selected by user})} \tag{7}
\]

Since some events occurred more than twice in the scenario, all users excluded them except one shot including one event. Assuming that shots with the same event and person are the same, the results were encouraging with average precision and recall values of 0.94. (They were 0.78 if we distinguished all shots by shot#.) Here, the number of shots selected by users and those selected by the proposed method were the same.

We conducted a subjective test to evaluate the performance of the method. Two questions were asked of ten university students. The questions are as follows:

- Q1. Does the summarized video show your daily routine concisely?
- Q2. Does the summarized video show each office event clearly?

Answers for these questions range from −3 to +3, where ‘−’ and ‘+’ signs mean ‘No’ and ‘Yes’, respectively. With this test, we compared three view transition methods. The first one (1) alters the view as a person moves, so
Fig. 9. Changes in DOIs value by shot number.

Fig. 10. Changes in DOIp value by shot number.

Fig. 11. Changes in finally evaluated score by shot number.
one event contains several views. The second one (2) alters the view based on the person and the main object for that event. In this case, views one through four constitute one event. The last one (3) does not transit the view, so one event is shown by a single view selected using domain knowledge. The first transition method is good for events such as Entry, Leaving, and Conversation where persons move a lot or events can be recognized easily without objects. The second one is good for events such as Calling, Vacuuming, and Printing where objects are required to recognize events. For these events, selecting views with their main objects is important because there are several view options at one time. The last method is good for an event like Resting where the view scarcely changes.

Figs. 12 and 13 summarize the results of the subjective test. Here, the x-axis represents the view transition method and the y-axis represents the score of the subjects. Fig. 12 illustrates the evaluation of Q1. Here, the score of the third method is lower than those of the other two methods. Though three methods of 1, 2, and 3 used the same summary, which was summarized using the proposed method, users considered them different because they used the different view transition methods. Fig. 13 contains the evaluation of Q2. Here, the second view transition method shows the highest score, and the third one shows the lowest score. Our overall evaluation of the proposed method is good because the average scores are positive.

6. Discussion

Summarization and view selection of indoor video has many useful applications. As described in the introduction, video volume has increased dramatically with the popularization and advancement of digital cameras. In the near future, we might have multi-camera systems in many homes and offices. Automatic summarization can be a substitute for a diary. A multi-camera system provides diverse views of each event, and users can select the camera views. We introduced two scenarios in which the multi-camera system could be useful as follows:

- **Home scenario:** S throws a birthday party for his eight-year-old son and invites his son's friends. S leaves the children for a few moments while he prepares food. S turns on his multi-camera summarization system to create a video of the party.
- **Office scenario:** J missed an important meeting because of a business trip. The company stored a video log of the meeting, which includes an automatically generated summary. J watched the summary video. The camera view changed according to the speaker, which helped J to understand the flow of the meeting. Thanks to the summary, J obtained important information about the meeting though it is a busy day.

Queries do not serve the same function as automatic summaries. In many cases, people search for events in order
to watch details of a segment, but they need a summary when they want to review the overall events quickly.

Use of multiple cameras for one target can offer an alternative perspective, and it can be exploited usefully in cases introduced before, but it also can be an overload if users prefer a simple and economic system.

7. Conclusions

This paper collected indoor video life-logs with a multi-camera system and proposed a system for video summarization and view selection. It generated a single event sequence from multiple event sequences from the multi-camera system and summarized the sequence using a TSK fuzzy rule-based system. The domain knowledge is based on questionnaire surveys and literature used for view selection. In order to provide personalized summarization, we measured users' degrees of interest. With experiments on view selection and summarization as well as using user-based evaluation, we confirmed that the summarized event sequence was promising. We also discussed how summarization and view selection can have useful applications at home and in the office.

Future works will focus on applying the proposed method to several types of videos in indoor environments. The design and implementation of an application with a user-friendly interface are required for easy use and effective presentation of the proposed method. Studying automatic annotation techniques is another area of interest.

Acknowledgments

This work was supported by the Ministry of Education, Science and Technology under NRF, Korea (2010-0018948) and the IT R&D program of MKE/KEIT (10033807).

References

[22] AMI corpus, (http://corpus.amiproject.org/).
[26] POLYMNIA, (http://polymnia.pc.unicart.it/).