A Practical Video Digest Generation System Designed for Nursery Schools

Yu Wang, Tomoya Ishikawa, Jien Kato, Kenichiro Ishii and Shigeki Yokoi
Nagoya University, Japan
{ywang, ishikawa}@mv.ss.is.nagoya-u.ac.jp, {jien, kishii, yokoi}@is.nagoya-u.ac.jp

Abstract

In this paper, we present a video digest generation system that designed for use in nursery schools. The system utilizes plural surveillance cameras that set in different nursery rooms, each has a corresponding RFID receiver to catch the signal send by RFID tags. With every kid having a RFID tag in the pocket all the day time, the system automatically generate one daily video digest for each kid. The digest is generated through two types of processing. One is RDIF log analysis, which quickly picks out videos that expect to have the target kids' appearance. The other is visual feature analysis, which recognizes events in raw video materials, select video segments for each event and construct the digest. The practical performance of our system is confirmed in both quantitative experiment and questionnaire survey.

1 Introduction

Nowadays, kids spend most of their time in nursery schools or kindergartens. This directly brings a requirement from their parents: they want to see how everyday goes with their children. Some nursery schools in Japan have introduced remote surveillance camera system [6]. Such system records videos of the daily life in nursery school, and allow parents to access the generated video data through the internet. A limitation for such kind of system is that in order to watch desired videos, the parents have to do manual search in the whole video materials. Since the plural surveillance cameras generate a mass quantity of raw video data every day, search in them is difficult.

Actually, for most parents, they only care about their own kid and want to know what happened with him/her during the whole day time. To meet their needs, in this paper, we proposed a video digest generation system that designed for use in nursery schools. Our system takes the video data from the surveillance cameras and the log file of RFID receivers as input, and automatically generates one daily video digest for each kid. The digest covers most activities the target kid has participated and could well reflect how the day goes with that kid during the whole day time.

2 Related Works

Recently, there has been increasing needs of methods which could efficiently handle the ever-growing amount of video data from various sources, such like television broadcasting, surveillance videos as well as personal recordings. In the context of video digest generating, there were already some prior works. M. Amano et al. proposed a video editing support system in [1]. They believe that the video digest should be generated under a set of special rules called "video grammar", and it is necessary to extract and index the metadata such as shot size or camera work to make the video grammar applicable. Though their system could help to generate quite comprehensible video digest, it were rely on severe manual operations.

In the other work [2], K. Miura et al. proposed a method to automatically abstract cooking videos. Their method uses motion-related features to extract video segments that have cooking motions and existence of foods. These segments are used to construct the cooking digest. Similar method has also been applied on sport video digest generation task. In [3], N. H. Bach proposed to combine motion and image features with a Hidden Markov Model, in order to select highlight scenes in baseball match video.

In our work, we also deal with the task of video digest generation. However, it is different comparing to previous works in three aspects: 1).the raw video material we are dealing with is of a very large quantity; 2). different with cooking digest and sports digest, it is hard to define what is a "good digest" for the kid's daily life in nursery school; 3). different with the cooking video or sports video which have some kind of rules, the video of daily life is unconstrained and without clear rules.

3 System Overview

Our proposed video digest generation system utilizes plural surveillance cameras which are set in different rooms of the nursery school. In the nursery school we are working on, there are seven cameras set in nursing rooms, resting room, garden and passageway respectively, as shown in Fig.1. Each camera has a corresponding RFID receiver, which catches the signal send by nearby RFID tags. We let every kid has a RFID tag in the pocket after entering the nursery school every morning, then in the whole daytime, these tags will continuously send out signals and the receivers could capture them and write into a log file.

Take the raw video material and RFID log as input, the objective of our video generation system is: 1) to fast and automatically generate one video digest for each kid, and 2) the resulted video digest should well reflect the daily life of that kid. We think such a video digest should cover most individual chief events the kid have participated during the day time, and propose to utilize a cascade of modules to generate it.

The processing flow of our system is shown in Fig.2. The first step is RFID log analysis, which we use to obtain the temporal location of the target kid. Such location information is then used to pick out video segments which have the target kid's appearance. In case of 7 cameras are used, the raw video material for generating the digest can be reduced to less than 1/7 with

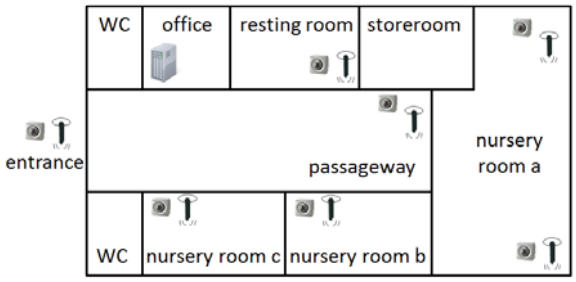


Figure 1. Hardware setting of the system.

such analysis. The remaining three steps are visual feature analysis, in which we assign a chief event label to each video segment, discover individual events in each type of chief event category, and generate the digest by connecting together a proper quantity of segments that contain the events from different individual chief events. In the following, we will introduce them in detail.

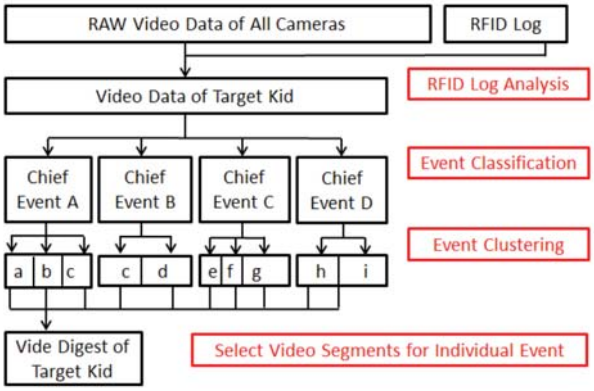


Figure 2. Flow of digest generation.

4 RFID analysis

The RFID log file has a set of records, each has a RFID tag id, a time stamp and a receiver ID. Every record could be considered as an evidence of the target kid appearing in a specify room at a given time. However, this kind of evidence is not reliable in our case because the receivers are set near to each other.

In Fig.3, we visualize a portion of practically generated RFID log. We can see that there exist disappearance and libation of the records, thus make it not able to be directly used. For efficient processing, we propose a weighted voting criterion to robustly estimate the temporal location for target kid.

We use one minute as a unit. For all the records R_n laying in each unit time interval T_i, it has an receiver ID denoted by $ID_n = k \in 1, 2...7$. We introduce a weight operator for each receiver i at time t denoted by x_{ik}, such operator reflects the priority of each receiver in different time period and could be assigned manually or learned from labelled data. By calculating the number of records which have the same receiver ID within the time interval, we get the votes for each receiver denoted by V_{ik}. We then select the k that maximizes $x_{ik}V_{ik}$, and take the video segment from the camera corresponding to the kth receiver if $w_{ik}V_{ik}$ excesses a predefined threshold. We get k for all the unit time interval and collect the corresponding video segments together. Such a set of segments is expected to have the appearance of the target kid.

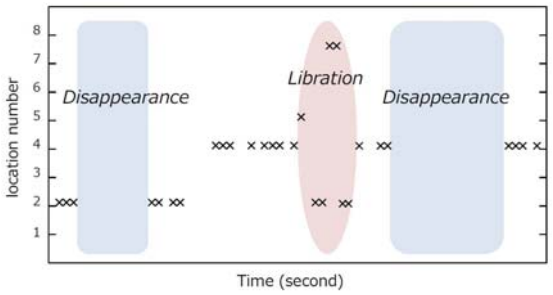
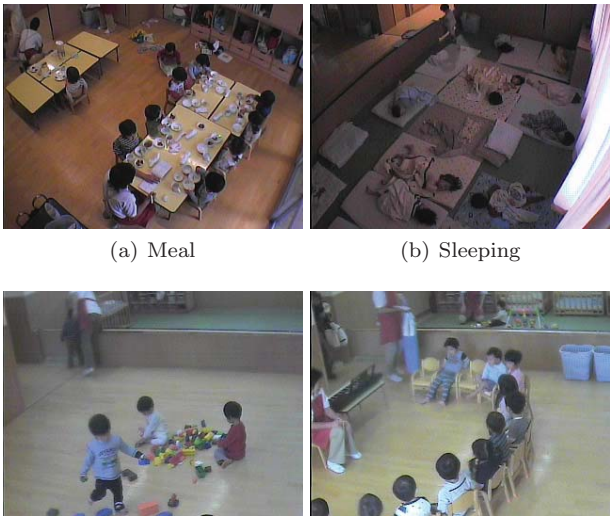


Figure 3. The practically generated RFID log is noisy.

5 Event Recognition and Clustering

After adapting RFID log analysis, the remaining videos are expected to have the target kid's presence. However, since the total amount of these videos is still very large, it is necessary to abstract it into a shorter daily digest for a convenient watching.

We think the daily life of nursery school kids consists of a limited number of chief events, therefore, an ideal video digest for them should cover individual chief events that the target kid participated in. In our work, we define four kinds of chief events, they are: playing, playing, meal and group activity. In Fig.4, we display some sample images of these events.



(a) Meal (b) Sleeping

(c) Playing (d) Group Activity

Figure 4. Example of events.

In our work, we find individual chief events and construct the digest through learning based event recognition and event clustering. For efficient processing, we use one minute as the unite time interval, and conduct the event recognition and cluster on one minute video segments.

5.1 Features

For every frame I_n in the video segment i, we calculate: 1).the number of changed pixels N_{nc} by inter-frame differencing; 1).the number of foreground pixels

Figure 5. Image Differencing and Erosion (first row: background subtraction; second row: inter-frame differencing; first column: input continuous frames; second column: grey scale image; third column: binarized image by Otsu's thresholding; last column: after erosion)

N_{nf} by background subtraction; and 3). the number of foreground black pixels N_{nb}. N_{nf} could implicitly reflect the global condition in the scene, while N_{nc} gives some hint about the motion intensity. Finally, the number of foreground black pixels N_{nb} reflects the number of kids in the scene.

While calculating the N_{nc} and N_{nf} from the result of background subtraction and inter-frame differencing, we adapt Otsu's global binarization method [4] and binary erosion [5]. Otsu's method finds the threshold that minimizes the within-class variance between changed and unchanged pixels, therefore is robust against changes in illumination. The following erosion step could help to remove bridges, branches and noise. We visualized the examples of our visual feature computation in Fig.5.

The mean of the N_{nf}, N_{nc} and N_{nb} within a video segment construct the three kinds of visual features. For some event, time is also a important feature. We use the time as the forth feature T, and compute it for a video segment of "a hour b minute" as: $T = 60 \times a + b$. These features are constructed in to a four dimensional vector for each video segment.

5.2 Event Recognition and Clustering

In the event recognition step, every video segment is assigned one of the four chief event labels. This is done by using a pre-trained classifier. In our work, we use AdaBoost [7] to train an "one vs.all" classifier for each kind of chief event. When novel video segment comes, we apply all the four classifiers on it, and apply majority rule to determine its category.

After every video segment have been assigned a chief event label, we collect the segments that have the same label together. In order to discover individual events in each chief events data, we apply k-means clustering. The resulted k clusters represent each individual event. Finally, video segments of the k cluster centers are picked out, and are then connected together to generate the digest.

6 Experimental Results

In this section, we evaluate the practical performance of our proposed system in two aspects. One is the accuracy of event recognition, which is the core step in our whole digest generation process. The other is the quality of the generated digest. For this, we distributed the generated digests to the parents, collected the feedbacks, and did a questionnaire survey.

6.1 Evaluation of Event Recognition

The objective of event recognition is to classify each video segment into one of the chief events. In this work, we defined four kinds of chief events, namely, playing, group activity, meal and sleeping. The experiment is to evaluate the practical accuracy of event recognition.

In order do such an experiment, we collect video segments and manually provided label for each one. The final data set consists of 2281 video segments. Among these segments, 1506 segments (meal: 213, sleep: 290, play: 600, and group activity: 403) were used to train the classifier and 775 (meal: 128, sleep: 143, play: 296, and group activity: 208) were used for test.

Using the training segments, we trained a "one vs. all" strong classifier for each chief event using AdaBoost. The number of weak classifier of each strong classifier is set as 10. For the test segments, we apply all the four classifiers and adapt the majority rule to predict its category. The experiment results is shown in Table.1. As we can see, our method could provide over 80% accuracy for the recognition. Though it is not perfect, we will show in the following section that with the event clustering, such accuracy is sufficient for the practical usage in our proposed system.

Table 1. The results of event recognition.

Test(num)	meal	sleep	play	group	rate
Meal(128)	103	19	0	2	80.4%
Sleep(143)	20	119	0	0	83.2%
Play(296)	1	0	259	31	87.5%
Group(208)	12	0	36	154	74.0%
Total(775)					81.8%

6.2 Evaluation of Digest Generation

In order to evaluate the quality of generated digests, we invited three kids to participate the experiment, and generate three specific digests for them. We define the number of individual event in each chief event category as three, and select a one minute video segment for each individual event according to the way we described in Section.5. The processing takes about two hours and the resulted digests are 12 minutes.

Table.2 shows the repeatability of the generated digest. For digest (1) and (3), all the segments were correctly chosen, while in digest (2) one segment that should belong to "playing" is recognized as "group activity" by mistake. As a whole, we usually can reach a higher repeatability of events than the accuracy for individual event recognition. This may mainly benefit from the event clustering phase, which helps to avoid choosing these segments that are misrecognized.

Table 2. The repeatability of events.

	meal	sleep	play	group	rate
Digest 1	3/3	3/3	3/3	3/3	100%
Digest 2	3/3	3/3	3/3	2/3	92%
Digest 3	3/3	3/3	3/3	3/3	100%

To evaluate the quality of the generated digest subjectively, we also produced three digests manually using the same raw data. While manually generate the digest, we follow the following three rules: 1) the digest should consist of different chief events; 2) it should be easy to understand the daily life of the target kid through the digest; 3) it is preferable to make the digest contains more variations.

We invited 24 parents (each group contains 8 persons) to participate our questionnaire survey. They are invited to watch both videos and then answer the following two questions: A) Which digest is better? and B) Which digest gives better description for the events occurring in one day? Through question 1, we want to obtain an overall evaluation of the digests, and through question 2, we want to get the evaluation about our generated digests as daily life digests.

The answers to these questions are asked to choose from five levels: (1)automatically generated digest is better, (2)automatically generated digest is slightly better, (3)same, (4)manually generated is slightly better, (5)manually generated is better. In Table.3, we summarized the result of the questionnaire.

From the results we can see most parents think that the automatically generated digests is almost the same as the manually generate ones. For both questions, more parents chose the (4) or (5) and there were still about five people think the automatically generated digest is better or slightly better.

For a more detail analysis, we find in the answers for digest 1 and digest 3, the number of people who chose (1) or (2) is almost same as the number of the people who chose (3) or (4). The bad evaluation on digest 2 might be partially caused by the error in event recognition.

Additionally, through the questionnaire survey, we obtained some comments from the parents such as "through the digests, we got to know the daily life of our kid in the nursery school", "we got to know something about our kid that can not be seen at home", etc. Also, we received some feedback with useful information like "scenes that including joyful look of kids are desirable", "we do not want to see the kids to be reprimanded", etc.

Table 3. The results of questionnaire survey.

Question A	Persons	(1)	(2)	(3)	(4)	(5)
Digest 1	8	0	2	5	1	0
Digest 2	8	0	2	1	3	2
Digest 3	8	1	1	3	3	0
Total	24	1	5	9	7	2
Question B	Persons	(1)	(2)	(3)	(4)	(5)
Digest 1	8	0	1	4	1	2
Digest 2	8	0	2	0	3	3
Digest 3	8	0	2	4	2	0
Total	24	0	5	8	6	5

7 Conclusion

We presented a video digest generation system that designed for use in nursery schools. Our system integrates a cascade of modules and could efficiently generate the daily digest for nursery school kids. The experiment results has shown that our goal of "generating a video digest that well reflect daily life in a nursery school for a particular kid, and it is interesting to the parent" has been basically realized.

As the future work, to meet severer needs, we will try to enhance the event accuracy by introducing some new features related to color, motion, etc. and take into account multi-view/dynamic-view images to achieve more interesting digest videos.

References

[1] M. , K. Uehara, M. Kumano, Y. Ariki, S. Shimojo, K. Shunto and K. Tsukada: "Video Editing Support System Based on Video Grammar and Content Analysis" IPSJ Journal, Vol.44, No.3, 2003.

[2] K. Miura, R. Hamada, I. Ide and H. Tanaka: "Motion Based Automatic Abstraction of Cooking Videos" IPSJ Journal, Vol.J86-D-II, No.11, 2003.

[3] N. H. Bach, K. Shinoda and S. Furui: "Robust Scene Extraction Using Multi-Stream HMMs for Baseball Broadcast" Proc. of the 8th Meeting on Image Recognition and Understanding, 2005.

[4] N. Otsu: "A threshold selection method from graylevel histograms" IEEE Trans. Sys., Man., Cyber, Vol. 9, No. 1, 1979.

[5] R. Gonzales and R. Woods: "Digital Image Processing" Prentice Hall, 2001.

[6] http://www.livekids.jp

[7] C.M. Bishop: "Pattern Recognition and Machine Learning" Springer, 2006.