Retrieval of Personal Experience Records Using Relevance Feedback with a Structuring Filter

Takahiro KOIZUMI†, Yoshinari KAMEDA††, Yuichi NAKAMURA†

† Faculty of Engineering, Kyoto University
†† Graduate School of Systems and Information Engineering, University of Tsukuba

Keywords: Personal Experience Record, Lifelog, Video Retrieval, Relevance Feedback

Abstract

In this paper, we propose a novel method of "relevance feedback with a structuring filter" that efficiently retrieves personal experience records in the form of videos taken by a head-mount camera or other devices. Because such data can be redundant and shaky, they are not suitable for thorough perusal and we therefore require an efficient method of browsing and retrieving these data. Relevance feedback is a technique that supports flexible content retrieval; however, it is often difficult to select from a vast amount of data appropriate samples that are similar to the records being sought. To solve this problem, our method associates contents based on semantic structures, and returns closely related portions in conjunction with portions obtained by relevance feedback. This process makes content retrieval more flexible and efficient. In this paper, we define the semantic structures of personal experience records, describe the relevance feedback technique with a structuring filter, and present experimental results that show the efficacy of our proposed framework.

1 Introduction

Over the past few years, a number of studies have been conducted on taking personal experience records with cameras or other devices. These devices include, for example, a head-mounted camera and a microphone, which capture what a person sees and hears, and other sensors that acquire other types of information for enriching the captured data [1], [4]. The stored personal experience records can be used in various ways for various purposes. For example, data can be shared not only as working records or video manuals but also as a simple outside memory for personal use. Developments in storage devices have made it possible to record such personal experience records for a considerable period of time such as a day, a week or even a year. Thus, it is important to develop a method of efficiently retrieving portions of data from a vast amount of data.

For this purpose, we previously proposed two approaches: the use of an "environmental view" and the use of "scenes of attention" (SoAs) [5]. The former method utilizes both "personal view" videos captured by a head-mount camera and "environmental view" videos taken by a wide-angled camera placed on a ceiling or wall. The environmental view is used to compensate for the limitations of the personal view, since a wide-angled survey cannot be not captured by a head-mount camera. The two types of data are associated with each other by time and place, and are graphically represented in our graphical user interface (GUI) as shown in Figure 1[6]. This method is capable of defining where, how and for what purpose a person acted while taking records, which makes the retrieval process efficient. The SoAs method summarizes personal view records by scenes of attention, which are defines as those in which the person, equipped with a head-mount camera paid attention to something [5].

Together with the above techniques, we still need a quick and efficient retrieval method that can handle a

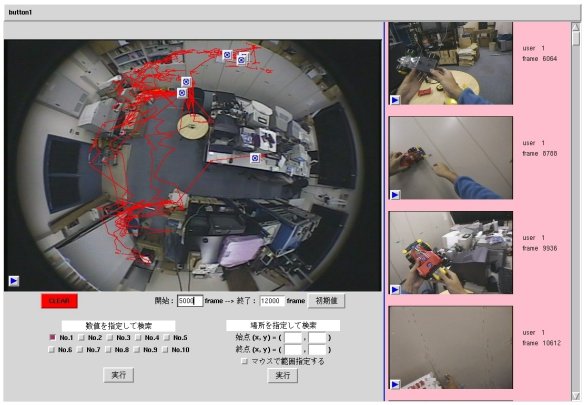

Figure 1: Presentation example of personal experience records: the image on the left shows the view of an environmental camera and that on the right shows the retrieved scenes of attention (SoAs).

large amount of data, since we expect that data will be accumulated every day and everywhere. Moreover, the retrieval process should be flexible since it needs to be able to respond to a variety of requests, *e.g*, remembering something, learning something, reporting something, etc. Considering this problem, we introduce *relevance feedback (RF) with a structuring filter (SF)*, which is an efficient content retrieval method that associates data portions based on the semantic structures of the recorded data. This method searches for relevant data portions even if a user cannot imagine exact target instances. In the following sections, we will briefly explain the idea of RFSF, and give simple examples of how the method effectively works.

2 Retrieval of Personal Experience Record

2.1 Personal Experience Record Retrieval

Personal experience records consist of personal view records and environmental view records which are associated with each other. We can assume that each record is an aggregation of still images or short video segments. Based on this idea, we consider each *SoA* associated with a particular time and place as a unit of the experience record, since we previously showed that a collection of SoAs can be a concise summary of personal experience records [5]. The present method improves the efficiency of retrieval by removing insignificant scenes such as a scene of nothing but walking, a scene in which the person in question simply holds still, etc. Thus, the problem of experience record retrieval is treated as a problem of searching for relevant SoAs from a vast amount personal experience records.

2.2 Similarity Search

When a user searches for something in personal experience records, he cannot usually specify the exact time or place, or supply a sample image. In most cases, the user's requests takes a more general form, such as "how did someone work around that place?" or "where is a book like this?", etc. Relevant portions of experience records, therefore, cannot always be successfully retrieved by a single fixed criterion, and we thus need flexible adjustment of similarity measurement. Suppose that a user wants to retrieve a scene that includes one of the user's books. The user cannot always specify an exact key image or similarity measure for retrieval, even though he may know quite precisely what he wants. For this purpose, a similarity search with RF is a useful function. In RF for image retrieval, the distances or similarity measures among images are adjusted based on the user's response, i.e., feedback. A user simply needs to choose images that may be similar to the image(s) that he wants. Nevertheless, we can easily imagine many cases in which a user cannot specify similar images or experience records. For instance, a user might want to know which tools are necessary for a particular task. In this case, the user probably does not know the shape or color of the tools, and it is difficult for him to guess where and when the tools might have been captured on video or to specify sample images that might include images of the tools. Thus, we need a more sophisticated method for the retrieval of personal experience records.

2.3 Retrieval by using Semantic Structure

Most human activities do not occur randomly or independently of other activities, and they have close relationships

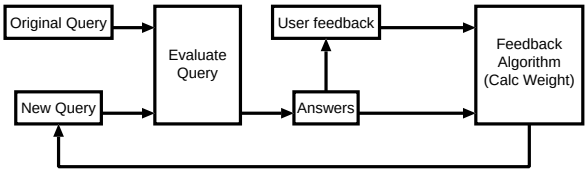

Figure 2: Retrieval process with Relevance Feedback (RF).

to preceding or succeeding events or even to events occurred at a completely different time. For instance, the activity of manipulating a device is related to its preparation, the result obtained by the manipulation, the same operation on different days, operations on related devices, the training for the manipulation, etc. Behaviors around the same time of day on different days are likely to be similar, though with small differences. Hereafter, we denote such a portion closely related to a specified portion as *related segment*, and the key aim of the present research was to achieve efficient retrieval by associating related segments. The following are typical relationships, each of which is semantically important:

(a) relationships among portions within a task or a meaningful behavior.

(b) relationships between objects and their ordinary states, such as that between a tool and its toolbox.

(c) relationships among records across different persons who worked cooperatively.

(d) a variety of temporal relationships.

Many other relationships are also possible, especially when we consider the use of speech. RFSF provides associations based on these relationships in conjunction with relevance feedback, and realizes flexible data browsing by retrieving related segments.

3 Personal Experience Record retrieval by Relevance Feedback

Figure 2 shows the typical flow of RF. First, the system chooses the initial set of samples from a data set. This set may be randomly chosen, or it can be selected by some criteria given by a user. Given the initial set, the user then chooses samples (denoted as $\boldsymbol{Rel}$) that are relevant or close to relevant with respect to time, location, or image appearance. The similarity measure is modified by using the statistical characteristics of $\boldsymbol{Rel}$, and the similarity measure is used by the system to select the samples for the next iteration. Again, the user chooses samples from the data retrieved by the new similarity measure. The system and the user repeat this process until the user finds relevant segments. A typical example of a similarity calculation is given by the following formulas:

$$r(\boldsymbol{x}_i) = \sum_{\boldsymbol{x}_j \in Rel} D(\boldsymbol{x}_i, \boldsymbol{x}_j)^{-\alpha} \quad (1)$$

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\frac{1}{N} \sum_{d=1}^{N} w_d |x_{id} - x_{jd}|^{\beta} \right)^{\frac{1}{\beta}} \quad (2)$$

where $r(\boldsymbol{x}_i)$ represents the relevance, $D(\boldsymbol{x}_i, \boldsymbol{x}_j)$ indicates the distance between segment $\boldsymbol{x}_i$ and segment $\boldsymbol{x}_j$, w_d represents the weight for the d-th dimension, and α and β are constant values, each of which usually ranges from 2 to 5. Smaller values of $D(\boldsymbol{x}_i, \boldsymbol{x}_j)$ indicate more similar records between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. A larger value of $r(\boldsymbol{x}_i)$ indicates that $\boldsymbol{x}_i$ is more relevant, since $r(\boldsymbol{x}_i)$ is the sum of distance values raised to $-\alpha$. This relevance value is calculated for each segment in the personal experience records, and the segments are then sorted according to their relevance values. Next, the top k segments are shown to the user as an intermediate result. Each weight w_d is adjusted between 0 and 1 by Equation 2. A small distance value has a stronger effect on the relevance value if α in Equation 1 is large. On the contrary, large distance value has stronger effect if β in Equation 1 is large. In a typical RF, the weight for each dimension is updated by the following formula:

$$w_d^{new} = \frac{\sigma_d^{rel}}{\sigma_d^{all}} \quad (3)$$

$$w_d = \gamma \cdot w_d^{new} + (1 - \gamma) \cdot w_d^{old} \quad (4)$$

where γ is a constant that controls the strength of feedback, whose value is empirically determined between 0.6 and 0.9, σ_d^{all} is the variance of the d-th attribute value for all segments, and σ_d^{rel} is that for the segments included in Rel. This method is based on the assumption that the variance of x_{id} must be small if its value is an essential

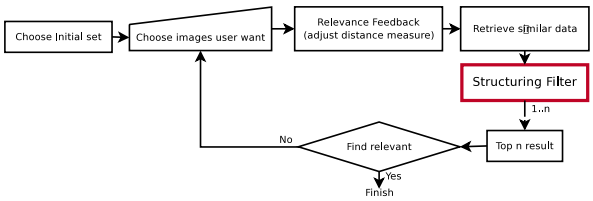

Figure 3: Process of retrieval of personal experience records

condition for user selection. Therefore, when the ratio of σ_d^{rel} to σ_d^{all} is small, the d-th attribute value has much importance while, in contrast, when the ratio is large, the dimension has less importance.

4 Structuring Filter

4.1 Relevance Feedback with a Structuring Filter

In the present study, we used our proposed RFSF method, which automatically associates related segments that are semantically proximal to already retrieved segments. Figure 3 shows the revised process flow for RFSF. We first introduce an association matrix A_γ that expresses the strengths of the relationships among data segments. The value of the formula A_γ.

$$A^\gamma = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

a_{ij} ranges from 0 to 1, which represents the strength of the relationship between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The value of a_{ij} is determined based on the type of involved relationships. An association matrix is prepared for each relationship, and the weighted sum of those matrices described below is used to compute actual association. Let us posit a relevance vector whose elements denote the relevance values of each segment, which are calculated by Equation 1; this relevance vector is denoted as $\boldsymbol{r} = \{r(\boldsymbol{x}_1), r(\boldsymbol{x}_2), \cdots, r(\boldsymbol{x}_N)\}$. Our revised relevance vector $\boldsymbol{r'} = \{r'(\boldsymbol{x}_1), r'(\boldsymbol{x}_2), \cdots, r'(\boldsymbol{x}_N)\}$ is calculated by using the association matrix

$$\boldsymbol{r'} = \sum_\gamma (w_\gamma A_\gamma \boldsymbol{r}) \tag{5}$$

where $\boldsymbol{w}_\gamma$ is the weight for the γ-th association matrix. If we give 1 as the weight of the identity matrix and 0 for all other weights of association matrices, we get a revised relevance $\boldsymbol{r'}$ equal to $\boldsymbol{r}$. The revised relevance r'_i becomes higher, the more related segments in Rel given by a segment $\boldsymbol{x}_i$. This sometimes produces a serious drawback. For example, the following association matrix represents a relationship in which $\boldsymbol{x}_i$ occurred before $\boldsymbol{x}_j$, and this can be given as an upper triangular matrix if the segments are sorted in chronological order:

$$A^{past} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

The revised relevance values then tend to be $r(\boldsymbol{x}_1) > r(\boldsymbol{x}_2) > \ldots > r(\boldsymbol{x}_n)$, regardless of the relevance values $r(\boldsymbol{x}_i)$ In such a case, we use only the top k related segments to each segment to calculate the revised relevance in order to reduce the effect of the number of related segments. As a result, a segment that is closely related to the segments with high relevance values receives a high revised relevance value regardless of the number of related segments.

4.2 Semantic Association by the Structuring Filter

In this section, we explain typical relationships for association matrices and how to obtain them.

(a)Relationship among task portions within a task or a behavior

In many cases, a user may wish to retrieve task portions including more than a single shot of a moment, even though the query is based on a single image, *i.e.*, that is, a snapshot of a moment. Thus, it is often preferable that mutually related portions within the same task be retrieved simultaneously. For this purpose, if records are

partitioned into a collection of tasks, we can efficiently use this relationship by utilizing rough approximations although precise partition is often difficult. One simple method is partition by entering/exiting a room, which is based on the assumption that a person perfomes a single task between entering and exiting. Segments in the same partition are considered to have a close relationship of this type, and we can denote the association matrix as

$$a_{ij} = \begin{cases} 1, & if\ j \in \boldsymbol{A}_i \\ 0, & otherwise \end{cases}$$

where $\boldsymbol{A}_i$ represents a set of $\boldsymbol{x}_j$ that is included in the same partition based on entering/exiting a room.

(b)Relationship between an object and its ordinary state

A user often requires information on where and how an object is located or stored, however it may be difficult to search for this information using only a similarity search: the object may be stored in a box, for example, in which case the color and shape of the search object may differ from those of the object in question. Thus, it is useful for tasks involving finding or storing an object to associate the object with its ordinary state. The following is a typical behavioral pattern: "Unless a person prepares everything before performing a task, the person must eventually go to the place where the necessary object is located/stored in order to bring it to the work space." This pattern may be used to identify the main workspace and secondary places where the person stopped. The main work space is considered to be the place where the person stayed for the longest time, and we denote the set of recorded segments taken at this place as $\boldsymbol{M}$. Secondary places are then identified by locating the places the person moved to directly from the main work space and stayed for a short period, returning directory to the main work space. The set of record segments taken at any secondary place is denoted by $\boldsymbol{S}$. If the image of any segment in $\boldsymbol{M}$ and the image of any segment in $\boldsymbol{S}$ are similar, these two segments are considered to have the object-ordinary state relationship, based on the assumption that the same object is captured in both images if the images are similar. This can be denoted as follows:

$$a_{ij} = \begin{cases} 1, & if\ sim(\boldsymbol{x}_k, \boldsymbol{x}_j) < thr_d\ for\ (\boldsymbol{x}_k \in \boldsymbol{S}) \wedge (\boldsymbol{x}_j \in \boldsymbol{M}) \\ 0, & otherwise \end{cases}$$

$$sim(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{16} \sqrt{(h_{ik} - h_{jk})^2}$$

where $sim(\boldsymbol{x}_i, \boldsymbol{x}_j)$ represents the similarity between two images. During the current step, we use color histograms of attention areas, which are detected from SoAs,[1] to calculate this similarity.

(c)Relationship among records across different persons

When we perform a task, we often work cooperatively with other people: someone brings us something, gives us advice, etc. We can retrieve such data by checking the recording time and location of colleagues. If the locations of two or more people are close enough to each other at a certain time, we consider the recorded segments to be mutually related. Let X^{my} be a segment in my experience record, and let X^{other} be that in the record of another person. We then obtain the following association:

$$a_{ij} = \begin{cases} 1, & if\ tdist(\boldsymbol{x}_i^{other}, \boldsymbol{x}_j^{my}) < thr_t \wedge dist(\boldsymbol{x}_i^{other}, \boldsymbol{x}_j^{my}) < thr \\ 0, & otherwise \end{cases}$$

$$tdist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{t \in \boldsymbol{t}} \sqrt{(t_i - t_j)^2}$$

where $dist$ is a function used to calculate the distance between two people, and $tdist$ is used to calculate time difference between two segments.

(d)A variety of temporal relationships

We can also imagine a variety of temporal relationships that may be used to associate segments. For instance, we often do similar things at a similar time of day, and this relationship is not well counted as proximity in a chronological sense. A certain interval, *e.g.*, 30 minutes, may be significant in a given task. Likewise, the initial or final state of a task may have great importance. As a typical example, we can define a past-future relationship for the last case. The association matrix of a future association is a lower triangular matrix and the past association

[1] In the previous works in our group, attention areas are detected from scenes of attention. Although this detection is not sufficiently accurate for precise indexing, we can nevertheless expect a considerable increase in accuracy.

Figure 4: Personal view camera

matrix is an upper triangular matrix, since segments are stored in chronological order:

$$A^{future} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix} \quad A^{past} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

5 Experiments

5.1 Environment

We conducted a number of experiments for verifying the efficiency of our proposed RFSF system. First, personal experience records were taken for tens of hours both in the form of personal view records and as environmental view records. In taking records, a person placed a small camera on his head, as shown in Figure 4, and personal views were recorded using a notebook PC installed with a hardware MPEG2 encoder card. Environmental view records were captured by an SXGA digital video camera (1280 x 960, 7.5fps, IEEE1394) fixed on the ceiling, and the videos were recoded on a PC through a software MPEG2 encoder. The captured personal view records and environmental view records were processed, and the SoAs and location(s) of the person were tracked. The SoAs and human locations were associated with each other, and personal experience records that included SoAs, locations, times and index numbers were generated.

5.2 Retrieval Experiment

For our retrieval experiment, we gave a subject user the following purposes of retrieval:

- Where did the person last use the digital video camcorder (DV cam)?

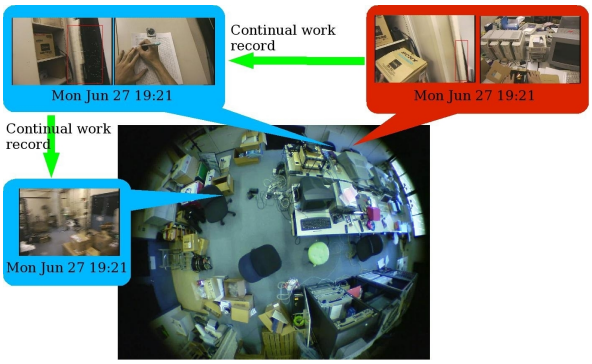

Figure 5: Retrieval result of a query about sequential work after taking out the DV cam: SoAs with red backgrounds were retrieved as a result of a similarity search, and those with blue backgrounds were retrieved by the association of the sequential task.

- Did the person mark a check sheet when he took out the DV cam.
- Where is the package of this software which was installed in this PC?
- Where is the DV cam usually stored?

These examples are typical retrieval purposes that are not well handled with a simple similarity search. In the present experiment, we used association matrices generated automatically as described in section 4.2 above. By using the following process based on several association matrices, we can efficiently retrieve objective segments:

1. retrieving most recent segments by past-future relationship.
2. retrieving preceding and succeeding segments by the relationship among task portions.
3. retrieving the location of an object used in the task by the association of the object and its ordinary state.

The effect of (1) is obvious, so we will illustrate here that (2) and (3) work effectively.

(2)Retrieving by the relationship among task portions

For process (2), suppose that a person remembers that he did some work using the DV cam, but does not remember whether he marked the check sheet for taking out camera. In this case, if we choose images on the DV cam as key images for a similarity search, we cannot expect that the scene of marking the check sheet will be directly retrieved. Similarly, it may not be efficient

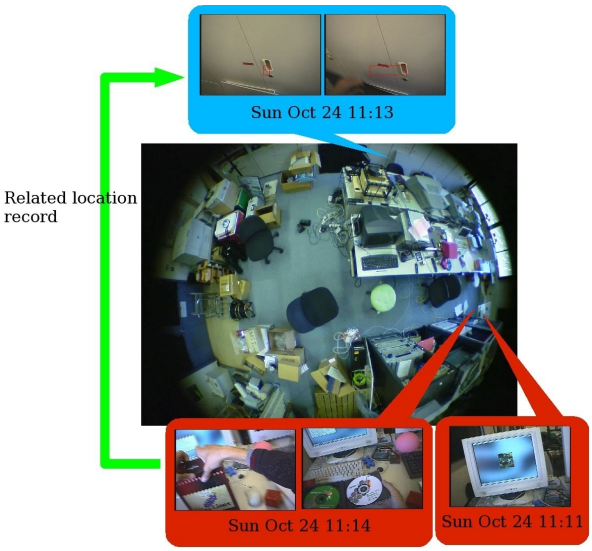

Figure 6: Retrieval result of a query about where the software was taken from: SoAs with red backgrounds were retrieved as similar portions, and those with blue backgrounds were retrieved by the association of an object and its ordinary state.

to directly retrieve the scene by requesting images of the check sheet, since the check sheet may have been accessed many times by many persons. On the other hand, if we can use relationships among task portions, we may reach the objective portion based on our memory of what we did in the task or we can guess what portion is part of the same task. Thus, any portion of the task using the DV cam can potentially be a good key for retrieval. In the present experiment, the objective scene was efficiently retrieved as shown in Figure 5.

(3a)Retrieving the location of an objective tool

For process (3), the retrieval of the location of an objective tool, as explained above, we are able to search for the location of an object by using the association between the main work space and secondary places. In the present experiment, the scenes of taking/storing the software package that was installed in the task were retrieved based on this type of association. Figure 6 shows the results in which the scene of installing the software and the scene of a locker were retrieved. Then, based on a query about the scene of a locker, the scene of taking out the software package was retrieved, as shown in Figure 7.

Figure 7: The result of a repeated query at the related place which is denoted at the time before query: SoAs were retrieved by the similarity of the location where denoted by the blue background that appears in Fig. 6

(3b)Retrieving the place where an object is usually stored

In order to retrieve a scene identifying the location where the DV cam is usually stored, actual retrieval was performed as follows. At the initial step of retrieval iteration, we expect to obtain some images related to the target object. If none is given, random selection is repeated. Next, by choosing those images as feedback to the system, we can expect to obtain segments showing the location where the target object is stored, segments showing the location to which the object is taken or brought, or other related segments. Then, by choosing certain segments and excluding segments clearly undesired, the objective segments tend to appear at a higher rank. In our experiment, choosing segments that included the image of the DV cam without considering the time and place worked well. Figure 8 shows our results.

6 Conclusion

In this paper, we proposed a novel "relevance feedback with a structuring filter" method that allows efficient content retrieval for data sets with internal structure. This technique efficiently uses relationships inside content, gives appropriate candidates to a user, and accelerates content retrieval. Experiments showed that the technique is applicable to the retrieval of personal experience records. In future work, we will examine the automatic construction of structuring filters as well as weight control for

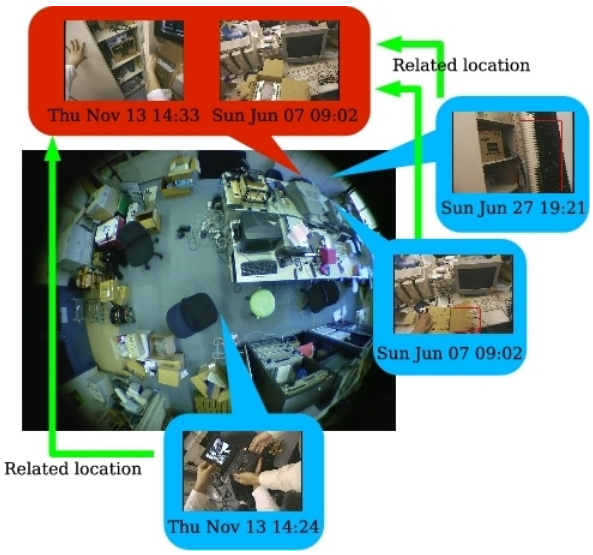

Figure 8: Location where the target item is usually located: by selecting the SoAs which include the images with blue backgrounds from the DV cam, SoAs with red backgrounds are retrieved by the relationship of each location.

combining structure filters.

References

[1] T. Kawamura, Y. Kono, M. Kidode Wearable Interfaces for a Video Diary: towards Memory Retrieval, Exchange, and Transportation, In *ISWC2002*, pp. 31–38, 2002.

[2] Nobuchika Sakata, Takeshi Kurata, Takekazu Kato, Masakatsu Kourogi, and Hideaki Kuzuoka WACL: Supporting Telecommunications Using Wearable Active Camera with Laser Pointer, In *Proc. 7th IEEE International Symposium on Wearable Computers (ISWC2003) in NY, USA*, pp. 53–56, 2003.

[3] T. Kawashima, K. Yoshikawa, K. Hayashi, Y. Aoki Situation-based Selective Video-Recording System for Memory Aid. In *IEEE Proc. of Int. Conf. on Image Processing, III*, pp. 835–838, 1996.

[4] K. Aizawa, T. Hori, S. Kawasaki, T. Ishikawa Capture and efficient retrieval of life log In *Pervasive 2004 Workshop on Memory and Sharing Experiences*, pp.15–20, 2004.

[5] Y. Nakamura, J. Ohde, and Y. Ohta. Structuring personal activity records based on attention - analyzing videos from head-mounted camera. In *15th Int'l Conference on Pattern Recognition Track4*, pp. 220–223, 2000.

[6] S. Kubota, Y. Nakamura, and Y. Ohta. Detecting scenes of attention from personal view records – motion estimation improvements and cooperative use of a surveillance camera. In *IAPR Workshop on Machine Vision and Applications*, pp. 209–213, 2002.

[7] Michael Ortega-Binderberger, and Sharad Mehrotra. Relevance Feedback in Multimedia Databases.

[8] Leejay Wu and Christos Faloutsos and Katia P. Sycara and Terry R. Payne. FALCON: Feedback Adaptive Loop for Content-Based Retrieval. *The VLDB Journal*, pp. 297–306, 2000.