A Novel Video Retrieval Method to Support a User's Recollection of Past Events Aiming for Wearable Information Playing

Tatsuyuki Kawamura, Yasuyuki Kono, and Masatsugu Kidode

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-Cho, Ikoma, Nara 630-0101, Japan
{tatsu-k,kono,kidode}@is.aist-nara.ac.jp

Abstract. Our system supports a user's location-based recollection of past events with direct input such as in always 'gazing' video data, which allows the user to associate by simply looking at a viewpoint, and providing stable online and real-time video retrieval. We propose three functional methods: image retrieval with motion information, video scene segmentation, and real-time video retrieval. Our experimental results have shown that these functions are effective enough to perform wearable information playing.

1 Introduction

This paper suggests a novel application of wearable computers for location-based recollections of past events. Our aim is to realize a framework that supports a direct, intensive, and automatic extension of human memory. Since a wearable information player [1] in the near future should be available for a daily uses, it must be portable and independently usable at any time/place. Our System can retrieve an associable video in previous recorded video data set, which is stored in the wearable computer, triggered by current video data, which is captured by a head-mounted camera. Users do not have to know that they are storing video data of their daily lives, but do not plan on utilizing such video data in the future.

Our intelligent framework should understand locations that are recorded in video data as scenes to be searched for in the future. If the user tracks a moving object, its video scene might look as if the object has stopped and its background is moving. In order to retrieve a video even in this kind of moving objects, we introduce effective functions to detect the movement of the user's camera and compensate its effect. We suggest three functional methods that use a retrieval method of location-based similar images of past events using motion information, a method of segmenting video scenes, and a method of dividing video data for real-time process.

2 Our Approach

Our system provides associable video retrieval in a previous data set triggered by current video data. The similar images of associable video images are shown in Figure 1. To achieve high speed and an appropriate location-based video retrieval

method, our system must efficiently pick up location information from the video data achieved by a face-on wearable camera placed at the center of a head-mounted display (depicted in Fig.1.). Tracking the user's head movements and moving objects in a scene and avoiding above two motion on the video retrieval process.

Fig. 1. Similar images in terms of location (left) and our wearable camera (right)

Motion information exclusion from video data is performed using the wearable camera by:

- tracking of yaw and pitch head movements with two mono-axis gyro sensors,
- tracking of moving objects in a scene using a block matching method, and
- excluding motion information by masking moving areas in the scene.

Video scene segmentation from continuous input video data is changed by:

- detecting scene changes continuously from current video data and two gyro data, and
- indexing each scene for easy viewing.

Real-time video retrieval from large sequential video data is retrieved by:

- dividing small segments from video data for stable and high-speed video retrieval, and
- retrieving a associable scene from a segment similar to the current video data.

2.1 Location-based Video Retrieval

Our aim is to achieve a support level for user's memories with a wearable computer. This support system retrieves an associable video data set with current video data from approximately the same viewpoint. This approach retrieves video scenes that trigger user's memory such as about persons, objects, actions, reasons, and time, which relate to a location.

The user's ideal recollection support system must include several functions. The "Forget-me-not" system [2] can detect the person whom a user met, or the time when the user give/take the person a document. The remembrance agent system [3] supports the editing of documents related to a particular time/place from the history of editing by the user. The use of these two studies is limited to an indoor environment, because sensors placed on the sides of the room. In contrast to the above two studies, the following studies use a video and stand-alone type wearable system. Clarkson's system [4], however, cannot directly retrieve previous associable video data for a user who wants to know detailed location information. Aoki's system [5] also cannot select similar video of approximately the same place/viewpoint quickly from continuously recorded video because an offline training sequence is used.

2.2 Image Retrieval using Motion Information

Wearable computers must treat motion information. We divided motion information into two types. The first type is the user's head motion information. The second type is moving object information. Each type of information must be removed to retrieve location-based images.

In the image retrieval method, we adopt three processes. The first process is to exclude head motion information from merged motion information. Motion information is made from two sequential images. We attached perceptual sensors near the wearable camera to recognize the user's head motion. These sensors are two mono-axis gyro sensors, which detect both yaw and pitch axis head rotation information. The second process is to recognize moving objects in a scene using a block matching method. This method divides an image into small blocks. The third process calculates the similarity of images using motion information detected by the prior processes. Location-based image similarity in this paper is defined as follows:

- an image recorded from approximately the same viewpoint, and
- moving objects that do not influence the similarity of an image.

Tracking Head Movements

User's head movements directly influence video data from the wearable camera. This head motion information must be removed from video data for recognizing moving objects. We adopt two mono-axis gyro sensors and place these sensors at the center of the wearable camera as shown in Fig. 1 (right). In order to remove the user's head motion information, an examination of the relationship between the amount of value transition with the gyro sensor and the amount of shift with images is necessary. Fig.2 show both relationships (left: yaw, right: pitch). These results can remove the user's head motion information from video data.

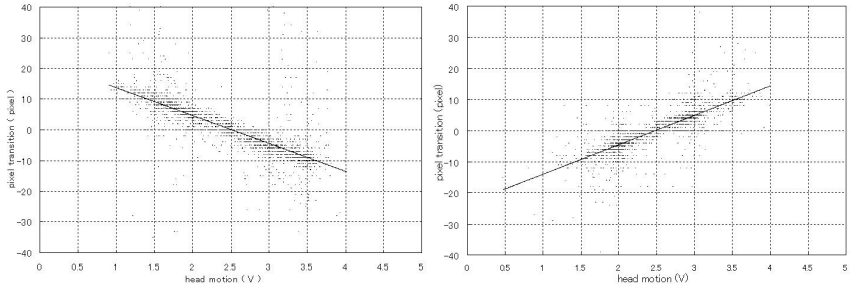

Fig. 2. A gyro sensor value and the amount of image shift value (yaw and pitch)

Tracking Moving Objects

We adopted a block matching method that detects areas, each of which includes moving objects in a scene. The method divides an image into small blocks. The matching process compares divided blocks and an area of the same size. The method is normally limited in calculation amount. The pixel data to recognize moving objects in a scene is the (r, g, b) pixel. The system selects the ($I = r + g + b$, $I_r = r / I$, $I_g = g / I$, $I_b = b / I$) data obtained from the (r, g, b). Our method is defined in the following formulae:

$$Mr_{i,j}(u,v,t) = \sum (Ir(i_0, j_0, t) - Ir(i_u, j_v, t))^2 , \quad (1)$$

$$Mg_{i,j}(u,v,t) = \sum (Ig(i_0, j_0, t) - Ig(i_u, j_v, t))^2 , \quad (2)$$

$$Mb_{i,j}(u,v,t) = \sum (Ib(i_0, j_0, t) - Ib(i_u, j_v, t))^2 , \quad (3)$$

$$M_{i,j} = Mr_{i,j} + Mg_{i,j} + Mb_{i,j} . \quad (4)$$

The calculated minimum value of comparisons in a limited area shows an estimated block motion vector $(u_{\min}, v_{\min})$. The block motion vectors calculated by the above method are redefined into five simple states (up, down, left, right, and non-movement). If a motion vector is adequately small, this block is named as "non-movement."

Exclusion of Motion Information

The current query image and the previous images each have specific motion information. In order to remove mutual motion blocks in each image from target searching blocks, a motion block mask should be made. First, the image matching process compares the same address block in two images with blocks called "non-movement" states. The block matching method uses the same method mentioned in the section "Tracking Moving Objects." The second process divides a value, from the summed values calculated by the previous process, by the number of "non-movement" blocks. We adopt the divided value for an evaluation of image similarity. This value is derived from the value calculated by using the block matching method.

2.3 Video Scene Segmentation

We construct scene changes using color differences appearing on the entire screen and a moving average method with two mono-axis gyro sensors. Unlike a video data used in television, there are no clear scene changes in the wearable video data we use in our research. If the difference and the amount of the value transition of gyro sensors are large, we choose the point to divide the scene. We then merge some sequential short scenes into one scene for easy viewing.

In the moving average method, continuously input gyro sensor values are added from the past value in T frames prior to the current value and T divides added value. This method can obtain a meta-trend of captured data. The moving average method equation is as follows:

$$MA_T(t) = \frac{\sum_{i=t-T}^{t} f(i)}{T} . \quad (5)$$

In this paper, four values are calculated by the moving average method: Two values are calculated with yaw-axis gyro value, and other two values are calculated

with pitch-axis gyro value. The following three states are defined to detect scene changes:

- Stable state: Both the moving average value of a short interval (e.g. T=30) and that of a long interval (e.g. T=300) are under a certain range
- Side1 state: The moving average of the short interval is higher than the long interval.
- Side2 state: The moving average of the long interval is higher than the short interval.

By using the parameter, $MA_T(t)$, a scene change is detected by a state transition. The minimum length of a segmented scene is limited to 30 frames. If the color difference between adjacent images is under a threshold, this frame does not make a new scene.

2.4 Real-time Video Retrieval

The proposed video retrieval method is based on similarity predictions, which divide video data into small segments and retrieves the associable video data, because the cost of the retrieval process increases as the video data set becomes large. In this retrieval method process, all images in a segment l are compared with a current query image from the wearable camera. The next process for video retrieval is changed from l to the next segment $l+1$, when the maximum image similarity, $HM(l)$, is under a threshold th. We consider the following hypothesis: Similar images form clusters in a sequential video set.

$$l = \begin{cases} l, & for\ HM(l) \geq th, \\ l+1, & for\ HM(l) < th. \end{cases} \quad (6)$$

3. Experimental Results

We use an IBM ThinkPad X20 (Intel Pentium III 600MHz, 320 MB Memory). The wearable camera has an I-O DATA USB-CCD 320x240 triad resolution. We selected a video see-through head-mounted display, the OLYMPUS EYE-TREK FMD700. The motion detection sensor, which has two mono-axis gyro sensors, is from Silicon Sensing Systems (max 100deg/s).

3.1 Video Retrieval with Motion Information

The experiment took place during the daytime, outdoors, in a hall, and in a room. The experimental tasks were recorded four times in each place. The first frame image in the recorded video data is defined as the retrieval query. The experimental task consists of two parts. One task was for the subject to wave his/her hand several times.

The other task required that the subject turn his/her head to the left. The “normal” method, which does not consider motion information, was performed to compare with our proposed method. The result is shown in Fig.3. In the figure, a higher location-based image similarity corresponds to a lower evaluation value. Our method clearly shows a higher similarity than the normal method in the hand waving task. Our method removes larger distance from the evaluation values in the hand waving task to the evaluation values in the head turning task than in the normal method.

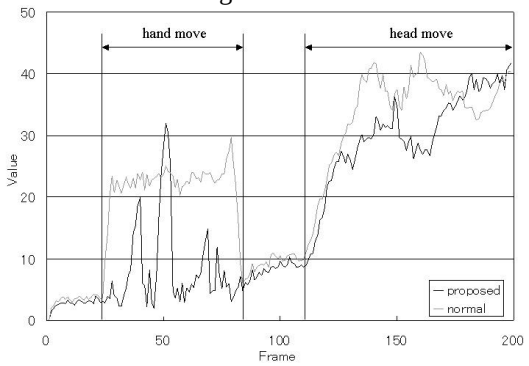

Fig. 3. Evaluation and targeting of video similarity between base images

Table 1 illustrates the relevance and recall rates of both methods. The relevance rate is the rate of retrieved data for all correct data. The recall rate is the rate of correct data for all retrieval data. Our method is well suited to retrieve location-based similar images, because the relevance rate of the proposed method performed 1.3 times as well as the normal method.

Table 1. Relevance and recall rate

	relevance		recall	
	proposed	normal	proposed	normal
outdoor	0.90	0.54	0.98	0.96
hall	0.97	0.56	0.92	0.90
room	0.88	0.57	0.97	0.96
average	0.92	0.56	0.96	0.94

3.2 Video Scene Segmentation

Both image and gyro data, which consist of 5584 frames (191.36 seconds, 29.19 frame/second), were recorded for evaluation. A subject walked two times around a specified route in our laboratory. We set intervals of the moving average as 30 frames (about 1 second) and 300 frames (about 10 seconds). The process limited the minimum scene length to 30 frames. The remarkable result of the experiment is shown in Fig.4. Upper lines in the figure are the moving average values. The lower line shows the scene changes. The scene changes took place 9 times in the figure.

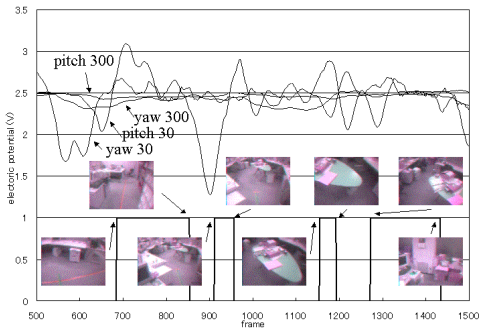

Fig. 4. Making scene changes using two mono-axis gyro sensors

The following results (5584 frames) ware obtained: 41 scenes were segmented from a video data set. The average length of the segmented scenes was 119.51frames (4.09 seconds). The minimum and maximum length of segmented scenes were 31 frames (1.06 seconds) and 756 frames (25.90 seconds), respectively. The minimum scene was made when the subject walked quickly through a narrow path. The maximum scene change was segmented on a long straight and un-diversified hallway. From the experimental results, we conclude that a user's motion information can be used to segment scene changes on a wearable computer environment.

3.3 Real-time Video Retrieval

We compared our method and the normal method. The data set is the same as that for the scene segmentation. We used the first half of the recorded data set as an associable data set, and the latter half as a query of one. We divided the associable data set into 30 small segments. We set a process condition that retrieves location-based similar images from the same segment when the evaluation value is over or equal to 0.2.

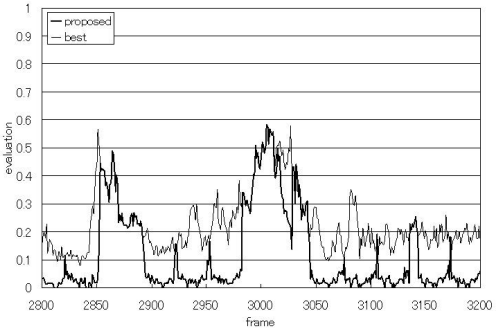

Fig. 5. Proposed and full data sequence search

The remarkable part of the results of the experiment is shown in Figure. 5. In the figure, a higher similarity of location-based images corresponds to a higher evaluation

value. The thick line shows the experimental results of our method, and the other line to the normal method. The calculation time of the process per frame reduced searching for full data sequence to 1/30. The best evaluation value of all data sets is tracked by comparing the same segment of divided video data to the query image when the evaluation value is maintained over or equal to 0.2. We conclude that the hypothesis regarding the clusters of similar images is correct.

4. Concluding Remarks

We have proposed three functions to support a user's location-based recollection of past events. First, we have realized the associable image retrieval method with head movements and moving objects in a scene. We adopted two mono-axis gyro sensors for tracking two-axis head movement. In the future, we are planning to implement other axes or the parallel movement, such as roll axis head movement or walking by various sensors. Second, we proposed a video scene segmentation method to make comprehensible video scenes for users. We adopted the moving average method using two mono-axis gyro sensors to recognize the user's action history. Finally, we proposed a real-time video retrieval method that divides a large video data set into small segments and selects the segment to search an associable video scene with a continuous input image. A future direction of this study will be to develop a faster, stabilized, and efficient video retrieval method to cope with longer continuous video memory.

Acknowledgements

This research is supported by CREST of JST (Japan Science and Technology).

References

1. M. Kidode: Advanced Computing and Communication Techniques for Wearable Information Playing (in Japanese). IEICE, SIG-PRMU2000-159, pp. 93-94, 2001.
2. M. Lamming, and M. Flynn: Forget-me-not: Intimate computing in support of human memory. In FRIENDS21: International Symposium on Next Generation Human Interface, pp. 125-128, 1994.
3. B.J. Rhodes: The Wearable Remembrance Agent: a System for Augmented Memory. Proc. ISWC'97, pp. 123-128, 1997.
4. B. Clarkson, and A. Pentland: Unsupervised Clustering of Ambulatory Audio and Videol. Proc. ICASSP99, 1999.
5. H. Aoki, B. Schiele, and A. Pentland: Realtime Personal Positioning System for Wearable Computers. Proc. ISWC'99, pp. 37-43, 1999.