3—14

HySIM: A Hybrid-space Image Matching Method for High Speed Location-based Video Retrieval on a Wearable Computer

Tatsuyuki Kawamura Norimichi Ukita Yasuyuki Kono Masatsugu Kidode*
Graduate School of Information Science
Nara Institute of Science and Technology

Abstract

We propose a high speed video retrieval algorithm with a hybrid-space image matching (HySIM) method. Our aim is to realize a location-based memory support system on a wearable computer suitable for everyday use. This HySIM algorithm has the following technical features; two different feature spaces and a control mechanism image matched against frequent input queries. Our experimental results show that the proposed algorithm can be used to quickly retrieve particular video images from a huge amount of video data.

1 Introduction

We propose a high-speed location-based video retrieval algorithm with a hybrid-space image matching (HySIM) method useful for retrieving particular images from a huge amount of video data.

In this hybrid-space, we have employed two different types of feature spaces: a time-sequential space with continuously recorded images, and an image-feature space with a similar feature value. A user wears a camera, which continually captures images taken from a user's viewpoint with a head-mounted display. The new algorithm operates from the user's viewpoint images, and a memory support system employs this algorithm on-the-fly, in a frame-by-frame video retrieval.

Memory-support using wearable computing has been studied extensively in recent years. However, in this paper, we consider a new memory support environment using the concept of "Experience Recycling". Experience Recycling allows us to realize a memory support system where we can retrieve and display video data from recorded data in daily life with a wearable computer. Users might not have to know that they are continuously recording video data.

We have conducted video retrieval experiments in terms of retrieval accuracy and speed. Experimental results have shown a very high-speed retrieval, which is approximately a 50 times faster retrival method, than the full-search method. This hybrid matching is effective enough to perform video retrieval on a wearable system.

2 Related Works

Clarkson et al. proposed a method for segmenting sequential audio and video data [2]. This method can recognize situations of abstract location. However, this method cannot directly retrieve previous associable video data for a user who wants to know detailed location information. Aoki et al. proposed a real-time personal positioning method with video data [1]. This system is suitable for navigating a user to an unfamiliar location. However, selecting similar video images quickly with a continuously recorded video is difficult because an offline training process is used. For this reason, we have proposed a location-based memory-aid system [3]. We improved a retrieval process of this system in order to conduct higher speed search.

*Address:
8916-5 Takayama-cho, Ikoma, Nara 630-0101 Japan. E-mail: {tatsu-k,ukita,kono,kidode}@is.aist-nara.ac.jp

3 Hybrid-space Image Matching

In proposing a novel, high-speed video retrieval algorithm as well as a Hybrid-space Image Matching (HySIM) method, we have assumed the following two conditions:

- A huge amount of video data, which is aontinuously recorded.
- The query image is input frequently (about 30fps) from a wearable camera.

Section 3.1 describes the HySIM method, the wearable equipment, and the peripheral conditions necessary for video retrieval using a wearable camera.

3.1 Equipment and Conditions

Figure 1: An wearable equipment (camera and display)

We have employed wearable equipment in this study. Figure 1 illustrates a user who wears a head-mounted display that is set with a wearable camera at a center of the display. The display and the camera are connected with a mobile PC that is worn by the user. The user's viewpoint images are always captured from the wearable camera, and the captured images are then drawn in the display for the user.

Continuously recorded video data from the wearable camera has the following two characteristic features:

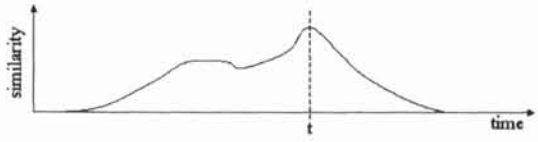

Figure 2: Indistinct video scene cluster

(1) **Continuity:** Adjacent images closely resembling each other. Similarity between the image of point t and another point changes smoothly with the expansion of a captured time difference (Figure 2). This feature creates indistinct video scene clusters.

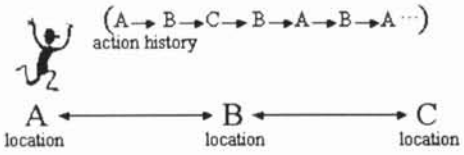

Figure 3: Spatial video feature

(2) **Spatiality:** Based on where the user is standing, similar images are captured by a similar spatial viewpoint. In any one day, for example, the user walks to places where he/she wants to go and. In Figure 3, for example, the user moves around places A, B, and C. A history of where the user walked can be shown as $A \rightarrow B \rightarrow C \rightarrow B \rightarrow A \rightarrow B \rightarrow A$. This example of action history represents how the user arrives at the same place within an odd interval.

In the study, we assume that the recorded video data set is huge. If the system recorded a color video of the user's entire life, the size of all recorded color video images would be over 17 PB ($320(width) \times 240(height) \times 3(bytes) \times 30(frames) \times 60(seconds) \times 60(minutes) \times 24(hours) \times 365(days) \times 80(years)$). When the system links similar images and searches for a video image along linked video sequences, the retrieval time decreases. The similar image linking method, however, does not resolve the video data size problem. In Figure 4(a), if the system runs an image matching process sequentially from the latest video frame to the goal frame, then the system time would sometimes be over 1/30 of a second according to distance between the goal point and the latest video frame. A video retrieval system with a huge video data set, therefore, has to contain a function for resetting the starting point near the goal point, per interval, as shown in Figure 4(b).

(a) From the latest image to goral image

(b) Previous tracking strategy

Figure 4: A sequential search and a tracking search

In our video retrieval system, query images from the user's viewpoint are input with high frequency (the same frequency as the video input rate). By using high frequency input, the system can quickly track the changes of the user's request to refer a location-based associable video data. The user can always slightly change an image with his/her own body control and input a query image to the system at the same time. The system, however, must have the function of an on-the-fly adaptive video search. The system prepares itself using a background processing system for a user's unforseen choices of a scene in the huge video data set.

3.2 An Overview of HySIM

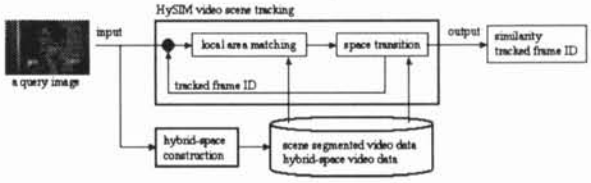

Figure 5: A overview of HySIM mechanism

The HySIM algorithm consists of two processes for high-speed video retrieval. One process includes the construction process of hybrid-space. The other process includes a process of a video scene tracking in the HySIM (Figure 5). In the construction of hybrid-space, the system makes a scene-segmented, video data set and constructs hybrid-space at the same time. In the video scene tracking process, the system searches for a video and then tracks a similar video using a current user's viewpoint query image from the scene-segmented video data set.

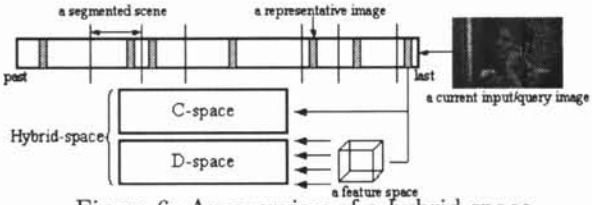

Figure 6: An overview of a hybrid-space

An overview of a hybrid-space is illustrated in Fig 6. The algorithm uses two spaces. These spaces consist of representative images, each of which well represent all images in a single image segment; continuously observed images are segmented so that similar images are stored in the same segmented scene. One of the spaces is a time-sequential space, which is constructed from representative images. We named this space the C-space. Another is a feature space that is also constructed from representative images. We named this discrete space the D-space.

3.3 The Process of a Hybrid-space construction

The construction process of hybrid-space is shown in Figure 7: 1) Segmentation of a video scene and representation of an image from a segmented scene.2) Linking the selected representative image to the last selected image in C-space. 3) Categorizing and linking the selected representative image to the last categorized image in the same space of D-space. In the i th segmentation and representation process, there are n numbers of a representative image candidate that include a current (t th) input/query image. A j th image in the i th segmentation process is shown $R_i(j)$. A temporal representative image R'_i is calculated by equation (1). Segmentation is performed when an error $\varepsilon_{max}(i)$, which is evaluated from equation (2), is higher than a threshold Th.

$$R'_i = \min_j \{\max_k (|R_i(j) - R_i(k)|)\}, (t - n \leq j, k \leq t, j \neq k) \quad (1)$$

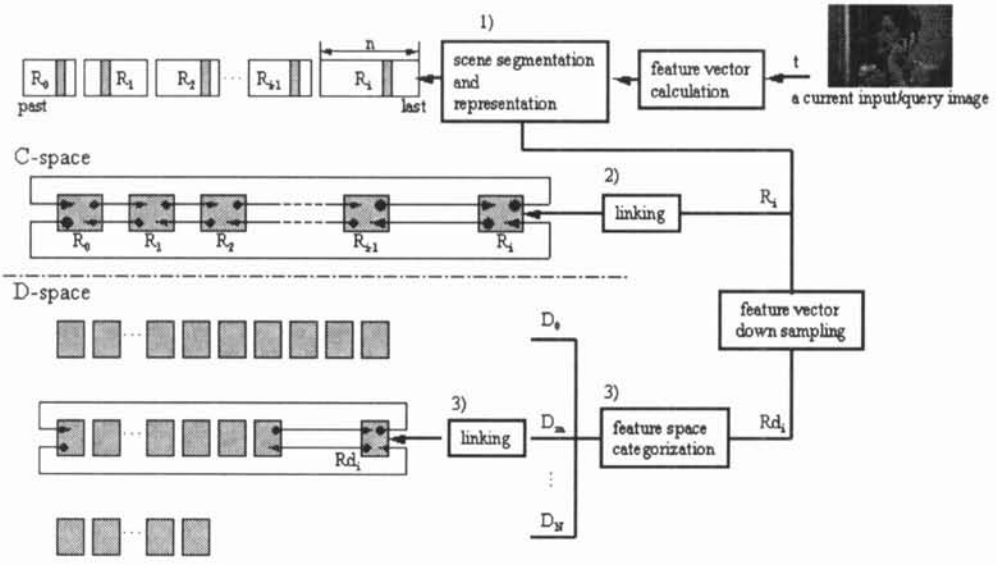

Figure 7: A Construction of hybrid-space

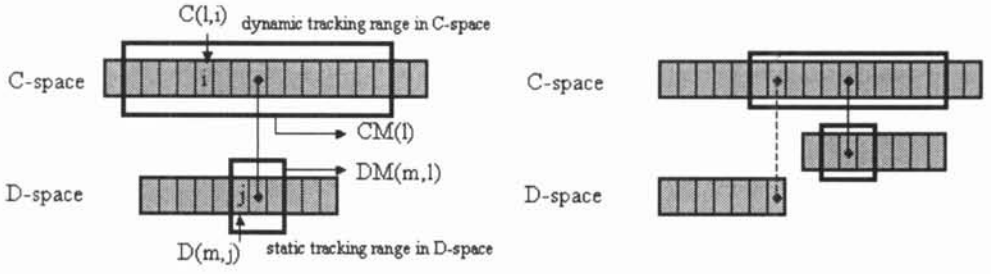

Figure 8: A Construction of hybrid-space

$$\epsilon_{max}(i) = \max_{k}(\left|R'_i - R_i(k)\right|).(t-n \leq k \leq t) \quad (2)$$

After the selection of the representative image, a hybrid-space is constructed using the representative image R_i. The C-space is reconstructed by linking representative images. R_i is linked to R_{i-1} and R_0, and R_{i-1} is unlinked from R_0. In the construction of the D-space, the down sampled Rd_i is categorized in a feature space. The representative image is then linked to the past representative image and the last image in the local space D_m of the D-space.

3.4 The Process of a HySIM video scene tracking

A basic calculation method for a HySIM video scene tracking is shown in Figure 8. The D-space has N dimension in this paper. A representative image is stored in both spaces. In the C-space and D-space, the tracking at a certain moment is performed within a dynamic range (CR) and a static range (DR), respectively. A value of the static tracking range in the D-space is set in advance. $D(m,j)$ represents a similarity between a representative image j and a current query image in dimension m. $D(m,j)$ is identical to $C(j,j)$. Let l denote a center frame of a tracking range in C-space and D-space. $C(l,i)$ represents a similarity between a representative image i and a current query image in the frame l. In this study, all similarities have a range from 0.0 to 1.0. In the tracking range determined by l, values of the $CM(l)$ and $DM(m,l)$ are calculated by the following equations:

$$CM(l) = \max_{i}(C(l,i)), (l + (CR+1)/2 \leq i \leq l - (CR+1)/2) \quad (3)$$

$$DM(m,l) = \max_{i}(CM(i)).(l+(CR+1)/2 \leq i \leq l-(CR+1)/2) \quad (4)$$

The dynamic tracking range CR depends on the value of a center point $C(l,l)$, similar to equation (5).

$$CR = \alpha \cdot A + (1-\alpha) \cdot A \cdot (1 - C(l,l))^2.(\alpha, A : const) \quad (5)$$

Here, A represents a basic range and α prepresents a flexibility rate.

The next tracking step can set a new center frame by using the evaluation of the previous step and the following rule of a space transition of the tracking area.

$$l = \begin{cases} i & \exists_i\{C(l,i) \equiv CM(l), l - \frac{CR+1}{2} \leq i \leq 1 + \frac{CR+1}{2}, i \neq l\} \\ j & \exists_j\{CM(j) \equiv DM(m,l), 1 - \frac{DR+1}{2} \leq j \leq 1 + \frac{DR+1}{2}\} \end{cases} \quad (6)$$

Table 1: A comparison between the full-search and the HySIM search

Data (frame)	10000	20000	30000	40000	50000	60000	70000	80000	90000	100000
R-frame (frame)	1257	2254	6235	10138	11142	12568	13724	15313	16520	21039
Full (sec)	.0106	.0189	.0515	.0835	.0920	.1037	.1138	.1263	.1374	.1744
HySIM (sec)	.0021	.0022	.0031	.0035	.0033	.0033	.0031	.0030	.0030	.0033
Accuracy rate (%)	69.48	69.90	34.07	35.73	34.70	31.93	34.69	41.92	37.61	30.33

4 Experiments

The following experiments were conducted to evaluate the video retrieval performance of the HySIM algorithm. In these experiments, the HySIM algorithm is compared with a full-search algorithm, which uses only representative images. Here, we have evaluated retrieval speed and retrieval accuracy.

4.1 Experimental Methods

In these experiments, one test subject wore a wearable computer system, and continuously walked around for approximately an hour in a building with three halls, and two laboratory rooms. Input query images were captured frame by frame. We used the same images for the query data set as for the recorded data set. 100,000 images were captured for approximately an hour and were divided into 10 data sets. The parameter α in equation (5) is set at 50 in this experiment. The parameter DR is equal to 15 per trial.

4.2 Results

Table 1 shows the performance result of these experiments, in which the data column shows the mumber of images. The R-frame illustrates the amount of representative images in a trial. Both full and HySIM methods show the video data processing times required to retrieve the similar image with an input query image, respectively. The accuracy rate is calculated by the number of frames, which have a similarity with over the threshold Th (set at 0.95 in the experiment) in the segmentation of a video scene or as the best similarity in all recorded images.

The processing time of the full algorithm shows a linear increase. The processing time of HySIM increases approximately 1.57 times when the data of 10,000 frames increases to 100,000 frames. The processing time of HySIM is 52.20 times as fast as faster that of the full algorithm in the data of 100,000 frames.

4.3 Discussion

In this paper we have discussed how the location-based video retrieval method with a huge amount of continuously recorded video data can operate stably and quickly. We are continuing to do experiments to evaluate the effectiveness of the HySIM method.

In this paper, we first discussed the results of the average search time for a 100,000 frame data set. 10,000 and 20,000 frames of data are faster than 30,000$\sim$100,000 frames in the HySIM method. In the two tests, the number of images matched by the system per query image are in most cases enough to search a video from this data set. The results of our experiment show an accuracy rate of approximately 70% in the above two test data sets. Although the scene tracking system may have failed occasionally, our result also show that the tracking point did not keep a local minimum of image similarity.

Secondly, we discussed the results of the accuracy rate, which showed roughly constant in the 30,000 to 100,000 frame data set. We believe that this result was caused by a decrease in search area per query in all data set. In the near future, we are planning to evaluate the performance by a new measure.

5 Concluding Remarks

We proposed a high-speed video retrieval algorithm with a hybrid-space image matching method, which can be used on a wearable computer system for daily use. We have conducted experiments to evaluate video retrieval accuracy and speed. The experimental results have shown that the proposed algorithm is good enough to retrieve a desired video from a huge video data set.

For future research to evaluate performance, we are working on the derivation of a new objective rating method with an appropriate data set. We are also considering about a new performance criterion for the on-the-fly video retrieval algorithm of HySIM, which can evaluate the total performance associated with high accuracy and speed. We are also planning to improve our video retrieval algorithm for higher accuracy, and with less cost of video retrieval.

Acknowledgment

This research is supported by Core Research for Evolutional Science and Technology (CREST) Program "Advanced Media Technology for Everyday Living" of Japan Science and Technology Corporation (JST).

References

[1] H. Aoki, B. Schiele and A. Pentland, "Realtime Personal Positioning System for Wearable Computers," *Proc. of The Third International Symposium on Wearable Computers*, pp.10-13, 2000.

[2] B. Clarkson and A. Pentland, "Unsupervised Clustering of Ambulatory Audio and Video," *Proc. of The International Conference of Acoustics, Speech and Signal Processing*, 1999.

[3] T. Kawamura, Y. Kono and M. Kidode, "A Novel Video Retrieval Method to Support a User's Recollection of Past Events Aiming for Wearable Information Playing," *Proc. of The Second IEEE Pacific-Rim Conference on Multimedia, Springer LNCS2195*, pp.24-32, 2001.