Video Retrieval System Using Handwriting Sketch

Akihiro SEKURA
Future University-Hakodate
116-2 Kamedanakano-cho,
Hakodate, Hokkaido, Japan
e-mail: m1205122@fun.ac.jp

Masashi TODA
Future University-Hakodate
116-2 Kamedanakano-cho,
Hakodate, Hokkaido, Japan
e-mail: toda@fun.ac.jp

Abstract

It is difficult to represent video scenes using keywords. Therefore, in video retrieval, it is not easy to design keyword queries that enable a viewer to represent and find the desired scenes. In this paper, we proposed a video retrieval system that uses handwriting sketches as queries. The user queries the system by drawing a handwriting sketch that includes object shape and motion. The proposed system matches the query and target video by using a histogram of the relative positions of shape edges and the trajectory of moving objects. Experimental results showed that simple videos can be fined by this proposed system.

1 Introduction

the popularity of making videos has increased, because of the proliferation of digital cameras and cell phones with a video cameras. On the web, there are many new sites for sharing these personal videos. There are many videos with varied content. However it is not easy to search for and retrieve specific desired scenes among this rapidly growing accumulation of videos. Text search is the general technique used by most search services, but it is difficult to find a scene based on keywords. It is possible to use the names of objects in a video as keywords, but it is not possible to represent the content of a video by using only these keywords. Therefore, a search technique of using visual features of a video (not keywords) and a system that can use these features as a query are necessary.

There has been much research on content-based image retrieval [1, 2]. However, there has been little research on video retrieval. Video can be thought of as an application of image retrieval. Therefore, although, some features of image retrieval can be used video retrieval, video retrieval is more difficult about representing queries.

In this paper, we proposed a video retrieval system that uses handwriting sketches for a queries. Drawing a picture makes it possible to create a more precise query than using keywords, when the user imagines and draws a scene he wants. The user can include visual features of the desired video by using a handwriting sketch as a query.

2 Related Work

Some researches has proposed ideas using visual features [3, 4]. Mikami [3] focused on appearance-based scene retrieval. In this study, a user indicated a camera motion in a prepared 3D environment model to create a query of line history image (LHI) as a feature. This system only works in video of a determinate environment such as a soccer field, because a 3D model of environment must be prepared. Therefore it is difficult to search within varied personal videos.

Fujisaki [4] focused on the user interface and proposed a video retrieval system using motion and color histograms as features. In this study, features of a query of a video scene were selected from features shown in the system. A user creates a query by selecting features iteratively. However, with this type of indirect operation, it is not easy for a user to create the query he wants.

3 Proposed Approach

3.1 System Overview

Our proposed system uses information about the shape and trajectory of a moving object in a video to search for a desired video. Figure 1 is an overview. First, the system extracts two features from each video and registers them in a feature database. Then, a region around a moving object is estimated and the feature extraction process is applied to this region. The user inputs a handwriting sketch as a query. In this sketch, a shape is represented by a line drawing and its trajectory is represented by an arrow. Finally, the system compares features of the sketched query with videos registered in the database and displays the results of based on comparisons of degrees of similarity. In the following subsections, we describe the details.

3.2 Region Estimation of Moving Object

A moving object's region is estimated by using moving vectors. Moving vectors in each frame image are calculated by Block Matching. Moving vectors in the region of a moving object are different from vectors in a background region. In fact, if the vectors in either region can be estimated, each region can be divided. Thus, we estimate the vectors of the background region.

Cammera movement determines these background vectors, so we use a robust estimate (the least median of squares method) to estimate camera movement vectors block matching vectors. We suppose camera movement can be approximated by affine transform, and calculate the transformation parameter a (1). Pa-

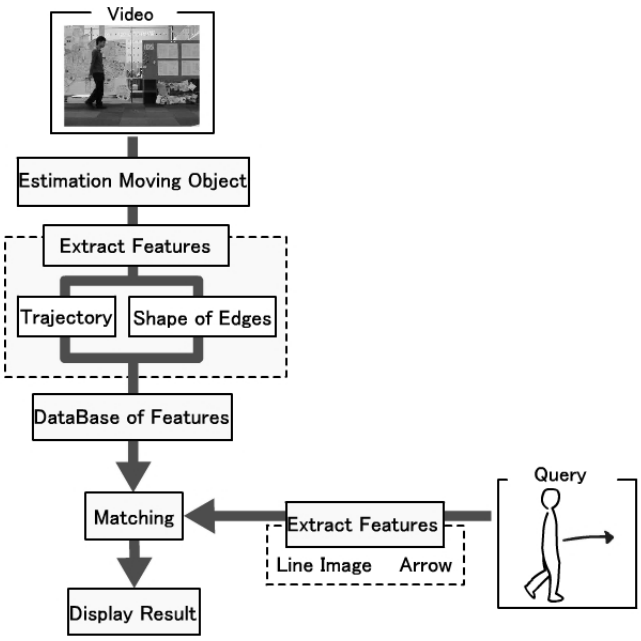

Figure 1: Overview

rameter estimation needs three vectors.

$$\begin{aligned} u &= a_0 x + a_1 y + a_2 \\ v &= a_3 x + a_4 y + a_5 \end{aligned} \tag{1}$$

After some repetition of estimates from three random vectors, the parameter that best minimize the equation(3) is selected to represent camera movement.

$$r_{i,j} = |q_{i,j} - f(a, p_{i,j})| \tag{2}$$

$$med = \text{ median}\{r_{i,j}^2\} \tag{3}$$

where $p_{i,j}$ is the start coordinate of the block matching vector, $q_{i,j}$ is the end coordinate and $f(a, p_{i,j})$ is the end coordinate of camera movement vector calculated from parameter a.

Next, the region of a moving object is divided by using the following three conditions [5].

- In a series of two-frame images, the differences in brightness exceed the threshold in a block region of same location on two frames.
- The directions and lengths of vectors calculated by block matching are different from those of the vectors calculated by affine transformation parameter.
- There are vectors calculated by block matching that have the same direction around.

After these, isolated blocks are deleted. Figure2 shows the results of this process.

3.3 Extract Two Features.

Two identifiable video features are shape and trajectory of a moving object in the video. Each feature

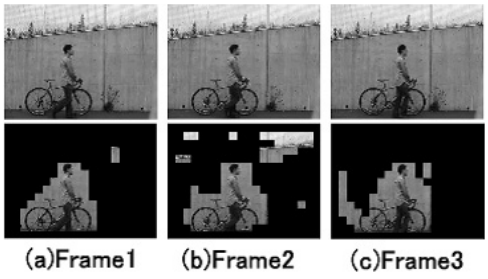

Figure 2: Region of Moving Object

is extracted from the region of a moving object. In this subsection, we describe each of the two features.

First, we will describe the shape feature. To define the shape feature, we use the shape of edges feature[6]. The relative position of the edge pixels can be represented on a 256-dimensional histogram. This feature is shift- and scale-invariant. The same shape feature is extracted from all frame images. The shape feature of a video is a set of all of the frames in the video. Now, a shape feature of frame i is defined as Fm_i and the video feature is defined as Fm. The shape feature of a video can be expressed as (4).

$$Fm = \{Fm_1, Fm_2, ..., Fm_i\} \tag{4}$$

where i is the number of frames.

Second, we will describe the trajectory feature. We suppose the region of a moving object has coordinated moving vectors. So, in each frame, the average of vectors in the region of the moving object is calculated as a single vector. However, since camera movement affects these, difference vectors are calculated from getting the difference between camera movement vectors and moving object vectors. These vectors are used to calculate average vectors. These average vectors of each frame are consolidated and represented as a time-series vector, which represents the trajectory feature. Figure 3 shows a frame format of this process. Now, a average vector of frame i is defined as Vm_i and the video feature is defined as Vm. The trajectory feature of a video can be expressed as (5).

$$Vm = \{Vm_1, Vm_2, ..., Vm_i\} \tag{5}$$

where i is the number of frames.

Finally, Fm and Vm are saved in the database as a video feature.

3.4 Matching Process

In this subsection, we describe the process of matching features of input sketches to feature of videos. When a sketch query is input as in Figure 4, the system separates the sketch into a line image and an arrow, and compares each feature of the query with the feature of a video.

First, we explain the matching of the line image part of the sketch. The shape of the edges feature Fq is also calculated from the line image. Fq is compared with a video feature Fm in the database by calculating

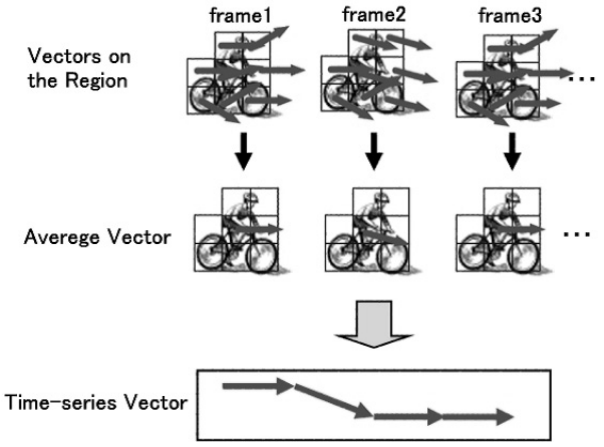

Figure 3: Process of Extract Time-Series Vector

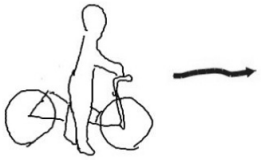

Figure 4: Example of a query

the Euclidean distance. If there are k 256-dimensional features per video, then distance D_F is the shortest one.

$$D_F = \min_{k=1,2,...,K} \left\{ \sqrt{\sum_{j=0}^{255} (Fq_j - Fm_{kj})^2} \right\} \quad (6)$$

Next, we describe the matching of the arrow on the sketch. The arrow drawn by the user is divided into several vectors at constant length such as Figure 5. In the same way as the trajectory feature, the divided vectors are considered as a time-series vector, so the time-series vector of the arrow is defined as Vq. The distance D_V is calculated by using DP matching between Vq and Vm. In this case, two vectors are compared by the angle of the vectors such as in expression (7). A correspondence relation of the two vectors is defined as $C_n = (i, j)$, and the warping path is defined as $Warp = \{C_1, C_2, ..., C_n\}$.

$$d(i,j) = 1 - \frac{Vm_i \cdot Vq_j}{\mid Vm_i \mid\mid Vq_j \mid} \quad (7)$$

$$e(C_n) = d(i,j) + \min \left\{ \begin{array}{c} d(i,j-1) \\ d(i-1,j-1) \\ d(i-1,j) \end{array} \right\} \quad (8)$$

$$d(0,0) = 0, d(i,0) = d(0,j) = \infty \quad (9)$$

$$D_V = \frac{1}{N} \sum_{n=0}^{N} e(C_n) \quad (10)$$

where n is the number of correspondence relations of two vectors.

In DP matching, the result of matching between Vq and Vm is calculated to minimize the expression (10).

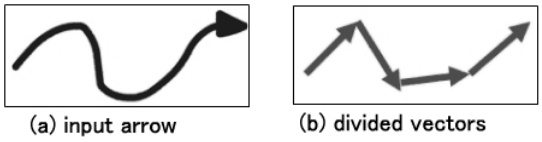

Figure 5: Divide of Arrow

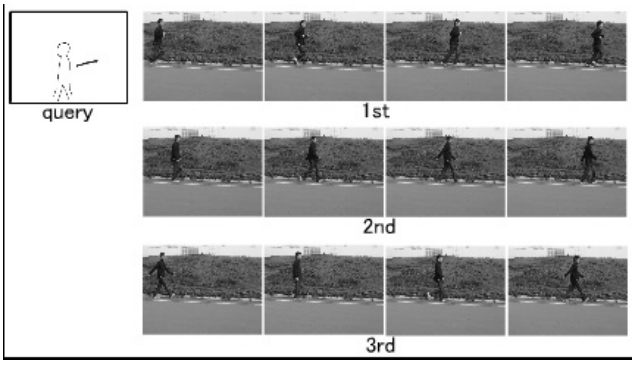

Figure 6: Results of Test Video

Finally, the distance D_M is calculated to sum D_F and D_V. D_M is the final difference between a query and a video. In this instance, D_F multiplied by 0.5 as a result of trial and error. The smaller the difference, the higher the similarity.

$$D_M = 0.5 * D_F + D_V \quad (11)$$

4 Experiment

We experimented with video retrieval using our proposed system. The first experiment was conducted using fourteen test videos we had taken. These videos included scenes of people walking horizontally, walking while pushing a bicycle, walking while pushing a chair, and turning while walking. All were taken with the camera in a fixed position. The results are shown in Figure 6. Videos which were similar to the sketch in both shape and trajectory features are at the top. In this experiment, the results were relatively good.

Second, we described the experiment using videos on YouTube [7]. We collected twenty-two videos from YouTube and extracted scenes from these videos to use in this experiment. After watching a target video, a user searched for it by sketching a query. This experiment was conducted with three subjects. Figure 7 and Figure 8 show the examples of good result. In this experiment, the average of rank of all subjects results is 3.227. This means that the video that the user wants is included in about higher 14 % of results.

5 Discussion

In the experiment using our test videos, the results were relatively good. In the first experiment, there are some videos that have the similar feature of edges. For example, in Figure6, there are two similar videos

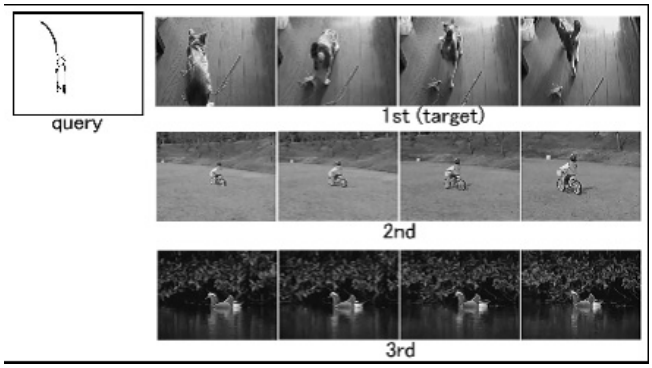

Figure 7: Results of YouTube Video (1)

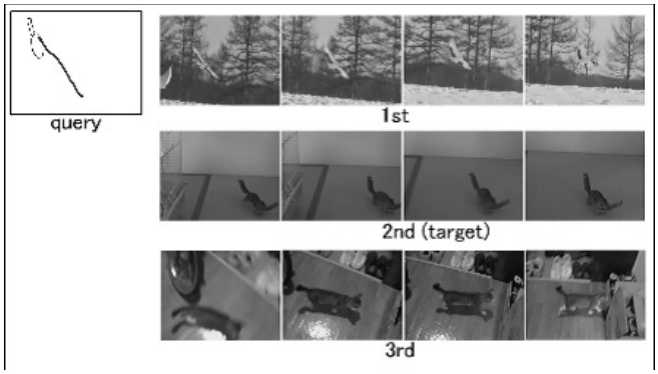

Figure 8: Results of YouTube Video (2)

of the target that included scenes of people walking from right to left and turning while walking. Table1 shows the difference of each features of three videos. It is appeared that the feature of edges is very similar among the three videos, but feature of trajectory is different clearly. Therefore, it is possible to distinguish the target videos. It is thought that the objects that has the similar shape can be distinguish by using two features.

Table 1: Difference of Each Features

	D_M	D_F	D_V
left to right (Fig.6 1st)	0.1232	0.0899	0.0331
turing	0.8849	0.0915	0.7933
right to left	2.0149	0.0949	1.9200

In the second experiment, the average of rank of all subjects results is 3.227. We think that this result can be considered reasonable as the result of a retrieval system. In this experiment, there are two groups of video. The videos of one group was retrieved at the relatively higher rank when each subjects retrieved. The videos of another group was not retrieved at the relatively higher rank. The reasons may be the following two.

1. The movement that included in the video (the movement of the object and the camera movement) is simple. Therefor, a noise is hard to be included and the system calculates the feature with high accuracy, because the blur is less. And because the user sketches the videos simply, it is easy that the feature of user's query corresponds with the feature that the system calculated. Therefor, the video is retrieved at the relatively higher rank.
2. The movement that included in the video is hard and complex. If the movement is hard, the video includes many blurring scenes. Therefore, the correct feature is not calculated. If the camera movement is complex, the regions of the moving object were not estimated with high accuracy, because the camera movement is approximated by affine transform in the proposed system. In addition, it is not easy that the feature of user's query corresponds with the feature that the system calculated in the complex movement. Therefor, the video is not retrieved at the relatively higher rank.

6 Summary

In this paper, we proposed a video retrieval system that uses handwriting sketches as queries. We conducted experiments using controlled test videos and videos from YouTube. In this study, it was shown that it is possible to retrieve the video by using handwriting sketch. And it is also possible to retrieve from the YouTube videos under the regulated condition. But there are many problems that needed to solve. The proposed system can use only limited videos, and this system is not so robust.

We will continue to work on the following goals to improve our system's performance.

- To improve the accuracy of the process of estimating the region of the moving object
- To recognize this system and to experiment on more large video database
- To consider the relation of video and the user's mental picture

References

[1] M. Matuzaki, M. Kasimura and S. Ozawa: "Image Retrieval for Pictures Based on Feature Graph Using Simplified Image," *The transaction of the Institute of Electronics, Information and Communication Engineers. D-II*, vol.J87-D-II, No.2, pp.521-533, 2004.

[2] H. Takahiro, I. Atsushi and O. Rikio: "Retrieval of Landscape Images Using a User-drawn Sketch and Semantic Icons." *Technical report of IEICE. PRMU*, vol.104, No.573, pp.19-23, 2005.

[3] M. Dan, A. Shozo and M. Masashi: "An appearance-based video scene retrieval based on synthesized video query by target points indication and the line history image matching." *Techical report of IEICE. PRMU*, vol.106, No.606, pp.67-72, 2007.

[4] F. Yasuhiko, M. Takamichi, I. Yasuhiro, K. Aki, Y. Katunori and S. Yoshihiro: "Video Retrieval System with Query Generation Assistance Using Relevance Feedback." *IEICE technical report. Image engineering*, vol.105, No.610, pp.95-100, 2006.

[5] S. Kenichirou, N. Masaomi and S. Hitoshi: "Moving Object Detection from Video Sequence Using Local Motion Direction Histogram Features." *IEICE technical report. Neurocomputing*, vol.102, No.382, pp.47-52, 2002.

[6] O. Gosuke, N. Yasutake, M. Keita and S. Yoshifumi: "Edge-based Image Retrieval Using a Rough Sketch." *The Journal of the Institute of Image Information and Television Engineers*, Vol.56, No.4, pp.653-658, 2002.

[7] YouTube: http://www.youtube.com/