

3—18

## Proposal of Query by Short-time Action Descriptions in a Scene

Hisashi Miyamori†\*

Tomio Echigo‡

Shun-ichi Iisaku‡

†Communications Research Laboratory  
Ministry of Posts and Telecommunications‡Tokyo Research Laboratory  
IBM Research**Abstract**

This paper proposes a new video representation for content-based retrieval using short-time action descriptions. The conventional methods use feature values such as transformed coefficients, etc. in query matching, whereas the proposed method employs the short-time action description which corresponds to a set of temporally and spatially local 2D appearance images in a scene. The advantage of the proposed method is the ability of obtaining not only simple action patterns by a single person or object but also relatively complicated action patterns by several persons or objects in a scene. The simulation results applied to soccer scenes confirm that the proposed method improves the efficiency and flexibility in retrieving rather complicated actions, compared to the conventional methods.

**1 Introduction**

Recently, the amount of video information has been rapidly increasing in various fields. Visual databases, video browsing, and video surveillance are expected to play an important role as major applications in the field of academic education, medical sciences, and training for artists, dancers, and sports players, etc. in the future. In such an environment, indexing, retrieval, and summarization will be the key function to achieve efficient browsing, which means that the video annotation will become increasingly important. The standardization activity is also scheduled to begin as MPEG-7, which is mainly focused on video description method for content-based retrieval.

Previous approaches to video representation for content-based retrieval can be classified by retrieval key used for matching. Keyword-based approach is the most disseminated way of realizing content retrieval because of its ease of matching using an explicit word directly. However, it has an disadvantage of the human cost necessary to put indexes manually to the contents. So, this method is appropriate for retrieval by queries of outlines. Query-by-example approach[1]-[3] uses feature values such as

color, shape, texture, motion, and structural information, of a given example image as retrieval key, and can retrieve contents which are hard to describe explicitly in words. However, since the results of this query vary by the given example, most of the implementations of this method are limited to still image applications. As an extension to the queries of actions by humans and objects in a scene, there are retrieval methods using the time-varying symbol pattern obtained from a certain feature space[4]-[6]. These methods have been applied to human gestures or swinging actions in tennis. However, the target gestures have been limited to rather simple ones performed by a single person or object in a screen, and the extension to rather complicated actions by several persons or objects has not been achieved yet.

This paper proposes a new scene representation method which can describe the sequence composed of complicated patterns of actions by several persons/objects which successively change in time in a scene, as well as simple patterns of actions by single person/object. It is achieved by introducing a short-term action description as a unit to describe the whole context of the actions which last during a rather long term. It also discusses the automatic generation method of its scene representation data.

The rest of the paper is organized as follows; In section 2, the overview of the proposed method is introduced, and a few examples of the proposed method applied to soccer scenes are provided. In section 3, a full automatic generation method of the proposed representation data is tested with some discussions on the simulation results. The conclusion is summarized in section 4.

**2 Scene Representation by Short-time Action Descriptions**

This section proposes a query method using the descriptions of actions which occur in a short period of time, in order to retrieve complicated patterns of actions by several persons/objects which successively change in time. This section also verify the validity of the proposed method by applying it to the soccer scene.

\*Address: 4-2-1 Nukui-kitamachi, Koganei City, Tokyo 184-8795 Japan. E-mail: miya@crl.go.jp





description method	description unit	values of description unit	illustration <span style="float: right;">[time or space]</span>
appearance-based image method	whole action	feature symbols	
short-time action description method	short-time action	set of local feature symbols	
displacement description method	frame-based	physical values	
input video	frame-based	pixel values	

Figure 1: Classification of scene representation method and its data

## 2.1 Comparison with Conventional Methods

Figure 1 shows the comparison on the scene representation methods and their representation data.

The appearance-based method utilizes the characteristics of the object “appearance” in a scene, and is one of the popular methods often applied to gesture recognitions and pose estimation[4],[5]. The target gestures, however, have been limited to rather simple ones performed by a single person or object in a screen.

On the contrary, considering the applications like automatic monitoring and sports scene retrieval, the target gestures are required to cover rather complicated actions by several objects in a scene. For example, the actions of the players in a soccer game is thought to be linked with the keywords like “shoot - goal”, as a result of transition pattern of successive actions/states of each player and the ball.

The short-time action description method is the scene representation method using the descriptions of actions which occur in a short period of time, in order to retrieve complicated patterns of actions by several persons/objects which successively change in time. This method does not calculate nor match the feature symbols of the target action for the whole sequences as a whole, but calculate the basic action/state as a temporally-localized feature symbol and use its transition pattern for matching.

On the other hand, the displacement description method[6] uses frame-based physical values such as object position, length, velocity, etc. in the screen as indices, and retrieves temporal condition changes of object’s motion. Since the indices are low-level physical values, the method can be applied to a wide range of action patterns according to the definition of condition equations. However, since the indices from which the retrieval keys are generated are given on frame basis, the method essentially refers to huge amount of indices, which leads to a drawback in speeding-up of the processing performance.

The short-time action description method has an

advantage on the speeding up of the processing performance compared to the frame-based descriptions, because it describes local actions/positions based on temporal intervals.

If we suppose the input video including the target action patterns is the lowest level of scene representation method, the short-time action description method can be located between the appearance-based method and the displacement description method (figure 1).

## 2.2 Application to Sports Video

In this subsection, the proposed method is applied to the sports video. Our test domain is soccer game. Soccer is a game which does not consist of a single player, but comprises the team plays by several players. Thus, the integrated representation of the related players in a scene is indispensable for the description of the whole context of the game. Also, the representation of actions by each player is necessary to describe the context of the team plays. Furthermore, the transition of the situation needs to be represented in the description since the context of the game varies along time.

First, the state transition graph using short-time action descriptions is generated manually from the soccer sequence (figure 2).

In figure 2, the horizontal axis represents object ID, and the vertical axis shows time. Object ID consists of team name and ID number, whereas no ID number is given to two goal keepers and judges.  $(l, m)$  describes the lasting period of each state, and  $l, m$  denote the starting and ending time of the state respectively. As for  $[x, y]$ ,  $x$  shows object’s location ID, and  $y$  describes short-time action ID. Note that the soccer field is divided into clusters  $A-H$ , and the short-time action is composed of ten basic action categories beforehand.

First, take action category “heading - shoot” for example. This action category can be thought as firstly heading by a player around the goal and secondly overlooking or jumping or diving towards the

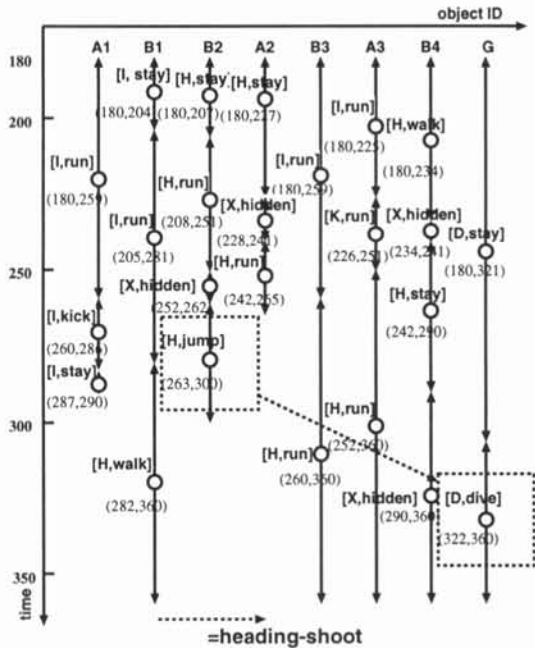


Figure 2: State transition graph of test sequence

ball by a goal keeper. Consider that the player of the ID=2 in B team did heading, and that the goal keeper dived towards the ball for instance. This can be described by the proposed method as;

$$\begin{aligned}
 & B2[H, jump] - G[D, dive] \\
 & \text{or} \\
 & B2[D, jump] - G[D, dive].
 \end{aligned}$$

Actually, there are 12 shoot scenes in the test sequence, among which 4 shoot scenes are done by heading. It is identified that all these scenes can be detected using the proposed method if retrieval is done so that arbitrary ID number is allowed as a player ID, and that these shoot-by-heading scenes can be discriminated from other shoot scenes such as “kicking - shoot”. On the other hand, the query method without using short-time action description has to use the location of each object and its transition for retrieval, while the simulation shows that the action category “heading - shoot” cannot be detected by the retrieval method using only location and its transition information.

Considering the action category “corner kicking”, the proposed method can distinguish its action from simple kicking by representing the action as  $[I, kick]$  or  $[E, kick]$ . However, the short-time action description without location information cannot tell this kind of difference. Thus, the usage of only short-time action ID is not enough to be applied to content-based retrieval.

Therefore, it follows that the retrieval method using only location information can obtain only the limited number of action categories, while the pro-



Figure 3: Short-time action descriptions can distinguish “heading-shoot” scenes from other shoot scenes

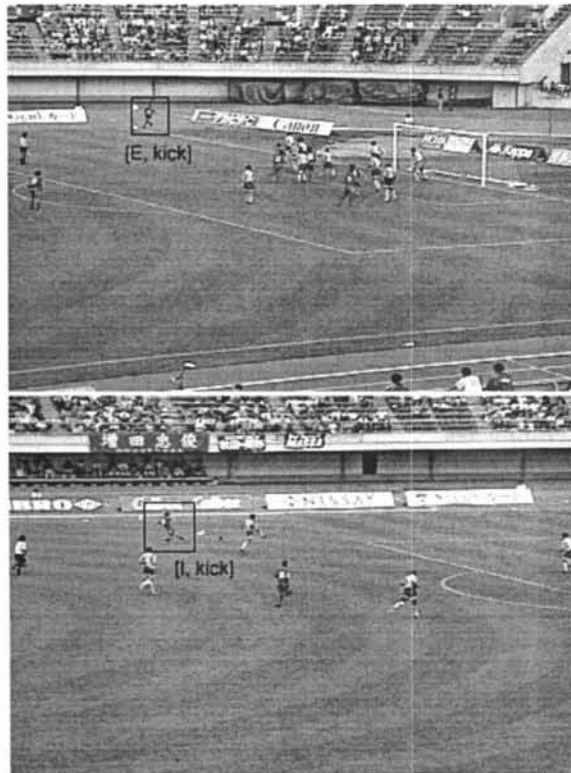


Figure 4: Location information is necessary to distinguish corner kicking (top) from simple kicking (bottom)

posed method can detect rather complicated action categories compared to the conventional method.

### 3 Automatic Generation Method of Scene Representation Data

This section tests on an automatic generation method of proposed scene representation data.

Although the automatic extraction method of contextual information from general video seems unlikely to be achieved with current vision applications, there are several approaches being well applied to a particular type of video utilizing a specific model created from the structure of the corresponding video scenes[7]-[10].

In this paper, the following domain knowledge and model matching approach is tested (figure 5).

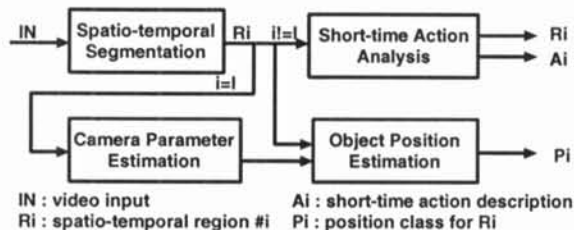


Figure 5: Generation diagram of scene representation data

First, the input image is divided into the background candidates and object regions by segmentation. Background candidates are defined as regions in which the number of the same motions in each segmented output is maximum, and the object regions are defined as the regions other than the background candidates. The segmentation employed here is spatio-temporal segmentation using color, texture, and motion[11].

Then, the relative parameters of camera motion are calculated using the displacement vectors of the background candidates. Each region is projected onto the same plane using the obtained parameters to generate a background mosaic. After that, the position of each object is estimated using the predefined clusters on the background field model. The output is the transition data of cluster IDs of the background.

Meanwhile, each object region is given to the analysis of short-time actions. The action analysis is done by generating the silhouette images of each object and judging its shape transition using eigen-space method and hidden Markov model[12]. The output of this analysis is the transition data of short-time action categories of each object.

The simulation results confirm that the limited action categories such as "run", "walk" and "stay" are successfully extracted to generate state transition indices. However, the indexing fails when players interact each other or swarmed round the ball, because the extracted silhouette images does not re-

fect the sufficient action characteristics of the players. The study of the robust feature analysis method of player's actions against the background remains as future work.

### 4 Conclusion

A new query method using short-time action description is proposed in order to obtain not only simple action patterns by a single person or object but also relatively complicated action patterns by several persons or objects in a scene. The application to other video materials and the improvement of automatic indexing method remain as future work.

### References

- [1] S.Ravela, R.Manmatha: "Retrieving images by similarity of visual appearance", In the Proc. of the IEEE Workshop on Content Based Access of Images and Video Databases, CAIVD'97, pp.67-74, June, 1997
- [2] M.Flickner, et. al.: "Query by image and video content: the QBIC system", IEEE Computer Magazine, pp.23-32, November, 1995
- [3] A.Nagasaka, Y.Tanaka: "Automatic video indexing and full-video search for object appearances", IPSJ Trans. Vol.33, No.4, pp.543-550, 1992
- [4] J.Yamato, J.Ohya, K.Ishii: "Recognizing human action in time-sequential images using hidden Markov model", CVPR, pp.379-385, 1992
- [5] T.Watanabe, C.W.Lee, A.Tsukamoto, M.Yachida: "Method of real-time gesture recognition for interactive system", ICPR, pp.473-477, 1996
- [6] S.Abe, Y.Tonomura: "Scene retrieval method using temporal condition changes", IEICE, Vol.J75-D-II, No.3, pp.512-519, 1992
- [7] S.S.Intille, A.F.Bobick: "Visual tracking using closed-worlds", MIT Media Laboratory Perceptual Computing Section Technical Report No.294, November, 1994
- [8] Y.Gong, L.T.Sin, C.H.Chuan, H.Zhang, M.Sakauchi: "Automatic parsing of TV soccer programs", Proc. Int'l Conf. on Multimedia Computing and Systems, pp.167-174, May, 1995
- [9] D.D.Saur, Y-P.Tan, S.R.Kulkarni, P.J.Ramadge: "Automated analysis and annotation of basketball video", Storage and Retrieval for Image and Video Databases V, SPIE-3022, pp.167-187, 1997
- [10] T.Kawashima, K.Yoshino, Y.Aoki: "Qualitative image analysis of group behavior", CVPR, pp.690-693, June, 1994
- [11] T.Echigo, H.Miyamori, S.Iisaku: "Object segmentation of the soccer video by using GMRF and optical flow", Meeting on Image Recognition and Understanding, MIRU'98, WP3-10, 1998
- [12] J.Maeda, T.Echigo, H.Miyamori, S.Iisaku: "Motion estimation and segmentation using continuous state hidden Markov models", Meeting on Image Recognition and Understanding, MIRU'98, TP3-04, 1998