November 26, 1999
RT0338
Multimedia 6 pages

# Research Report

## Video Enrichment: Retrieval and Enhanced Visualization based on Behaviors of Objects

Tomio Echigo, Masato Kurokawa, Alberto Tomita, Hisashi Miyamori, and Shun-ichi Iisaku

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

**Research Division**
**Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich**

# Video Enrichment: Retrieval and Enhanced Visualization based on Behaviors of Objects

**Tomio Echigo†, Masato Kurokawa†, Alberto Tomita†,**
**Hisashi Miyamori‡, and Shun-ichi Iisaku‡**

**†IBM Research**
**Tokyo Research Laboratory**
**{echigo, kurokawm, atomita}@jp.ibm.com,**

**‡Communications Research Laboratory**
**Ministry of Posts & Telecommunications**
**{miya, iisaku}@crl.go.jp**

## ABSTRACT

This paper presents a novel framework for video management called Video Enrichment. It is based on the description of objects that are extracted from the video image sequences and allows a semantical interpretation of the video contents, that is important to solve the problems inherent to content retrieval. The proposed technologies extract the objects by segmenting regions of the images, analyze the behavior of the objects and generate descriptions based on that behavior. Descriptions obtained through this analysis of the image sequences are stored using keywords and a pre-defined syntax based on a priori knowledge of a given domain. They can be searched and query results can be shown not only by retrieving the original sequence, but also by displaying a view from a different perspective, with the movements of the objects summarized for a short interval. This enhanced visualization makes possible for the user to explore new interests. In this way, the Video Enrichment framework provides tools for video management by repeatedly querying, searching, viewing and refining images sequences in the video database.

**Keywords:** video retrieval, object segmentation, MPEG-7, video mosaicing, knowledge discovery

## 1. INTRODUCTION

Digital video is being generated in ever-greater quantities, and is becoming pervasive in many emerging applications. Owing to the inherently large amount of data, it is very difficult to manage and access video sequences according to their contents in large video databases. Efficient ways of manipulating, annotating, searching, browsing, analyzing, and refining video data of interest are therefore important, and have attracted much recent attention.

Conventional video retrieval technologies require annotations attached by human, a scenario, closed caption, or speech and telop recognition[8]. The efficiency of the technical approaches depend on the contents, and the above information is not important for few annotated contents; that is, sports and surveillance. On the other hand, other methods by generic video processing; that is, color, texture, and shape extraction, have a limit to interpret videos semantically[1].

In this paper, we propose a new framework for managing video sequences, called Video Enrichment, which consists of two processes. One is the externalization process that is mining semantic information from binary data. The other is the augmentation process that provides the enhanced visualization of search results.

In the externalization process, firstly we segment regions that consist of moving objects and a stationary background, tracking the moving objects in a spatio-

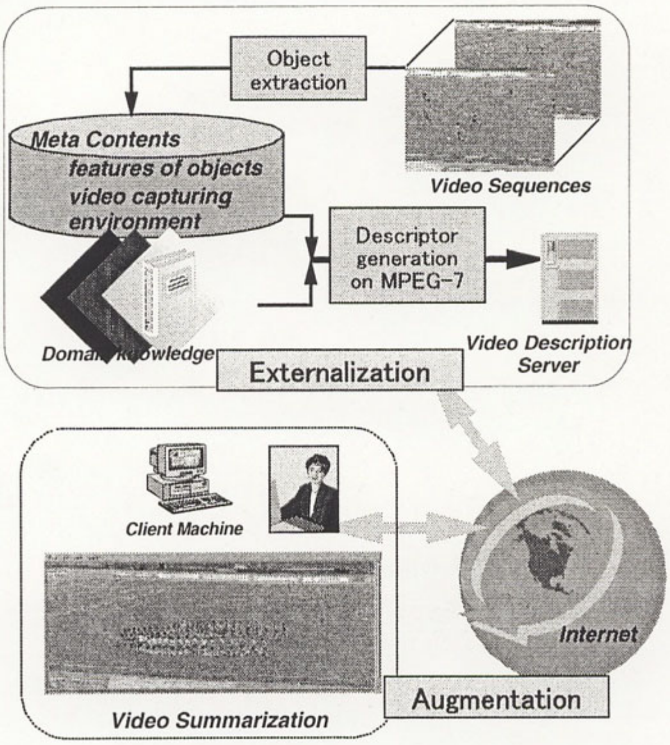


Fig. 1 The Video Enrichment Framework

temporal buffer over multiple frames. Secondly, motion parameters of the camera are estimated from motion vectors of feature points in the background region. These feature points are obtained from correspondence between frames. The coordinates in the image sequence can be transformed to the unified coordinates of a single reference plane by using parameters of camera motions. The position of an object on the image plane then can be transferred onto the real ground from a priori posture of the setting camera. The state of an object is classified into some primal motion categories from the speed of objects moving on the ground and from changing shapes in successive frames. Thirdly, descriptions of objects are generated by interval of object's behavior. The other descriptions that are equivalent to the results of semantic interpretation of an image sequence are also generated from evaluating the syntax defined from the beforehand knowledge on the specified domain.

In the augmentation process, an image sequence of the search result are shown in not only original sequence, but also a new view changed geometry and summarization of moving objects in a short interval. The enhanced visualization is possible to explore new interests for a user.

## 2. SPATIO-TEMPORAL REGION SEGMENTATION

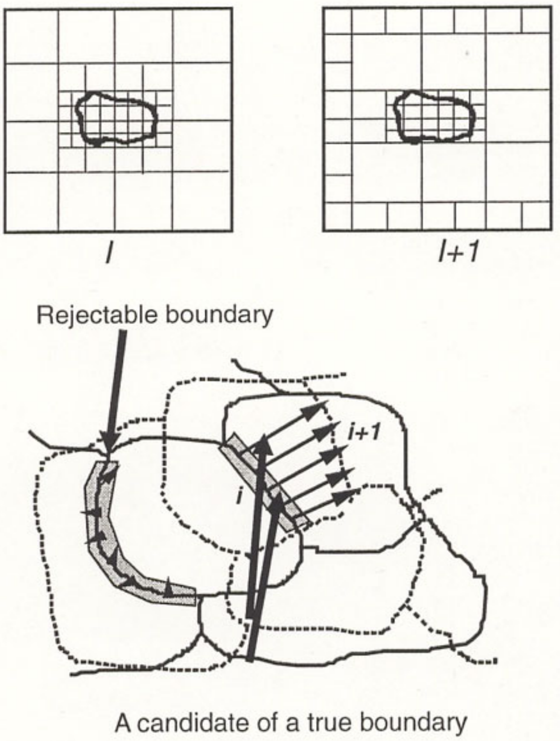


Fig. 2 Different configurations of quad-trees between successive frames

In order to solve the problem of segmenting and tracking an object, many conventional approaches use snakes and active tubes techniques of dynamic contour detection. They require a correct initialization, so that objects' initial contours include the objects themselves. However, it is not convenient to tie the initial contours to a particular frame of the video sequence.

The proposed technique segments regions and tracks them as predefined objects in a spatio-temporal buffer, where multiple frames are stored without initial contours.

First, all images in the buffer are spatially segmented by the split-and-merge method, which separates and merges regions on the basis of colors in the structure of a quad-tree. The shapes of the initial regions obtained by the split-and-merge method depend on the configurations of the quad-tree[2]. Our approach uses different configurations of quad-trees for each frame, that is, splitting home positions are different in successive frames. These images have different approaches of region growing, so that initial shapes of regions overlapped between successive frames are different from each other, as shown in Fig. 2. When the camera does not move and all objects are stationary, all pixel value differences between successive frames are nearly equal zero. In this case, the results of segmentation should depend on the spatial features.

Additionally, regions merged only by spatial features are not robust with respect to exact tracking over successive frames, because the gradient given by texture features often splits initial regions in overly small sizes, and motion is blurred by erosion of the shapes of the objects. Therefore, we use motion displacement along the boundaries calculated from spatio-temporal intensity gradients. When motion displacements are the same along a portion of the boundary that separates regions, that portion of the boundary can be considered as a part of the true contour of an object. Otherwise- that is, when motion displacements consist of disparate values along a candidate boundary- the candidate should be considered a false contour and be rejected[3]. A dominant motion of the camera movement is determined from corresponding features in the whole image between successive frames, as described in the next section. It can be judged from differentiation between successive frames shifted by a dominant motion whether a candidate boundary is rejected or not. All boundaries should be verified in the buffer of multiple frames not only between successive frames, but also between every third and every sixth frame. Regions inside the true contour can then be tracked as parts of the same object. To take account of occlusion, motion displacement of a region inside the contour is determined from its maximum value along the true contour.

In order to segment a player's region, adjacent regions that have close motion displacements should be merged into a single object, even if their colors and texture fea-

ID><Trajectory>

| Field | Type | Description |
|---|---|---|
| Action ID | Text | Action identifier |
| Time Interval | (ts, te) | Start and end time period |
| Object ID | Numerical | Object identifier |
| Trajectory | Time-stamped Polyline | Trajectory of the object |

The Action ID field identifies the object's action among the actions expected in the event domain. For example, in a soccer game these types could be "Kicking", "Jumping", "Diving", etc. The Time Interval field is a period defined by "start" and "end" times. The object ID corresponds to an object number assigned by the system. The Trajectory field is a time-series of two-dimensional coordinates (x, y, t), thus each point of the trajectory is composed by the (x, y) coordinates of a polyline node and a time-stamp relative to the trajectory starting point, i.e., the time elapsed until reaching the current position (by this definition, the position of an object in an arbitrary instant can be obtained by simple calculations).

### *4. 2. INTERACTION*

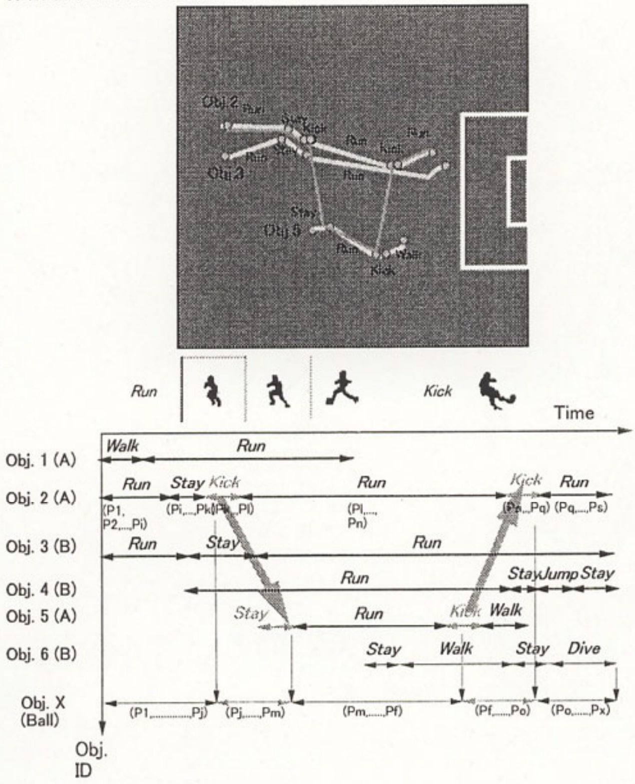


Fig. 3 The "Action" is described in an interval of an object's behavior with a trajectory on the pitch and the "Interaction" is involved the relationship of several "Action"s.

Next, an "Interaction" is built describing the meaning of a scene composed of several objects. Objects pictured in a scene can have different lifetimes and may be performing different actions, but their interaction is used to annotate a scene.

Fig. 3. represents the concept of objects in a time interval. (A) and (B) represents the teams, Obj. X is the ball. The annotation for the ball's movement is of an object without its motion identifier.

The definition of the "Interaction" representation is as follows:

Interaction ::= <Interaction ID><Time Interval><Object No><Object IDs>< Spatial Description >

| Field | Type | Description |
|---|---|---|
| Interaction ID | Text | Interaction identifier |
| Time Interval | (ts, te) | Start and end time period |
| Object No | Numerical | Number of objects |
| Object Ids | Numerical Array | Array of Object Ids |
| Spatial Description | Trajectory<br>Polyline<br>Polygon | Spatial description of this Interaction. (specified if necessary) |

The Interaction ID field identifies an event, such as "Pass", "Corner Kick", etc. These identifiers are completely domain-dependent. The next two fields indicate the number of objects involved in the Interaction been described, and their respective IDs. The Spatial Description field is used only when necessary. It shows the spatial positioning of an object or group of objects within a given Time Interval (in our prototype system, we use this field to describe the trajectory of the soccer ball).

The Interaction representation can also be used to describe more complex spatio-temporal relationships between objects. The Event Structure provides a set of logical functions, syntax and semantics to associate Actions and Interactions, assembling new Interaction representations. Instances of these newly created Interaction representations map the Action/Interaction associations they are composed of to events of higher complexity. For example, in a soccer game, a "Through Pass" is described by a "Pass" Interaction, combined with other Actions that satisfy a set of logical conditions

### *4. 3. EVENT STRUCTURE*

The Event Structure provides a scheme to describe events that comprise complex spatio-temporal relationships between objects. It allows Actions and Interactions tures are different, since the region may be a part of a face, an arm, a shirt, a pair of shorts, or a leg. We employ typical color map of players' regions of both teams that are cut out from a frame. Hence, the proposed algorithm is effective for segmenting and tracking predefined objects.

## 3. RECOVERY OF CAMERA MOTION AND CLASSIFICATION OF OBJECT'S BEHAVIOR

The position of a video object is usually expressed in image coordinates. However, as the camera moves, these coordinates have to be corrected in order to reflect an object's actual movements. For this purpose, the camera's movements are reconstructed and the resulting parameters are used to project the video images on a virtual plane. With the virtual plane as reference, it is possible to recalculate an object's position in a consistent coordinate system, using video obtained from a single camera. In addition, setting the virtual plane as if seen from a camera above the ground, it is possible to reconstruct an aerial view with the position of the objects projected on it, allowing the measurement of actual distances between objects in the ground instead of distances given in image pixels.

We employ a pin-hole camera model, that is generally used to model the camera perspective projection. The perspective model gives the exact representation to account for all the possible camera motions, compared with the other approximated models: the parallel transformation and the affine one (including para-perspective and weak perspective transformation), although its parameters are mathematically and computationally too hard to estimate. In our application, the camera is fixed on a tripod, giving images which have small translational displacements because the rotation axes does not coincide with the optical center of the camera (considering that the camera performs only rotational movements). However, the depth of the field of view is at least 40 meters far from the camera, so that the translation is much smaller than the scene depth. Therefore, our approach assumes a zero translation model between successive images of the video sequence.

Selecting image features in an image whose corresponding locations in consecutive images can be precisely measured is an important problem for estimating the exact parameters. In our approach, we employ Tan's method[6] that selects block features that have rich enough intensity textures and consistent inter-frame motions and finds correspondences between images. With this method, a quantitative measure can be obtained to select good motion features from images in the sense of the maximum likelihood estimation for estimating motion parameters about the multiple motion models of the rotation and scaling factors. In order to make estimation stable when a few data points are wrong, a robust estimation is also needed. We employ the M-estimation for robustness, which is applied with the Geman-McLure function[7]. Since M-estimation requires an initial estimate, firstly we use the initial estimate determined from the least squares method. Although the least squares method for estimating parameters has non-linear equation, we can assume that the rotational transformation should be small between successive frames, so that the initial estimate can be determined from linear equations. By using the initial estimate, the typical value of the standard residuals is defined as the median of the absolute residuals. The revised parameters are calculated by using the adjusted effective weight iteratively. When the residual becomes smaller than the threshold, the estimation process is finished.

Motion classification is performed based on changes in an object's shape through time. This is accomplished by discarding the color information inside the region of the image comprising the object, obtaining a silhouette. This silhouette changes as the object moves, generating patterns in eigenspace that characterizes given movements. Therefore, a motion can be recognized by matching a movement in eigenspace with previously recorded movement patterns. This process requires computing the changes in an object's continuous movement and mapping them to the eigenspace, however at the present development stage, these changes have been classified in three categories, that is, fast movement, slow movement, and stationary, based on the speed of objects moving on the ground.

## 4. OBJECT BASED DESCRIPTIONS

### *4. 1. ACTION*

It has been done by determining an object's motion during an interval, attributing a motion identifier and registering the frame numbers at the beginning and end of the movement[5]. Considering that an object performs the same motion in all frames within this interval, this data is inputted only at the boundaries, when the object changes to a different movement. This is done for all objects during their lifetime in the video. Thus, the essential description unit is the motion identifier. The description of an objects movements is called "Action", and the annotation comprises the motion identifier, start and end frames and the object's position observed through time (i.e. its trajectory, described as a series of discrete points in the time interval; the position of the object in an arbitrary point in time can be calculated by interpolating the points registered in the "Action")[4].

The definition of the "Action" representation is as follows:

Action ::= <Action ID><Time Interval><Object

of a set of objects to be combined with the use of domain-specific logical functions. Event Structures are processed when searching for a video for further retrieval. In this system, the scene database contains the descriptions defined by Action representations. Interactions are defined in the scheme provided by the Event Structures. Domain parameters represent the domain knowledge (such as the region of the “Goal area”, “Touch line”, etc.). A search is conducted by processing these Event Structures together with the domain-specific parameters. The retrieval engine searches for combinations of Actions and Interactions descriptions in the database, processing Event Structures. In the process, Actions and Interactions are assembled into new instances of Interaction representations, and those that satisfy the condition clauses of the search query are retrieved. The syntax to assemble Actions and Interactions in a Event Structure is:

```
Begin
Interaction Prototype
{Child Action}{Child Interaction}
where
{condition clause}
```

*The logical combination of the Domain-independent Function and Domain Support Functions which specify the necessary condition for defining this Interaction descriptors.*

```
fill
{value assignment}
end
```

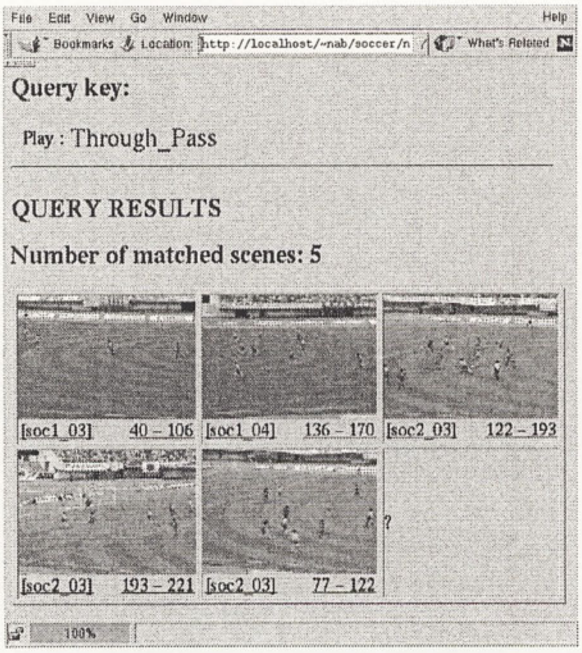
In the above syntax definition, "Domain-independent functions" is a set of functions that test generic spatial and temporal relationships. "Domain-Support Functions" is a set of logical functions that are necessary for describing or testing conditions that are domain-specific (for example, region definitions, team, and positions). For example, the “same-team(A,B)” function tests whether objects A and B belong to the same team.

Based on the combination of the two types of functions above, a user is able to specify in the condition clause the requirements a particular Interaction have to meet.

## 5. EXPERIMENTAL RESULTS

We have developed a prototype of the retrieval system based on the approach described in this paper. The data used in the experiments consist of 10 scenes from two professional soccer matches. Each scene contains between 300 and 1200 frames. The system was implemented on a Web environment, where user queries are issued on the client (Web browser) and the retrieval is performed on the server side. The server program was implemented on an RS/6000 system with AIX environment. Act representations involve 14 types of actions, such as Run, Walk, Kick, Jump, Stay, Sit, Hand-throw, Dive, Throw-in, etc. Almost 40 domain independent functions are provided to construct spatial and temporal relationships, and a few domain-support functions and domain-specific region definitions (such as "Goal Area", "Side Line", etc.). We have defined 20 types of Interactions, such as "Pass", "Shoot", "Centering", etc. The retrieval can be done upon queries to search scenes matching each Act and listed in a menu. Successful search and retrieval confirmed that the video sequences in the database can be described effectively by our approach. Fig. 4 shows an example with the retrieval results for the query of the Interaction "Through_Pass".

Search results from a query for a specific sequence can be displayed in a variety of formats. The simplest form is by playing back the original images of the sequence. In addition, images can be summarized in a short time by a mosaic image, which can display object's movements and changes in their position. Fig. 5 presents the realistic efficiency of the panoramic image in the wide angle from a long sequence of 141 frames and the stroboscopic painting on it. Fig. 6 shows the positions of the players when they score a goal. The circles surrounding each player illustrate the positions they can reach within 0.5 seconds, showing the advantages each player has from their location. Fig. 6 informs a user of the reason of scoring success an offensive player can use wide space near the goal apart from occupation of defensive players. With these visualization capabilities it is possible to capture the interest from the user, guiding him through the search-display-data analysis cycle.



Query key:

Play : Through_Pass

QUERY RESULTS

Number of matched scenes: 5

[soc1_03] 40 – 106 [soc1_04] 136 – 170 [soc2_03] 122 – 193

[soc2_03] 193 – 221 [soc2_03] 77 – 122

Fig. 4 Results of the retrieval for the query of "Through_Pass"

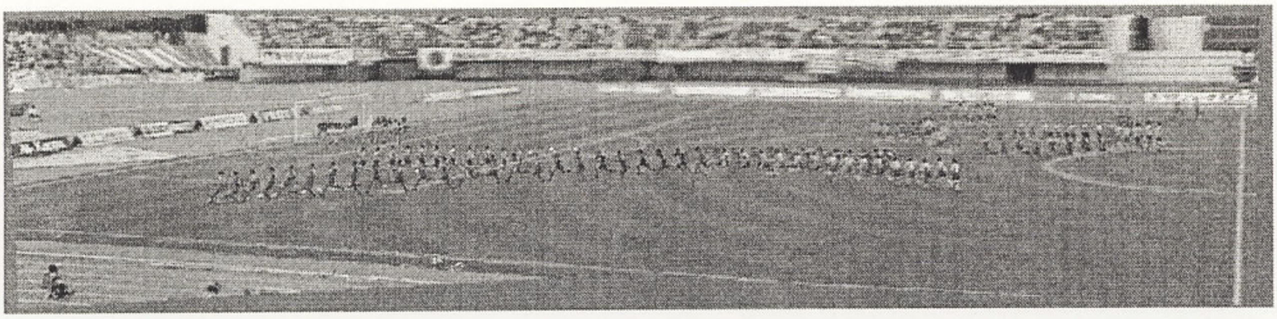


Fig. 5 Mosaiced background and superimposed moving objects

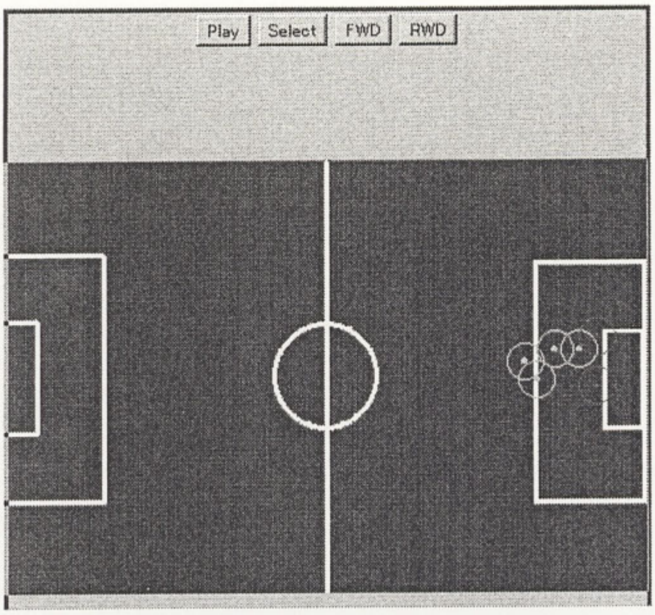


Fig. 6 A view changed geometry for analyzing a formation

## 6. CONCLUSIONS

We proposed a new framework for video management called Video Enrichment. It generated the descriptions based on behaviors of objects that are extracted from the video image sequences and allowed a semantical interpretation of the video contents, that is important to solve the problems inherent to content retrieval. The proposed description scheme will be contributed to the next generation international standard MPEG-7, aimed at applications of multimedia contents. A semantical interpretation is necessary to explore user's interests, so that we employed a priori knowledge on the specific domain in order to overcome the current limits of the generic video processing technologies. The domain knowledge was used for definition of objects and event structure. A subject for further study is the definition of the domain knowledge for covering many kinds of contents.

The Video Enrichment also provided tools for discovering and mapping knowledge of the contents by viewing and refining images sequences in the video database.

## References

[1] R. Bolle, B.L. Yeo, and M. Yeung, "Video Query: Beyond the Keywords," IBM Research Report, RC 29586, Oct. 1996.

[2] T. Echigo and S. Iisaku, "Unsupervised Segmentation of Colored Texture Images by Using Multiple GMRF Models and Hypothesis of Merging Primitives," Trans. IEICE D-II, vol. J81-D-II, no. 4, pp. 660-670, 1998

[3] T. Echigo, R. Radke, P. Ramadge, H. Miyamori, and S. Iisaku, "Ghost Error Elimination and Superimposition of Moving Objects in Video Mosaicing," IEEE ICIP-99, 28AO3.5, 1999.

[4] M. Kurokawa, T. Echigo, A. Tomita, J. Maeda, H. Miyamori, and S. Iisaku, "Representation and Retrieval of Video Scene by using Object Actions and Their Spatio-temporal Relationship," IEEE ICIP-99, 26AO2.1, 1999.

[5] H. Miyamori, T. Echigo, and S. Iisaku, "Proposal of Query by Short-time Action Descriptions in a Scene", IAPR Workshop on Machine Vision Applications, 3-18, pp.111-114, 1998.

[6] Y. P. Tan, ``Digital Video Analysis and Manipulation," PhD. Thesis, Princeton University, 1997

[7] H. S. Sawhney and S. Ayer, ``Compact Representations of Videos Through Dominant and Multiple Motion Estimation," IEEE PAMI, vol. 18, no. 8, pp. 814-830, 1996.

[8] H. Wactlar, T. Kanade, M. Smith, and S. Stevens,"Intelligent Access to Digital Video: The Informedia Project," IEEE Computer, vol. 29, no. 5, 1996.