PAPER *Special Issue on Machine Vision Applications*

A Method for Monitoring Activities of Multiple Objects by Using Stochastic Model

Nobuyoshi ENOMOTO[†], **Takeo KANADE**[††], **Hironobu FUJIYOSHI**[†††], *and* **Osamu HASEGAWA**[††††], ***Regular Members***

SUMMARY We present a method for estimating activities of multiple, interacting objects detected by a video surveillance system. The activities are described in a stochastic context because our method is concerned with humans and uses noisy features detected from video. To monitor activities in this context, we introduce the concept of an attribute set for each blob, consisting of object type, action, and interaction. Using probabilistic relations introduced by a specific Markov model of these attribute sets, the activity descriptions are estimated from surveillance video.

key words: *video surveillance,activity monitoring, Markov model*

1. Introduction

By newly introducing a method where activities of multiple, interacting objects are described in a stochastic model, we have realized a practical video surveillance system to monitor the activities mainly concerning with human actions.

Recent developments of low-cost video sensors and high-performance video processing hardwares made it possible to realize video surveillance systems. Surveillance cameras are already installed at many public facilities, such as banks, airports, stations and others. Video data, however, is commonly monitored and inspected by human operators, which costs very expensive. It is expected to implement automatic video understanding techniques which can not only detect moving objects but also extract unusual and meaningful events involving human activities.

Some automatic surveillance systems were already reported by Ohata et al. [1], and Lipton et al. [2]. In those systems, moving objects were automatically detected and classified to extract candidates of meaningful predetermined events. But activities involving object interactions, such as "A human entered a vehicle" could not be monitored to specify events in detail. These types of activities should be described as extended context to handle interactions between objects. And the context should not be deterministic but stochastic because human actions, which we mainly concern with, and their noise condition can be described stochastically.

To solve the problem of this sort, very recently Ivanov and Bobick [3]. and Oliver et al. [4]. have reported a new monitoring algorithm by parsing a SCFG and by using CHMM, respectively. In the first paper, Ivanov et al. have set some probabilities for the parsing which were not based on real observed data but were decided manually by experienced operators. It is very difficult to set those probabilities because an activity for a same event could be observed differently depending on position of camera, angle, etc. In the latter one, Oliver et al. generated training data by a multi-agent computer graphics simulator whose parameters were set by hand to account for situation-specific tuning against small numbers of training examples. But it seems very difficult to specify training data because their contexts in those systems are described by hidden inner-states, and not clearly predetermined.

To overcome these problems for practical applications, we have newly implemented a stochastic model for activities based surveillance system. In this paper, Sect. 2 introduces a new concept called "attribute set" to model and monitor activities mainly concerning with human actions,involving their interactions. Section 3 describes how to estimate parameters of the model from real training data. Section 4 describes an implementation of our method in the CMU VSAM test-bed system [5]. Experimental results for performance of our method are presented in Sect. 5.

2. Stochastic Estimation of Activities: Problem Definition

In a video surveillance system, we can detect, track and classify objects. Our goal is to form a stochastic representation of activities involving object interactions in data-driven manner by the system. Moreover, the context should be explicitly represented to allow for situation-specific tuning. To solve these problems, we

Manuscript received March 2, 2001.
Manuscript revised June 25, 2001.

[†]The author is with e-SOLUTIONS COMPANY, TOSHIBA CORPORATION, Kawasaki-shi, 212–8501 Japan.
[††]The author is with the Robotics Institute, Carnegie Mellon University, 5000 Forves Avenue, Pittsburgh, PA15213, U.S.A.
[†††]The author is with the Department of Computer Science, Chubu University, Kasugai-shi, 487–8509 Japan.
[††††]The author is with Neuroscience Research Institute, Advanced Industrial Science and Technology, Tsukuba-shi, 305–8568 Japan.

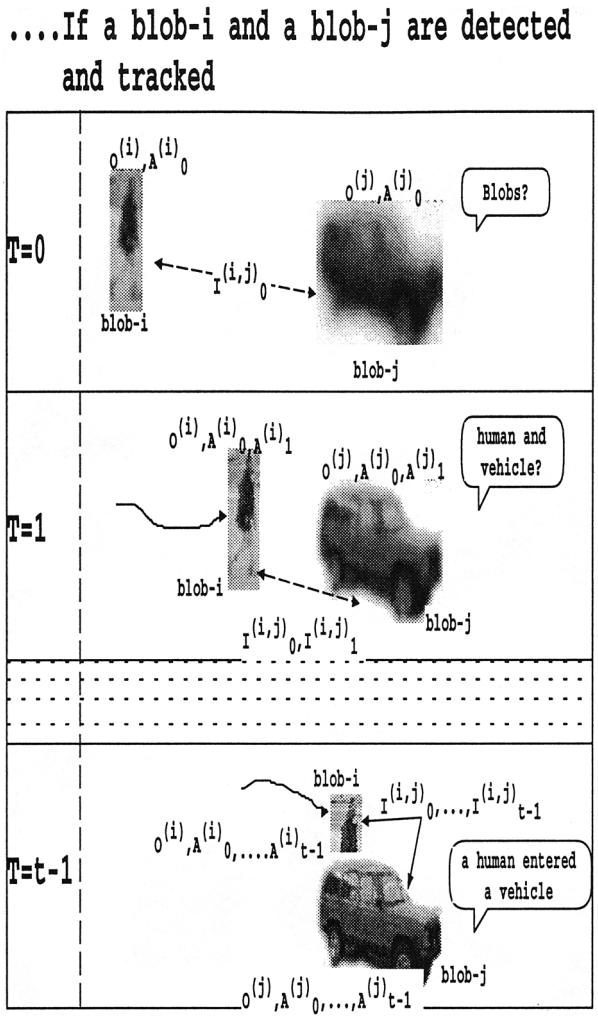


Fig. 1 Basic concept of activity description.

introduce the concept of an attribute set for each blob, consisting of object type, action, and interaction.

Figure 1 shows a basic concept of activity description based on the attribute sets. If a blob i and a blob j are detected in a frame and can be tracked for several frames, each trajectory consists of blob sequence $B_0^{(i)}, \ldots B_{t-1}^{(i)}, B_t^{(i)}$ and $B_0^{(j)}, \ldots B_{t-1}^{(j)}, B_t^{(j)}$. The sequence of blob i is considered as the observation of a real object which has a sequence of attribute sets of object-type $O^{(i)}$, actions $A_0^{(i)}, \ldots A_{t-1}^{(i)}, A_t^{(i)}$ and interactions $I_0^{(i,j)}, \ldots I_{t-1}^{(i,j)}, I_t^{(i,j)}$ between blobs i and j, frame by frame.

Our final goal is to obtain the most reliable description of the activity of each object. The object type $O^{(i)}, O^{(j)}$, action sequence $A_0^{(i)}, \ldots, A_t^{(i)}, A_0^{(j)}, \ldots, A_t^{(j)}$, and the interaction sequence $I_0^{(i,j)}, \ldots, I_t^{(i,j)}$ for each blob's sequence should be estimated by maximizing the conditional joint probability which is given as follows.

$$\widehat{\left(S^{(i,j)}\right)_0^t} = argmax_{\left(S^{(i,j)}\right)_0^t} P\left(\left(S^{(i,j)}\right)_0^t \mid \left(B^{(i)}\right)_0^t, \left(B^{(j)}\right)_0^t\right) \quad (1)$$

where

$$\left(S^{(i,j)}\right)_0^t \equiv \left\{O^{(i)}, O^{(j)}, \left(A^{(i)}\right)_0^t, \left(A^{(j)}\right)_0^t, \left(I^{(i,j)}\right)_0^t\right\},$$
$$\left(A^{(i)}\right)_0^t \equiv \left\{A_0^{(i)}, \ldots A_{t-1}^{(i)}, A_t^{(i)}\right\},$$
$$\left(I^{(i,j)}\right)_0^t \equiv \left\{I_0^{(i,j)}, \ldots I_{t-1}^{(i,j)}, I_t^{(i,j)}\right\},$$
$$\left(B^{(i)}\right)_0^t \equiv \left\{B_0^{(i)}, \ldots B_{t-1}^{(i)}, B_t^{(i)}\right\}.$$

$O^{(i)}$,$O^{(j)}$,$A_t^{(i)}$,$A_t^{(j)}$, $I_t^{(i,j)}$,$B_t^{(i)}$,$B_t^{(j)}$ are variables with states for the object types $o = o0, o1, \ldots$, actions $a = a0, a1, \ldots$, interactions $i = i0, i1, \ldots$, and observations $b = b0, b1, \ldots$.

3. Markov Model for Selecting the Most Probable Attribute Sequence

Consider a two frame sequence. If only two blobs, blob-i and blob-j, are in the scene, the conditional joint probability governing sequences of observations of these blobs is described as below.

$$\begin{aligned} &P(O^{(i)}, O^{(j)}, A_1^{(i)}, A_1^{(j)}, A_0^{(i)}, A_0^{(j)}, I_1^{(i,j)}, I_0^{(i,j)} \\ &\qquad \mid B_1^{(i)}, B_1^{(j)}, B_0^{(i)}, B_0^{(j)}) \\ &= P(B_1^{(i)}, B_1^{(j)}, A_1^{(i)}, A_1^{(j)}, I_1^{(i,j)} \\ &\qquad \mid O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}) \\ &\quad \cdot \frac{1}{P\left(B_1^{(i)}, B_1^{(j)} \mid B_0^{(i)}, B_0^{(j)}\right)} \\ &\quad \cdot P\left(O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)} \mid B_0^{(i)}, B_0^{(j)}\right) \end{aligned} \quad (2)$$

Using Bayes Rule, the first term on the right hand side of Eq. (2) becomes

$$\begin{aligned} &P(B_1^{(i)}, B_1^{(j)}, A_1^{(i)}, A_1^{(j)}, I_1^{(i,j)} \\ &\qquad \mid O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}) \\ &\cdot \frac{1}{P\left(B_1^{(i)}, B_1^{(j)} \mid B_0^{(i)}, B_0^{(j)}\right)} \\ &= P(B_1^{(i)}, B_1^{(j)} \\ &\quad \mid O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_0^{(i)}, A_0^{(j)}, A_1^{(i)}, A_1^{(j)}, I_1^{(i,j)}, I_0^{(i,j)}) \end{aligned}$$

$$\cdot\frac{1}{P\left(B_1^{(i)}, B_1^{(j)} \mid B_0^{(i)}, B_0^{(j)}\right)}$$

$$\cdot P(I_1^{(i,j)} \mid O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_1^{(i)}, A_1^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)})$$

$$\cdot P(A_1^{(i)}, A_1^{(j)} \mid O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}) \quad (3)$$

Equation (2) means that the conditional probabilities for times $t = 0$ to $t = t'$ can be described by using the conditional probabilities for $t = 0$ to $t = t'-1$ recursively. Practically, if t' is large, all of these conditional probabilities from $t = 0$ to $t = t' - 1$ can't be used.

To overcome this, we make the following assumption:

- $B_t^{(i)}$,$A_t^{(i)}$,$I_t^{(i)}$ are not decided by $B_{t-1}^{(i)}, \ldots, B_0^{(i)}$ because the objects with type $O^{(i)}$, action $A_t^{(i)}, A_{t-1}^{(i)} \ldots$, and interaction $I_t^{(i,j)}, I_{t-1}^{(i,j)} \ldots$, output the observed feature sequence $B_t^{(i)}$.

Equation (3) is then rewritten using this assumption, as

$$P(B_1^{(i)}, B_1^{(j)}, A_1^{(i)}, A_1^{(j)}, I_1^{(i,j)} \mid O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)})$$

$$\cdot\frac{1}{P\left(B_1^{(i)}, B_1^{(j)}\right)}$$

$$= P(B_1^{(i)}, B_1^{(j)} \mid O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, A_1^{(i)}, A_1^{(j)}, I_1^{(i,j)}, I_0^{(i,j)})$$

$$\cdot\frac{1}{P\left(B_1^{(i)}, B_1^{(j)}\right)}$$

$$\cdot P(I_1^{(i,j)} \mid O^{(i)}, O^{(j)}, A_1^{(i)}, A_1^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)})$$

$$\cdot P(A_1^{(i)}, A_1^{(j)} \mid O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}) \quad (4)$$

As we mainly concerned with the activity of humans, who can change their actions nondeterministically, we assume the attributes sets and the observations follow a 1st-order Markov model, and based on this, we can introduce the following assumptions.

- $I_t^{(i,j)}$ is only dependent on $O^{(i)}$,$A_t^{(i)}$, $O^{(j)}$,$A_t^{(j)}$,
- $A_t^{(i)}$ is only dependent on $O^{(i)}$,$A_{t-1}^{(i)}$, and $I_{t-1}^{(i,j)}$,
- $B_t^{(i)}$ is only dependent on $O^{(i)}$,$A_t^{(i)}$ and $I_t^{(i,j)}$.

Note that not all objects in a surveillance scene follow this assumption. For example, if observations of an object contain considerable error for some frames, the assumption will break. Nevertheless, using these assumptions and Eq. (4), the conditional joint probability of Eq. (2) is described generally as,

$$P(O^{(i)}, O^{(j)}, \left(A^{(i)}\right)_0^t, \left(A^{(j)}\right)_0^t, \left(I^{(i,j)}\right)_0^t$$

$$\mid \left(B^{(i)}\right)_0^t, \left(B^{(j)}\right)_0^t)$$

$$= \frac{P(B_t^{(i)} \mid O^{(i)}, A_t^{(i)}, I_t^{(i,j)}) \cdot P(B_t^{(j)} \mid O^{(j)}, A_t^{(j)}, I_t^{(i,j)})}{P\left(B_t^{(i)}, B_t^{(j)}\right)}$$

$$\cdot P\left(I_t^{(i,j)} \mid O^{(i)}, O^{(j)}, A_t^{(i)}, A_t^{(j)}\right)$$

$$\cdot P\left(A_t^{(i)} \mid O^{(i)}, A_{t-1}^{(i)}, I_{t-1}^{(i,j)}\right)$$

$$\cdot P\left(A_t^{(j)} \mid O^{(j)}, A_{t-1}^{(j)}, I_{t-1}^{(i,j)}\right)$$

$$\cdot P(O^{(i)}, O^{(j)}, \left(A^{(i)}\right)_0^{t-1}, \left(A^{(j)}\right)_0^{t-1}, \left(I^{(i,j)}\right)_0^{t-1}$$

$$\mid \left(B^{(i)}\right)_0^{t-1}, \left(B^{(j)}\right)_0^{t-1}) \quad (5)$$

Note, the conditional probabilities for $t = t$ can be described only by states and observations for $t = t$ and $t = t - 1$.

In Eq. (5), $t = 0$ corresponds to the initial state of an object's activity and is written as,

$$P\left(O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)} \mid B_0^{(i)}, B_0^{(j)}\right)$$

$$= P\left(B_0^{(i)} \mid O^{(i)}, A_0^{(i)}, I_0^{(i,j)}\right)$$

$$\cdot P\left(B_0^{(j)} \mid O^{(j)}, A_0^{(j)}, I_0^{(i,j)}\right)$$

$$\cdot\frac{P\left(O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}\right)}{P(B_0^{(i)}, B_0^{(j)})} \quad (6)$$

To calculate these equations, we need tables for the conditional probabilities $P(B_t^{(i)} \mid O^{(i)}, A_t^{(i)}, I_t^{(i,j)})$, $P(A_t^{(i)} \mid O^{(i)}, A_{t-1}^{(i)}, I_{t-1}^{(i,j)})$, $P(I_t^{(i,j)} \mid O^{(i)}, O^{(j)}, A_t^{(i)}, A_t^{(j)})$, and joint probability $P(O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)})$. The tables for the a priori probabilities can be obtained by counting events for each attribute set in training samples consisting of image sequences showing various types of activity.

The path that maximizes the probabilities described in Eq. (1) can be obtained by calculating Eq. (5) through the trellis diagram in Fig. 2. In this diagram, s1, s2 ... are state labels, and v, h, hg are labels of object-type. AP(i), MOVE(i), ... mean that blob-i has an action label "AP," "MOVE," ... and NEAR(i,j), ... means that blob-i and blob-j have an interaction label "NEAR," ... in a state. If we have a model (trellis diagram) for each activity, the most desirable description is obtained by selecting a model having the maximal posterior conditional joint probability.

To describe more than two blobs detected in a surveillance scene, we can generate a model for each

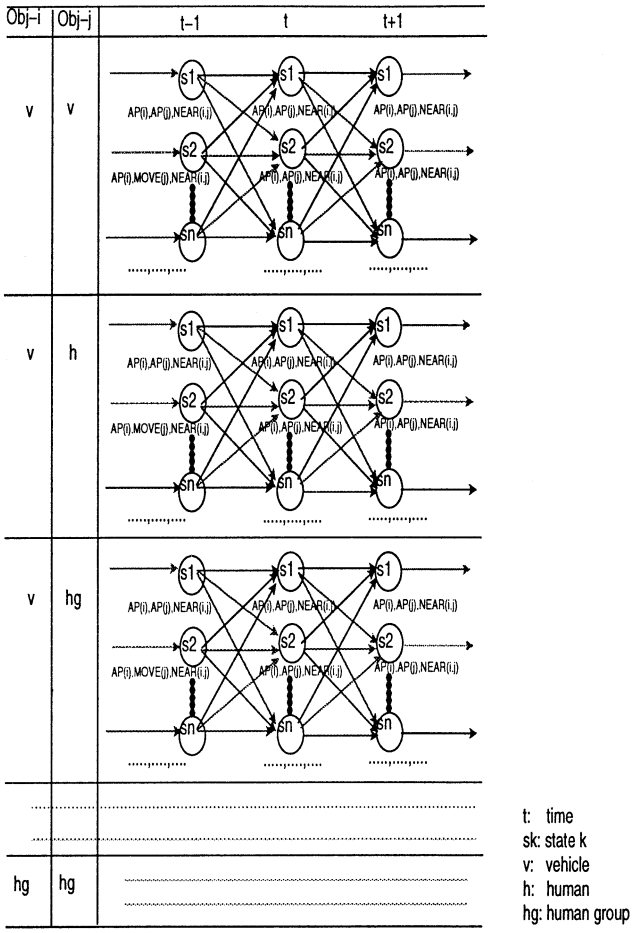


Fig. 2 Trellis diagram.

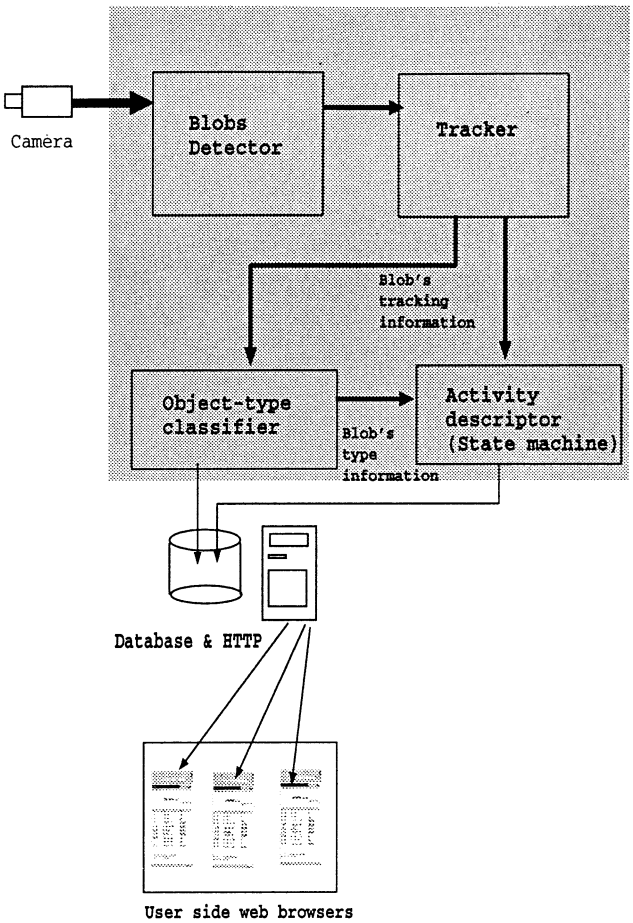


Fig. 3 The test-bed system.

pair of trajectories. These models continue to exist till the time when one blob of each pair is not detected for a certain duration (a few seconds). In order to prevent combinational explosion, the model has to have limited length, and should exist only for the pair for which the distance between blobs is smaller than some threshold.

4. Implementation

Our method for monitoring activity was implemented in the CMU VSAM test-bed system [5] as shown in Fig. 3.

In this figure, candidates of objects are detected as blobs by a blob-detector and trajectories of these blobs are detected by a tracker. Detected blobs are classified based on object type (semantic categories) such as human, human group, and vehicle by an object classifier. Finally, using the object's types and the actions and interactions obtained by observing their trajectories, activities for each pair of blobs are described by an activity-descriptor (a state-machine) which generates a trellis and calculates the path maximizing the a posteriori probability.

The first three of these functions run in real time (about 10FPS) in the test-bed system, and the last function runs off-line. Cropped subimages and classified types of detected objects are stored in a database of this test-bed with their activities to be retrieved. All activities can be explored by web browsing via CGI through HTTP server.

4.1 Blob-Detector and Tracker

To monitor activities between objects, such as a human entered or got out of a vehicle, we need to detect both objects even when the human overlaps the stopped vehicle. To achieve this capability, the blob-detector introduced "layered adaptive background subtraction" [5] based on analyzing whether a pixel is stationary or transient to detect moving and stopped blobs respectively. The tracker [5] extends the basic Kalman filter notion to maintain a list of multiple hypotheses to acquire the trajectories of multiple observed blobs. A blob can be tracked when it disappears for some frames, or when it splits into two blobs due to noisy background subtraction. When an object is occluded behind another one, and only becomes visible again after some time, a new trajectory is generated. In this test, we

Table 1 Blob's observations and attribute set.

(a)Observations

Object-type labels	Human/Vehicle/Human-Group/Uncertain
Action-type labels	Appear/Move/Stop/Disappear/Uncertain
Interaction-type labels	Near/From/To/No-Inter

(b)Blob's attributes set

Object-type	o0:Human, o1:Vehicle, o2:Human-Group
Action	a0:Appear, a1:Move, a2:Stop, a3:Disappear
Interaction	i0:Near, i1:From, i2:To, i3:No-Inter

o0, ... o2, a0, ... a3, i0, ... i3 are described in Eq. (1).

except such situations, even though such broken trajectories can potentially be remerged using classification results. The observations used in the test are described in Table 1 (a). The tracker extracts features for the action-type labels in the table for each detected blob if the length of the trajectory exceeds a threshold (for removing noise). These features are used as input for the activity-descriptor.

4.2 Object-Type Classifier

Another important observation used in the activity descriptor is an object-type label for each blob described in Table 1 (a). To obtain the label, we use an object-type classifier based on Linear Discriminant Analysis of the blob's appearance [5]. In the test-bed system, appearance features used for analysis were area, center of gravity, width and height of a blob, and 1st, 2nd and 3rd order image moments along the x-axis and y-axis.

4.3 Activity Descriptor and Tables for Conditional Probabilities and Joint Probability

In this test, the target activities to monitor were "A Human entered a Vehicle," "A Human got out of a Vehicle" and "Human Rendezvous."

During monitoring, the trellis diagrams for these three types of activities are generated for each pair of blobs whose distance is smaller than a certain threshold. To calculate Eq. (5) through trajectories, a label which is decided by the combination of object-type label, action-type label, and interaction-type label is used. The interaction label is calculated by referring to the inter-blob distance and the relative velocity between each blob in a pair.

Decision of activity for the input scene is made by selecting from the three trellis diagrams corresponding to the pair of trajectories the one that has the maximum posteriori probability.

Conditional probabilities and joint probabilities described in Sect. 3 for the activities are obtained through the following training steps:

- Blob pairs that correspond to each activity are collected from sampled scenes.

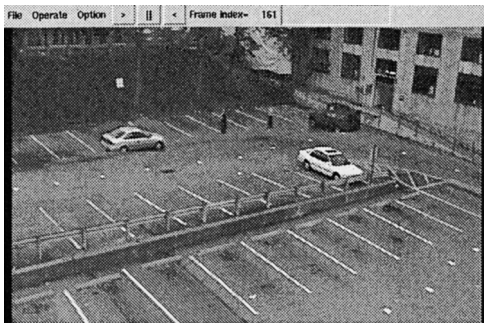


Fig. 4 A scene in test image sequences.

- Trajectories for each pair of blobs for each frame are detected by the blob-detector and tracker. For each detected blob, an object-type is obtained by an object-type classifier. An action such as Appear, Move, Stop, ... and an interaction like Near, From, To, ... are decided by using the trajectories.
- For each blob's pair, an observation label combining an object-type label, an action-type label, and an interaction-type label is assigned.
- For each pair of blobs, an attribute set label described in Table 1(b) is assigned to each detected blob in off-line teaching .
- A priori probabilities in terms of the right hand side of Eq. (5) and Eq. (6) are obtained for each blob's pair for each activity.

These probabilities are calculated by counting the number of events, like

$$\mathrm{N}(\mathrm{B}^{(i)}, O^{(i)}, A_t^{(i)}, I_t^{(i,j)})/N(O^{(i)}, A_t^{(i)}, I_t^{(i,j)}).$$

Here N(a) shows the number of event a.

5. Experimental Results

We have tested the functionality of our method with some image sequences acquired in a parking lot at Carnegie Mellon University during the daytime. These image sequences have activities including human-human and human-vehicle interactions. An example scene from these image sequences is shown in Fig. 4.

The image sequences for our performance test were acquired between 10:00 a.m. and 2:00 p.m. on a day in October 1999 by the test-bed system. Weather conditions ranged from fine to cloudy that day. The test sequences contained scenes ranging from 1 second to 8 seconds, and their total length was 10 minutes. The a priori probabilities for the test were obtained by using 10 additional minutes of image sequences acquired at the same site. The number of trajectories for each activity contained in the training sequences for the a priori probabilities are shown in Table 2.

The objective of this test was mainly to check the performance of the activity-descriptor whose function

Table 2 Number of trajectories for each activity in training data.

A Human entered a Vehicle	4
A Human got out of a Vehicle	3
Human Rendezvous	2

Table 3 Recognition results.

Ground Truth	Detected						
	En	Go	Rn	Rj	C	T	%
En	7	1	0	0	7	8	87.5
Go	0	8	0	0	8	8	100
Rn	0	0	2	1	2	3	66.6
Rj	0	0	0	27	27	27	100
TYPE-I					17	19	89.4
TYPE-II					44	46	95.7

En: A Human entered a Vehicle, Go: A Human got out of a Vehicle, Rn: Human Rendezvous, Rj: Reject, C: Correct, T: Total.

is not only a description of activity for each input trajectory pair but also rejection of false trajectory pairs caused by noise or that have no interaction. We define two type of recognition rates as follows.

$$TYPE - I \equiv CP/TP \tag{7}$$

$$TYPE - II \equiv (CP + CR)/(TP + FP) \tag{8}$$

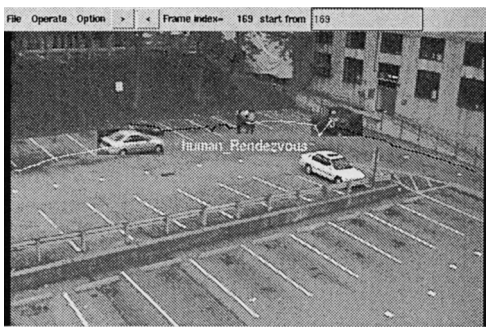
Here, CP is the number of correctly described pairs, CR is the number of correctly rejected pairs, TP is the number of true pairs with distance of blobs in the pair under two meters, and FP is the number of false pairs caused by noise or having distance over two meters.

Table 3 shows the number of activities detected by the activity-descriptor for test sequences consisting of three ground truth events and events to reject. Through the test, the number of blobs detected by the blob-detector was 2667 and 46 trajectory pairs were determined in the tracker. For the trajectory pairs, 17 pairs were correctly recognized when true pairs were 19. So TYPE-I recognition rate was 89.4%. Meanwhile, TYPE-II recognition rate amounted to 95.7% because all of 27 false pairs were correctly rejected.

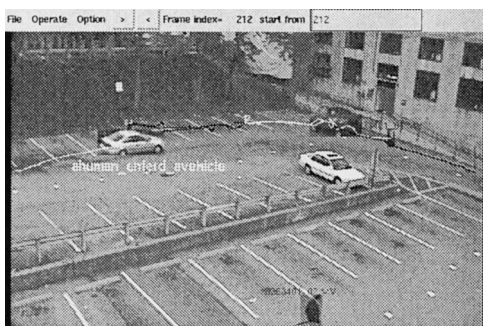
The detected results for human-vehicle interaction and human-human interaction are shown in Fig. 5. In these three scenes, events occurred near the center of the picture. Therefore, the observations of blobs described in Table 1 (a) were frequently (but not always) detected correctly through trajectories at the scenes.

The mis-detected results are shown in Fig. 6. In this case, an activity was mis-detected as "A Human entered a Vehicle," while the ground truth was "A Human got out of a Vehicle" as shown in Table 3. If each observations in the test were close to those in training data, typical sequences of attribute sets for training the activity and sequences selected for the test sequence should be close. In this case, typical context which we taught to describe the scene was as follows.

(a) A result for "a human got out of a vehicle"

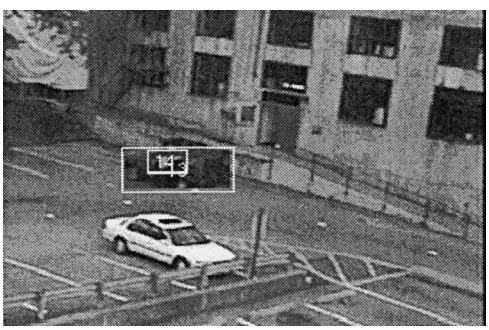


(b)A result for "human rendezvous"



(c)A result for "a human entered a vehicle"

Fig. 5 Typical results of activity monitoring.



—— Detected Observation Sequence ——
Vehicle,Stop <– Near –> Uncertain,Appear
Vehicle,Stop <– Near –> Uncertain,Move
.....,.... <– .... –> ....,....
.....,.... <– .... –> ....,....
Vehicle,Stop <– Near –> Uncertain,Move
Vehicle,Stop <– From –> human,Move
Vehicle,Stop <– From –> Uncertain,Move

Fig. 6 Typical mis-detection of activity monitoring.

Fig. 7 An example of wrongly-rejected scene.

Vehicle,Stop <– Near –> human,Appear
Vehicle,Stop <– Near –> human,Move
.....,.... <– .... –> ....,....
.....,.... <– .... –> ....,....
Vehicle,Stop <– Near –> human,Move
Vehicle,Stop <– From –> human,Move
Vehicle,Stop <– From –> human,Move

The selected attribute sequence for the test sequences were as follows.

Vehicle,Stop <– Near –> human,Stop
Vehicle,Stop <– Near –> human,Stop
.....,.... <– .... –> ....,....
.....,.... <– .... –> ....,....
Vehicle,Stop <– Near –> human,Stop
Vehicle,Stop <– Near –> human,Stop
Vehicle,Stop <– Near –> human,Stop

These two are different in the sense of context. Moreover, the observation sequence in the figure shows that a detected blob which corresponded to a human continued to be labeled as "unclassified" during 20 frames at the beginning of the event. In this case, the observations with poor samples for learning appeared in the scene and our method could not compensate them because of assumption of 1st-order Markov model. To force situation-specific tuning, we have to set the a priori probability of this model by using training data analogous to the test input. In this way, explicit context in our description helps can help with tuning.

Generally speaking, these additional training data are not always available. The system by Oliver [4] have generated such training data by a multi-agent CG simulator. In our method, as training is made by just counting the number of events through attribute sequence, it is expected that models can be tuned not only by training such synthetic data but also by changing a priori probability shown in Eq. (5).

The other case which caused inaccurate results is shown in Fig. 7. There was only one blob detected in the scene, which consisted of 2 persons walking together towards the same direction, with considerable amount of overlapping between the blobs corresponding to the 2 persons. Therefore, the scene was wrongly rejected as it was considered to be a scene which had no activities with interaction. It is quite difficult to segment overlapping objects. But if the targets can be segmented by using other information such as color or appearance pattern together with background subtraction, this type of scene could be described in our system.

6. Conclusion

A basic idea for monitoring activities of multiple objects in a video surveillance system was presented and the functionality of this method was tested by using 10 minutes of video.

In our method, activities of multiple, interacting objects were described by explicit context called attribute set for allowing situation-specific tuning, and parameters of these description were estimated from real training data.

In this test, recognition rate to monitor events correctly was about 89%, though a limitation caused by an assumption of a 1st-order Markov model was shown. And more, the availability of our explicit context description for situation-specific tuning was shown.

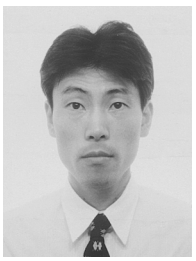
For our future work, we need to test the method by using longer video scenes with a larger variety of events, mainly for checking the limitation of the Markov assumption more precisely. To train this method correctly, we need video sequences which consists of relatively rare events. Another big issue is developing a method for training a priori probabilities efficiently when only a limited number of samples scenes are given.

Acknowledgments

The authors would like to thank Dr. Robert Collins, Dr. Alan Lipton, Mr. David Duggins, Mr. Raju Patil and other CMU VSAM members for their help and insightful comments.

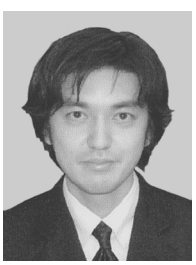
References

[1] H. Ohata, N. Enomoto, A. Okazaki, H. Kawasumi, S. Sudo, and Y. Yamada, "A human detector based on flexible pattern matching of silhouette projection," Proc. MVA'94, IAPR Workshop on Machine Vision applications, 1994.
[2] A. Lipton, H. Fujiyoshi, and R. Patil, "Moving target classification and tracking from real-time video," IEEE Workshop on Applications of Computer Vision (WACV), pp.8–14, Princeton NJ, Oct. 1998.
[3] Y. Ivanov and A. Bobick, "Parsing multi-agent interactions," M.I.T. Media Laboratory Perceptual Computing Section Technical Report, no.479, Nov. 1998.
[4] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," Proc. ICVS'99, Gran Canaria, Spain, Jan. 1999.
[5] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa, "A System for video surveillance and monitoring: VSAM final report," Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May 2000.

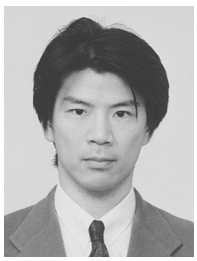


Nobuyoshi Enomoto received the B.S. and M.S. degrees in Electronic Engineering from University of Electro-Communications, Tokyo, Japan, in 1985, 1987 respectively. He joined TOSHIBA in 1987 where he has worked on the research and development of image proccessing and computer vision. Currently, he is a specialist of the Baseline Techonology Department, Yanagicho Operations-e-SOLUTIOS, TOSHIBA CORPORATION, Kawasaki, Japan. From 1998 to 2000, he was a visiting industrial scholar of the Robotics Institute, Carnegie Mellon University. His research interests include computer vision, multi-modal human computer interaction and intelligent transport systems.

Takeo Kanade received his Ph.D. in Electrical Engineering from Kyoto University, Japan, in 1974. After being on the faculty of the Department of Information Science, Kyoto University, he joined the Computer Science Department and Robotics Institute in 1980. Currently he holds the title of U. A. and Helen Whitaker University Professor of Computer Science and Robotics. He was the Director of the Robotics Institute from 1992 to Spring 2001. He served as the founding Chairman (1989 – 1993) of the Robotics Ph. D. Program at CMU, probably the first of its kind in the world. Dr. Kanade has worked in many areas of robotics, including manipulators, sensors, computer vision, multimedia applications, and autonomous robots, with more than 200 papers on these topics. He founded the International Journal of Computer Vision and had been the main editor till the end of 2000. Dr. Kanade's professional honors include: election to the National Academy of Engineering, a Fellow of IEEE, a Fellow of ACM, a Fellow of American Association of Artificial Intelligence; several awards including C & C Award, the Joseph Engelberger Award, JARA Award, Otto Franc Award, and Marr Prize Award.



Hironobu Fujiyoshi is a faculty member of the Department of Computer Science in Chubu University, Japan. He received his PhD in Electrical Engineering from Chubu University in 1997. For his thesis he developed a fingerprint verification method using spectrum analysis, which has been incorporated into a manufactured device sold by a Japanese security company. From 1997 to 2000 he was a post-doctoral fellow at the Robotics Institute of Carnegie Mellon University, working on the DARPA Video Surveillance and Monitoring (VSAM) effort and the humanoid vision project for the Honda Humanoid Robot. He performs research in the areas of real-time object detection, tracking, and recognition from video.



Osamu Hasegawa received the B.S. and M.S. degrees in Mechanical Engineering from Science University of Tokyo, Tokyo, Japan, in 1988, 1990 respectively. He received Ph.D. degree in Electrical Engineering from University of Tokyo, Tokyo, Japan, in 1993. Currently, he is a senior research scientist of the Neuroscience Research Institute, Advanced Industrial Science and Technology, Tsukuba, Japan. From 1999 to 2000, he was a visiting research scientist of the Robotics Institute, Carnegie Mellon University. His research interests include computer vision, multimodal human computer interaction, and cognitive brain science. He is a member of the IEEE, AAAI, IPSJ and others.