1—6

A method for monitoring activities of multiple objects by using stochastic model

Nobuyoshi Enomoto *
Information and Industrial Systems
& Services Company
Toshiba Corporation

Takeo Kanade†
The Robotics Institute
Carnegie Mellon University

Hironobu Fujiyoshi‡
Department of Computer Science
Chubu University

Osamu Hasegawa§
The Machine Understanding Division
Electrotechnical Laboratory

Abstract

We present a method for estimating activities of multiple objects which are detected in video surveillance systems.

In most existing video surveillance systems , the objects detection and classification sometimes cause inaccurate results. In addition to this, we want to monitor activities of objects including interactions between them for long term image sequence. To solve this problem, we newly introduce pre-defined knowledge that each blob has attributes set which consists of object's type, action, and interaction. Using probabilistic relations introduced by a specific Markov model of these attributes sets , the activity descriptions are estimated accurately .

1 Introduction

In most existing video surveillance systems like former VSAM test-bed system in CMU, the candidates of moving objects can be detected as blobs and some object's type can be classified. Lipton, et al developed a method to detect moving objects by adaptive background subtraction and a method to classify them by using Neural Network[1]. But these systems sometimes cause inaccurate results by the changing of lighting condition and the changing of object's appearance . In addition to these functions, we want to monitor activities of objects including interactions between them like "A human entered a vehicle" for long term image sequence. A system for detecting multi-agent interactions using SCFG was developed by Ivanov and Bobick[2]. In this system, they set the event likelihoods and the production rule probabilities manually. Oliver, et al also developed a system for detecting people interaction using Coupled Hidden Markov Model(CHMM)[3]. They used a multi-agent simulator to generate synthetic training data of CHMM. Meanwhile, we introduce pre-defined knowledge that each blob has attributes set which consists of object's type, action, and interaction. To estimate the activities for image sequences, the probabilistic relations of each attributes set for each blobs is used.

2 Stochastic estimation of activities: problem definition

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 $B_0^{(i)},....B_{t-1}^{(i)},B_t^{(i)}$ and $B_0^{(j)},....B_{t-1}^{(j)},B_t^{(j)}$. An sequence of blob i is considered as observation of an real object which has sequence of attributes set of object-type $O^{(i)}$, actions $A_0^{(i)},....A_{t-1}^{(i)},A_t^{(i)}$ and interactions $I_0^{(i,j)},....I_{t-1}^{(i,j)},I_t^{(i,j)}$ between the other blob j frame by frame. Our final goal is to obtain the most reliable description by maximizing conditional joint probability as,

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

where

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

$O^{(i)}$,$O^{(j)}$,$A_t^{(i)}$,$A_t^{(j)}$,$I_t^{(i,j)}$,$B_t^{(i)}$,$B_t^{(j)}$ are variables with states for the object types $o = o0, o1, ...$, actions $a =$

*Address: 70 yanagi-cho, saiwai-ku, kawasaki 212-8501 Japan. E-mail: nobuyoshi.enomoto@toshiba.co.jp.

†Address: 5000 Forves Avenue, Pittsburgh, PA 15213. E-mail: tk@cs.cmu.edu

‡Address: Kasugai, Aichi Japan 487-8501. E-mail: hf@cs.chubu.ac.jp

§Address: 1-1-4 Umezono, Tsukuba Japan 305-8568 . E-mail: hasegawa@etl.go.jp

$a0, a1, ...$, interactions $i = i0, i1, ...$, and observations $b = b0, b1, ...$.

3 Markov model for selecting most probable attributes sequence

If supposed only two blob sequences blob-i blob-j are in the scene which have only 2 frames, the conditional joint probability for these blobs is described as below.

$$P\left(O^{(i)}, O^{(j)}, A_1^{(i)}, A_1^{(j)}, A_0^{(i)}, A_0^{(j)}, I_1^{(i,j)}, I_0^{(i,j)} \mid B_1^{(i)}, B_1^{(j)}, B_0^{(i)}, B_0^{(j)}\right)$$
$$= \frac{P\left(B_1^{(i)}, B_1^{(j)}, A_1^{(i)}, A_1^{(j)}, I_1^{(i,j)} | O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}\right)}{P\left(B_1^{(i)}, B_1^{(j)} | B_0^{(i)}, B_0^{(j)}\right)}$$
$$\bullet P\left(O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)} \mid B_0^{(i)}, B_0^{(j)}\right) \quad (2)$$

Using Bayes Rule, the first term of the right hand of equation(2) is

$$\frac{P\left(B_1^{(i)}, B_1^{(j)}, A_1^{(i)}, A_1^{(j)}, I_1^{(i,j)} | O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}\right)}{P\left(B_1^{(i)}, B_1^{(j)} | B_0^{(i)}, B_0^{(j)}\right)}$$
$$= \frac{P\left(B_1^{(i)}, B_1^{(j)} | O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_0^{(i)}, A_0^{(j)}, A_1^{(i)}, A_1^{(j)}, I_1^{(i,j)}, I_0^{(i,j)}\right)}{P\left(B_1^{(i)}, B_1^{(j)} | B_0^{(i)}, B_0^{(j)}\right)}$$
$$\bullet P(I_1^{(i,j)} \mid O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_1^{(i)}, A_1^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)})$$
$$\bullet P(A_1^{(i)}, A_1^{(j)} \mid O^{(i)}, O^{(j)}, B_0^{(i)}, B_0^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}) \quad (3)$$

Equation(2) means that the conditional probabilities for 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 huge, all of these conditional probabilities from t=0 to t=t'-1 can't be used.

To overcome this, we make some assumptions below.

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

Attributes sets in each sequence follow Markov model as,

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

Using these assumptions and equation(3) , the conditional joint probability of equation(2) is described generally as,

$$P\left(O^{(i)}, O^{(j)}, \left(A^{(i)}\right)_0^t, \left(A^{(j)}\right)_0^t, \left(I^{(i,j)}\right)_0^t \mid \left(B^{(i)}\right)_0^t, \left(B^{(j)}\right)_0^t\right)$$
$$\simeq \frac{P\left(B_t^{(i)} | O^{(i)}, A_t^{(i)}, I_t^{(i,j)}\right) P\left(B_t^{(j)} | O^{(j)}, A_t^{(j)}, I_t^{(i,j)}\right)}{P\left(B_t^{(i)}, B_t^{(j)}\right)}$$
$$\bullet P\left(I_t^{(i,j)} \mid O^{(i)}, O^{(j)}, A_t^{(i)}, A_t^{(j)}\right)$$
$$\bullet P\left(A_t^{(i)} \mid O^{(i)}, A_{t-1}^{(i)}, I_{t-1}^{(i,j)}\right)$$
$$\bullet P\left(A_t^{(j)} \mid O^{(j)}, A_{t-1}^{(j)}, I_{t-1}^{(i,j)}\right)$$
$$\bullet P(O^{(i)}, O^{(j)}, \left(A^{(i)}\right)_0^{t-1}, \left(A^{(j)}\right)_0^{t-1}, \left(I^{(i,j)}\right)_0^{t-1}$$
$$\mid \left(B^{(i)}\right)_0^{t-1}, \left(B^{(j)}\right)_0^{t-1}) \quad (4)$$

where if t=0,

$$P\left(O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)} \mid B_0^{(i)}, B_0^{(j)}\right)$$
$$\simeq P\left(B_0^{(i)} \mid O^{(i)}, A_0^{(i)}, I_0^{(i,j)}\right)$$
$$\bullet P\left(B_0^{(j)} \mid O^{(j)}, A_0^{(j)}, I_0^{(i,j)}\right)$$
$$\bullet \frac{P\left(O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}\right)}{P(B_0^{(i)}, B_0^{(j)})} \quad (5)$$

To calculate these equations , we need tables for 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 priori probabilities can be obtained by counting events for each attributes sets in sampled image sequences.

The path which maximize posterior conditional joint probabilities described in equation(1) can be obtained by calculating equation(4) through the trellis diagram in Figure 1. In this diagram, s1,s2....are state's 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),... mean that blob-i and blob-j have an interaction label "NEAR",... in a state.

4 Experiment

We tested the functionality of our method with some image sequences which was acquired in a parking lot of Carnegie Mellon University at daytime and these have activities including human-human, human-vehicle interaction. A scene in these image sequences is shown in Figure 2.

4.1 Test-bed system

The test-bed system is mainly consists of an blobs-detector, an tracker, an object-type classifier and a state-machine which calculate the path which maximize the posterior. These are implemented on CMU VSAM test-bed system[4] and the first three modules run in real time(about 10FPS).

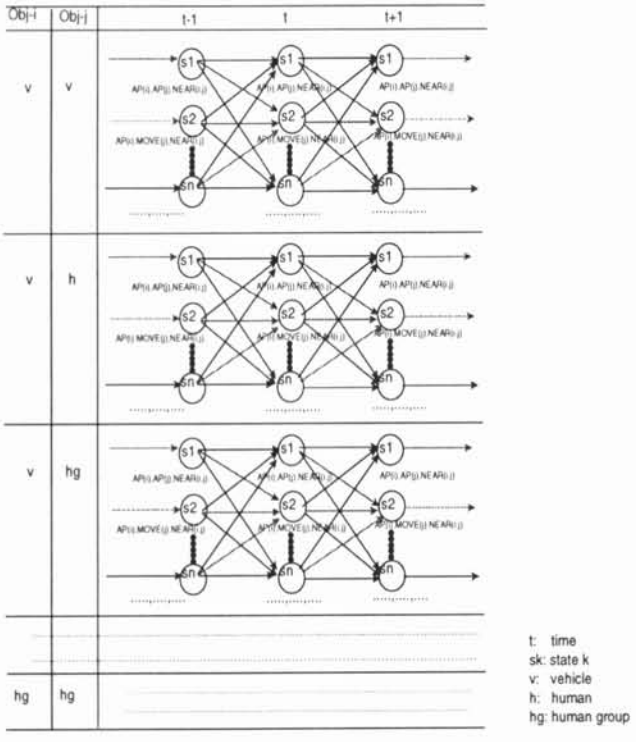


Figure 1: trellis diagram

Figure 2: A scene in test image sequences

4.1.1 blobs-detector and tracker

To monitor activities between objects needs to recognize when objects have stopped and even disambiguate overlapping objects. To capture these functionality, the blobs-detector introduced "layered adaptive background subtraction"[4] based on processes to analyze whether a pixel is stationary or transient to detect moving and stopping blobs respectively. The tracker[4] extended the basic Kalman filter to maintain a list of multiple hypotheses to acquire multiple blob's trajectories as observations. The observations used in the the test are described in Table:1(a). The tracker made up the feature for action-type label, blob's distance, and relative velocity between each blobs from the detected blobs.

4.1.2 object-type classifier

Another important observation used in activity monitoring is a object-type label for each blob described in Table:1(a). To obtain the label, we used the object-type classifier based on Linear Discriminant Analysis with blob's appearance[4]. In the test-bed system, features of the appearance to analyze 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.

4.1.3 Table for conditional probabilities and joint probability

In this test, target activities to monitor were " A Human entered a Vehicle" , "A Human got out of a Vehicle" and " Human Rendezvous". In on-line monitoring, decision of activity for input scenes were made by selecting maximum posterior which corresponds to the activities. Conditional probabilities and joint probability described in section:3 for the activities were obtained through following step:

- Sample scenes which correspond to each activities is collected.
- The observations for each blob for each scene are detected and quantized to 80 labels.
- The attributes set for each blob described in Table: 1(b) is assigned to each detected blob in off-line teaching .
- Events in each scenes are counted to make up probabilities

4.2 Experimental results

For the 10 minutes image sequences whose scenes were not used for learning probabilities , our system could detect correct activities for 89% of events even

Object-type labels (Human, Vehicle, Human-Group, Uncertain)
Action-type labels(Appear, Move, Stop, Disappear,Uncertain)
Distance and velocity between each objects

(a)observations

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

(b)blob's attributes set
o0,...o2, a0,...a3, i0,...i3 are described in equation(1).

Table 1: blob's observations and attributes set

though the observed features were often inaccurate. The typical detected results for human-vehicle interaction and human-human interaction are shown in Figure 3.

5 Conclusion

A basic idea for monitoring activities of multiple objects in a video surveillance system were presented and the functionality of this method was tested by using 10 minutes of video. In this test, activities for 89% of events are monitored correctly. For our future work, one thing which we should address is to test the method by using longer and various video scenes. To train this method correctly needs certain amount of video scenes. Our second issue which is planning is a method to tune priori probabilities when only limited amount of scenes are given.

6 Acknowledgments

Authors would like to thank Robert Collins, Alan Lipton, David Duggins and other CMU VSAM members for their help and insightful comments.

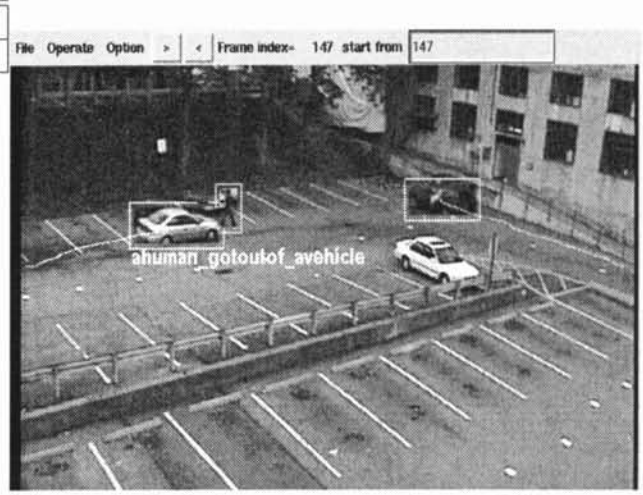
References

[1] Lipton, Fujiyoshi and Patil, "Moving Target Classification and Tracking from Real-time Video" IEEE Workshop on Applications of Computer Vision (WACV), Princeton NJ, October 1998, pp.8-14.

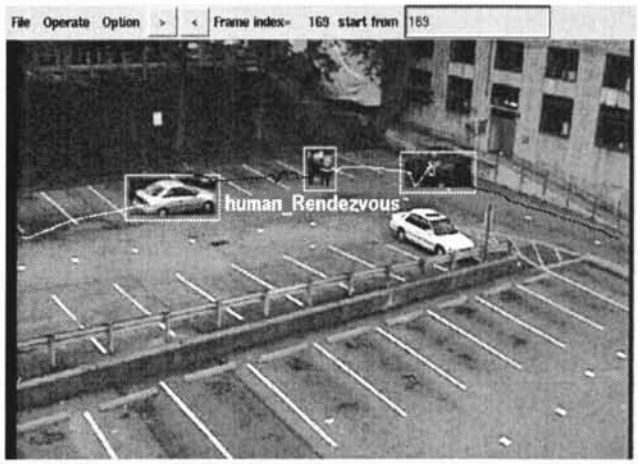
[2] Y. Ivanov, A. Bobick, "Parsing Multi-Agent Interactions", M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 479. Nov. 1998.

[3] N. Oliver , B. Rosario, and A. Pentland, "A Bayesian Computer Vision System for Modeling Human Interactions", Proceedings of ICVS'99 , Gran Canaria, Spain, Jan 1999.

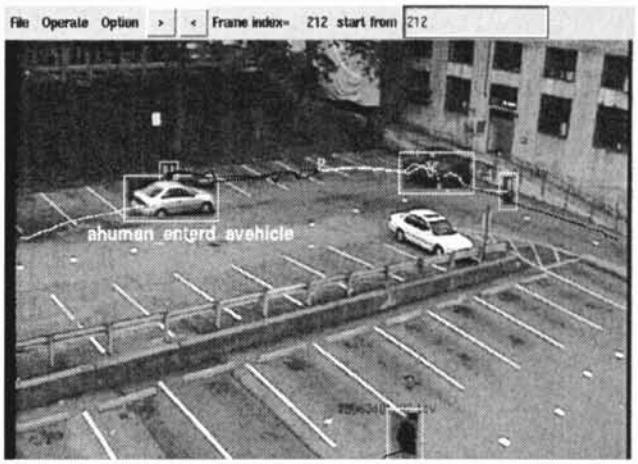
[4] Collins, et al "A System for Video Surveillance and Monitoring: VSAM Final Report", Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May, 2000.



(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"

Figure 3: Typical results of activity monitoring