

## **Video-Based Event Detection**

**Ph.D. research of Somboon Hongeng at the  
University of Southern California (2003)**

**Slides adapted from a talk of S. Hongeng at  
Hamburg University in October 2003**

1

## **Goals and Motivation**

- **Retrieves semantic information from video**
  - Determines if it contains any interesting events
  - When and Where? (i.e., spatial and temporal dimensions)
- **Applications include Video Surveillance, Video Summarization, Human-Machine Interaction, Intelligent Living Spaces**

2

## Monitoring of Vehicle Behaviors



"go through checkpoint"

- Checkpoint is the area between the two tanks

3

## Monitoring of Activities in a Crowd

- Multiple actors and objects
- Interaction among individual actions



"theft at phone-booth (PB)"

4

## Challenges of Event Detection

- **Generic event representation**
- **Effective and robust event recognition**
  - Bridges the gap between pixel values and symbolic event description
  - Computation of uncertainties
    - imperfect tracking of “objects” in noisy videos
    - similar activities must be distinguished
  - Variation in execution styles, temporal durations
  - Generic object recognition
  - Use and acquisition of scene and task context
  - ...

5

## Prior State of the Art

- **Action Recognition using Bayesian networks**
  - Remagnino et al. (1998), Buxton & Gong (1995)
  - Only handles static or simple events
- **Action Recognition using HMMs**
  - Ohya (1992), Starner (1998), Oliver et al. (2000)
  - Parameter space becomes too large in complex events
- **Syntactic Pattern Recognition of Actions**
  - Pinhanez (1998), Ivanov & Bobick (2000)
  - Action units are assumed to be detected and segmented reliably

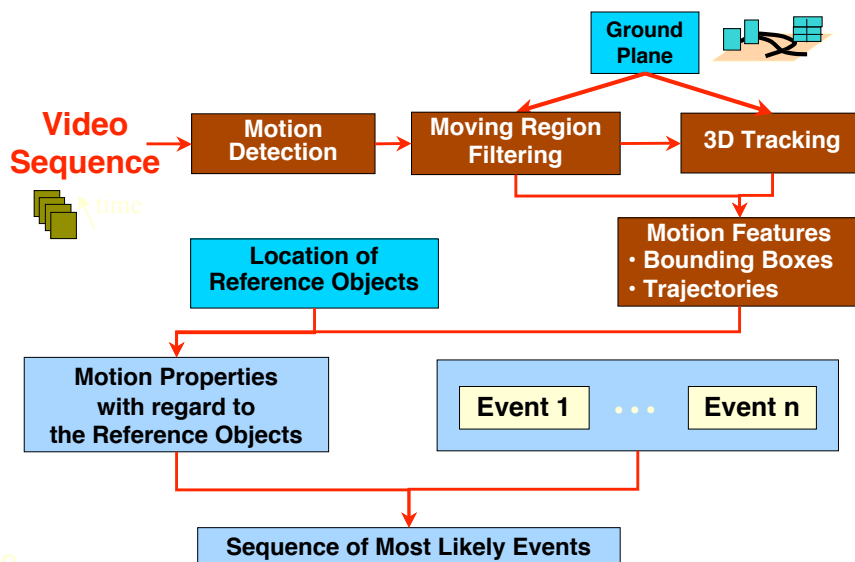
6

## Large-Scale Event Detection System

- Videos taken by a single, calibrated camera
- Moving objects are observed from a distance
  - Closely coordinated movements of body parts cannot be observed reliably
  - Blob shapes and trajectories are main sources of info
- Scene and task contexts are given
  - Interesting events to be detected are known and can be modeled a priori
  - Locations and types of scene objects are known

7

## System Overview



8

## Motion Detection & Tracking

- **Statistical background modeling**
  - Pixel-wise mode computation
- **Detects moving regions by background subtraction**
- **Tracks objects by making correspondence between moving regions at different times**
  - Moving regions may split due to low contrast, noise
  - Uses distance on ground plane to select blob correspondence across timeframes
  - Filters split regions based on color distribution consistency

9

## Tracking “Theft at PhoneBooth”



- **Ground tracks are noisy in low camera angle**
  - Few pixels mistake projects to several meters

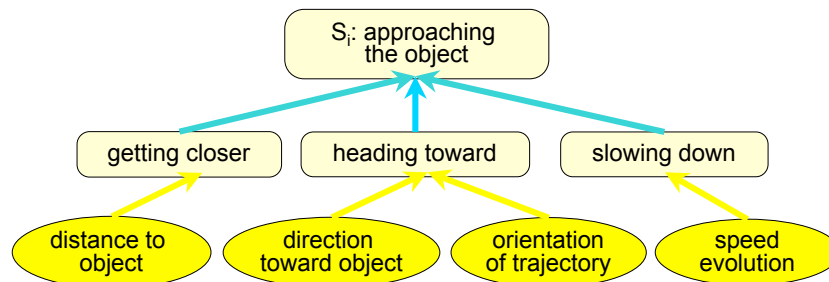
10

## Event Classification

- **Single Thread Events**
  - Simple Event
    - Short, coherent unit of movement (e.g., “going toward”)
    - Static poses (e.g., “stand”, “crouch”)
  - Composite Events
    - Linearly ordered continuous sequence of events
    - Long-term (normally longer than 30 frames)
- **Multiple Thread Events**
  - Temporal and logical combination of two or more single thread events

11

## Object Class and Simple Event Modeling



- Object classes and simple events are modeled by a **Bayesian Network** of sub-events or properties of shape and trajectory of the actor
- Recognized by computing  $P(S_i | O_t)$  at each time frame

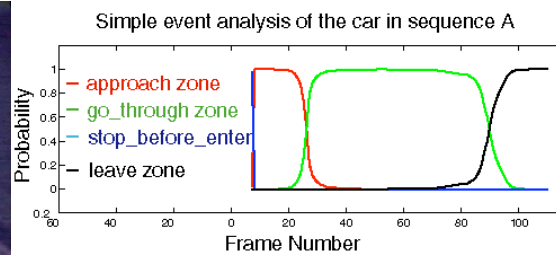
12

## Inferring $P(S_i|O_t)$

- **Compute “evidence”:  $O_t$** 
  - Properties related to object trajectories
  - Properties related to bounding boxes
- **Compute  $P(S_i|O_t)$  from  $O_t$  using Bayes’ rule**
  - Assume conditional probabilities are Gaussian
    - Estimate Gaussian parameters from 600 frames of event samples
  - Normalize  $P(S_i|O_t)$  based on all alternative events ( $S_j, S_k, \text{etc...}$ )

13

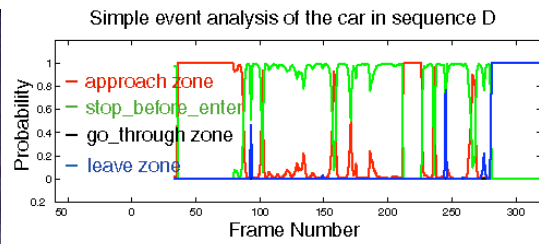
## Simple Event Analysis of “Checkpoint A”



- **Evolution of the output of Bayesian networks  $P(S_i|O_t)$  of four simple events**
- **The “zone” is shown by the quadrangle**

14

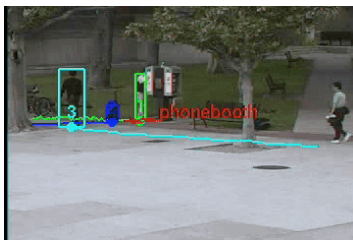
## Simple Event Analysis of “Checkpoint D”



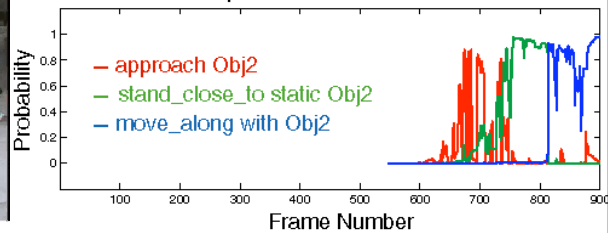
- Evolution of the output of Bayesian networks  $P(S_i|O_i)$  of four simple events

15

## Simple Events of “Take Object”



Simple event analysis of Obj4 with regard to Obj2 (luggage) in sequence PhoneBooth02

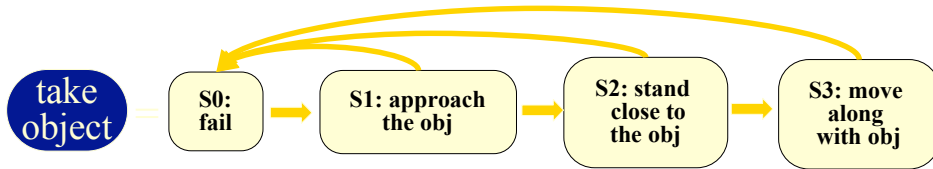


- Output of Bayesian networks  $P(S_i|O_i)$  of three sub-events of “take object”

16



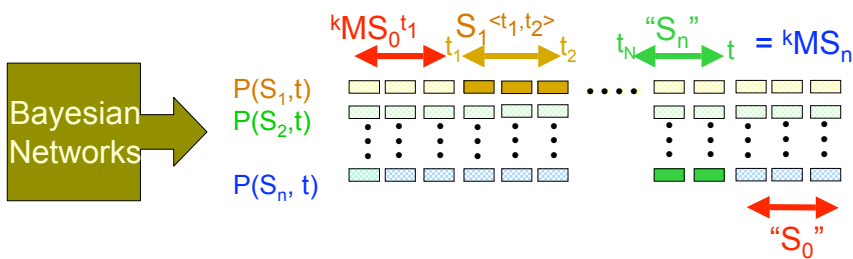
## Composite Event Modeling



- Finite state automaton is used to represent long-term composite events
- Dynamics of composite event are modeled by the transitions from one event state to another
  - Durations of event states can vary
  - Given the Bayesian probabilities of each event state computed for a period of time, a sequence of event states must be segmented appropriately

17

## Recognition of Composite Event $k$



- Let  $S_0, \dots, S_n$  be states of composite event  ${}^kMS$ ;  $O=(O_1, \dots, O_t)$  be the observations;  ${}^kMS_i$  be the fact that  $S_i$  is the current state of  ${}^kMS$
- ${}^kMS_n$  is recognized at frame  $t$  by computing

$$P({}^kMS_n^t | O) = \alpha_0 \sum_{V(t_1, \dots, t_n)} P(O | {}^kMS_0^{t_1} S_1^{<t_1, t_2>} \dots S_n^{<t_n, t>}) P({}^kMS_n^t)$$

Note: we drop  $k$  in the next slides for clarity

18

## Factoring $P(O|MS_n^t)$ and $P(MS_n^t)$

- Under semi-HMM assumption that
  - $O^{<t_m, t_{m+1}>}$  is independent of  $S_n^{<t_n, t_{n+1}>}$  given  $S_m^{<t_m, t_{m+1}>}$
  - probability of  $S_m$  making a transition to  $S_n$  depends on the duration of  $S_m$ ,

we have:

$$P(MS_n^t|O) = \alpha_0 \sum_{\mathcal{V}(t_1, \dots, t_n)} P(MS_0^{t_1}) P(O^{<1, t_1>} | MS_0^{t_1})$$

$$a_{1,0} P(d_{s_1} = t_2 - t_1) P(O^{<t_1, t_2>} | S_1^{<t_1, t_2>}) \dots$$

$$a_{n,n-1} P(d_{s_n} = t - t_n) P(O^{<t_n, t>} | S_n^{<t_n, t>})$$

- $a_{n,m}$  : the probability of the path from  $S_m$  to  $S_n$
- $P(d_{s_m})$  : the distribution of event duration of  $S_m$ ,
  - estimated using direct method, assuming a Gaussian
  - uniform distributions for highly variable event durations

19

## Computing $P(MS_n^t|O)$

- Assuming that  $O^t$  and  $S_n^t$  are independent given  $S_m^t$ ,  $P(O^{<t_m, t_n>} | S_m^{<t_m, t_n>})$  can be computed from Bayesian probabilities as:

$$P(O^{<t_m, t_n>} | S_m^{<t_m, t_n>}) = \beta_{<t_m, t_n>} \prod_{t_m \leq t \leq t_n} P(S_m^t | O^t)$$

$$\beta_{<t_m, t_n>} = \prod_{t_m \leq t \leq t_n} \frac{P(O^t)}{P(S_m^t)} \quad \text{is a normalizing constant}$$

- Let  $P'(MS_N^t|O)$  be the normalized  $P(MS_N^t|O)$ ;  
 $Bel(S_i^{<t_i, t_{i+1}>}, O^{<t_i, t_{i+1}>})$  be  $P(d_{s_i} = t_{i+1} - t_i) \prod_{t_i \leq t \leq t_{i+1}} P(S_i^t | O^t)$ ;

We have:

$$P'(MS_N^t|O) = \sum_{\mathcal{V}(t_1, \dots, t_N)} P'(MS_0^{t_1} | O^{<1, t_1>}) \prod_{1 \leq i \leq n} a_{i,i-1} Bel(S_i^{<t_i, t_{i+1}>}, O^{<t_i, t_{i+1}>})$$

20

## Computing $P'(MS_n^t|O)$ Efficiently

- Direct computation of  $P'(MS_n^t|O)$  is  $O(nT^n)$
- Efficient recursive algorithm based on Dynamic Programming can achieve  $O(nT)$

$$P'(MS_n^t|O) = \sum_{F(t_1, \dots, t_n)} P'(MS_0^t|O^{<1,t_1>}) \prod_{1 \leq i \leq n} a_{i,i-1} Bel(S_i^{<t_i,t_{i+1}>}, O^{<t_i,t_{i+1}>})$$



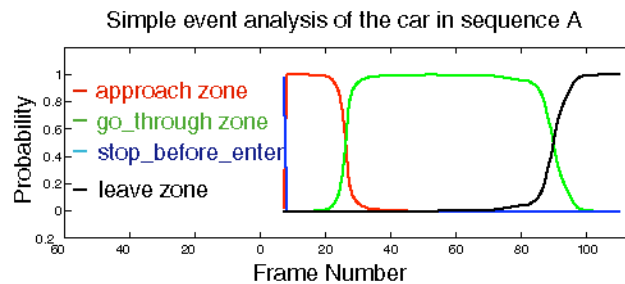
$$P'(MS_n^t|O) = \sum_{F(t_n)} a_{n,n-1} Bel(S_n^{<t_n,t>}, O^{<t_n,t>}) P'(MS_{n-1}^{t_n} | O^{<1,t_n>})$$

At frame  $t$ , for all  $S_i$ , update  $Bel(S_i^{<t_i,t>}, O^{<t_i,t>})$  with Bayesian probability  $P(S_i^t|O^t)$ ; multiply it with  $P'(MS_{i-1}^{t_i} | O^{<1,t_i>})$  that is already computed

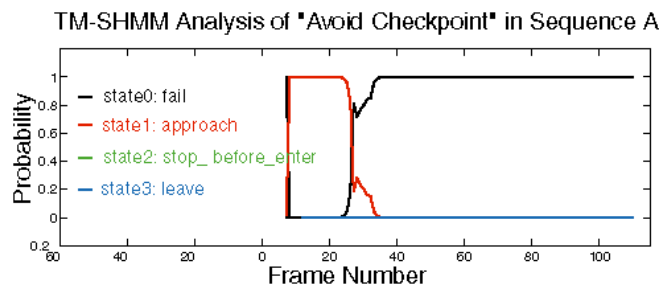
21

## Analysis of “Go Through Checkpoint”

$P(S_i|O^t)$  →



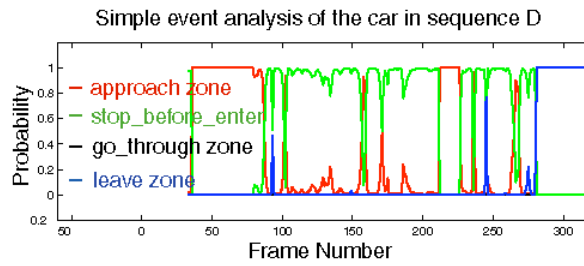
$P'(MS_i^t|O)$  →



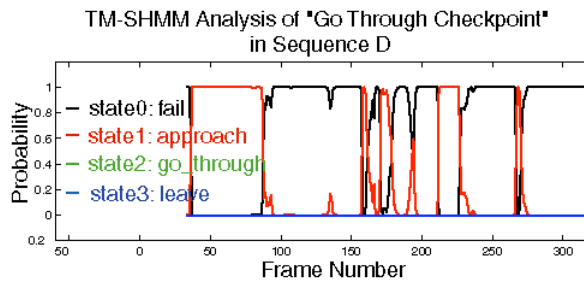
22

## Analysis of "Avoid the Checkpoint"

$P(S_i|O_t)$  →



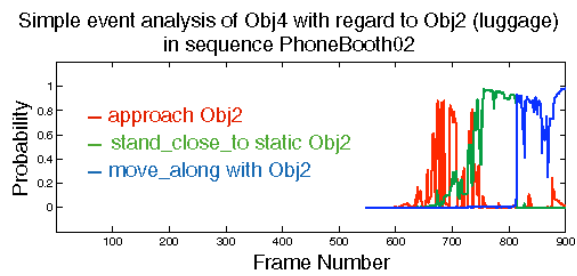
$P'(MS_i^t|O)$  →



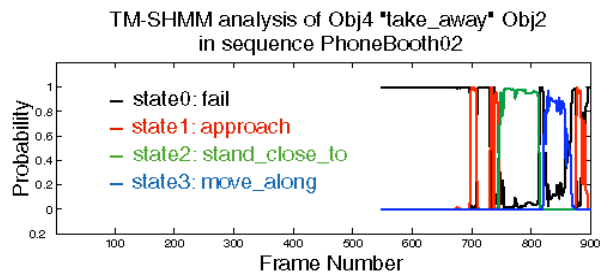
23

## Composite Event: "Take Object"

$P(S_i|O_t)$  →



$P'(MS_i^t|O)$  →



24

## Segmenting Composite Events

- Set a prob threshold to detect ending times ( $k_{t_1}^e, \dots, k_{t_p}^e$ ) of event instances 1, ..., p of  $kMS_n$
- At time frame  $t$ , compute  $P(MS_n^t | O)$  :

$$P(MS_n^t | O) = \alpha_0 \max_{F(t_1, \dots, t_n)} P(O | MS_0^{t_1} S_1^{<t_1, t_2>} \dots S_n^{<t_n, t>}) P(MS_n^t)$$

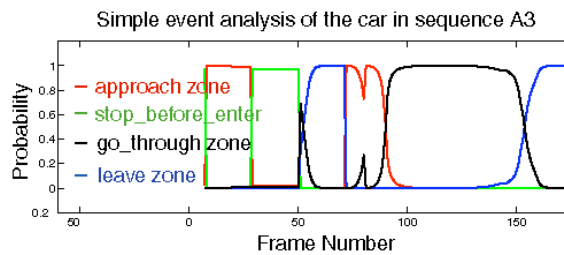
- Backtrack the transitions to  $t_1$  and keep track of  $q$  most likely starting times ( $k_{t_1}^s, k_{t_2}^s, \dots, k_{t_q}^s$ ) during  $k_{t_{i-1}}^e$  and  $k_{t_i}^e$
- Likelihood of event instance  $i$  that ends at  $k_{t_i}^e$  is defined as the maximum value of  $P'(MS_n^t | O)$  during ( $k_{t_{i-1}}^e, k_{t_i}^e$ )

25

## Simulation: Concatenation of “Avoid” and “Go Through”

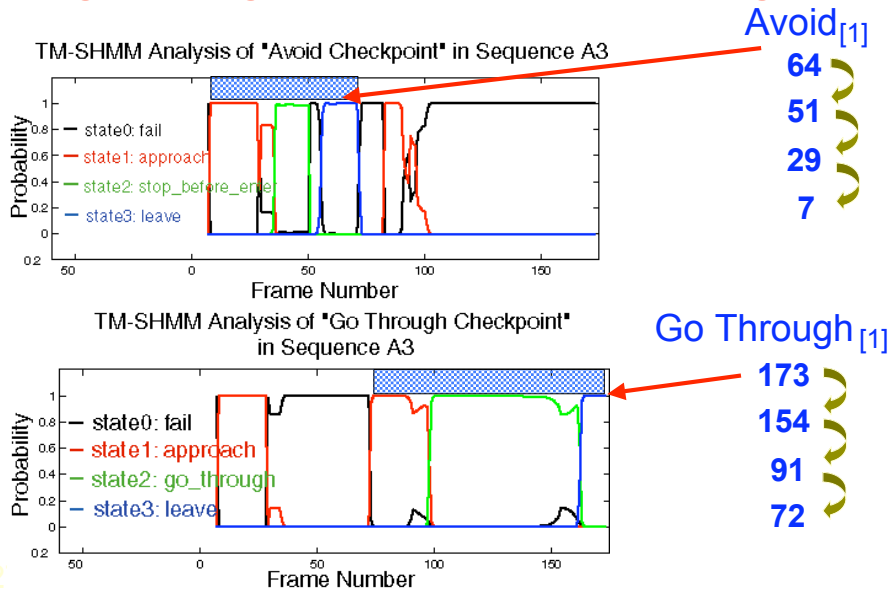


Sequence A3



26

## Segmenting “Avoid” and “Go Through”



## Multi-Thread Event Modeling

- **Global activities can be described by several actors performing related actions ...**
  - Action threads are related by temporal/logical constraints
  - May overlap in a non-linear fashion
- **... represented by an event graph**
  - Nodes are single-thread events
  - Links indicate temporal relations represented by *Interval-Based Temporal Logic*
    - “starts”, “meets”, “during”, “before”, “overlaps”, ...

## “Theft at Phone Booth (PB)”

- **Defines five action threads:**

- Obj1 *bring-in* Obj2
- Obj1 *use-phone*
- Obj3 *take* Obj2
- Obj3 *leave* Obj1
- Obj2 *taken-away-from* Obj1

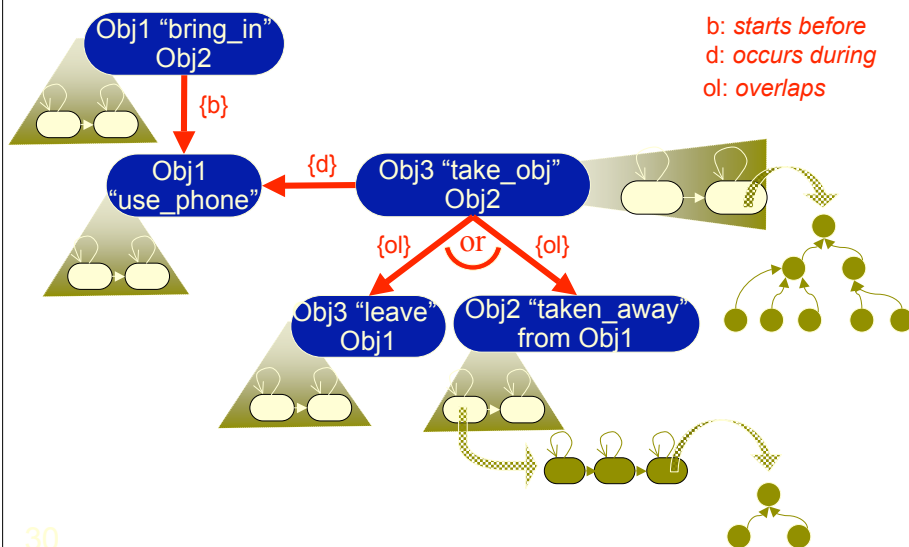


- **Defines the appropriate temporal relations**

- Obj1 *bring-in* Obj2 *starts before* Obj1 *use-phone*
- Obj3 *take* Obj2 *occurs during* Obj1 *use-phone*
- ...

29

## Event Graph for “Theft at PB”



30

## Multi-Thread Event Recognition

- Individual event recognition is uncertain
- Several instances of events may be detected during a period of time
  - “approaches”, “stops”, “approaches”....
- Search for the event threads that best fit the required “*interval-based relations*”
  - How to evaluate the relations of event instances?

31

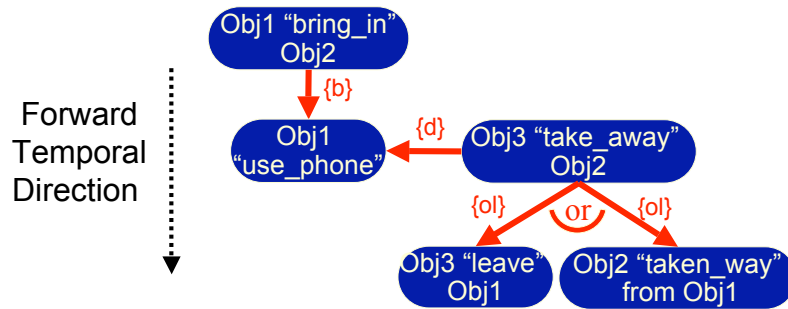
## Evaluation of Temporal and Logical Relations

- **Temporal Relations** are evaluated by combining the probabilities of event instances subject to the corresponding temporal constraint
  - $P(\text{“A starts before B”}) =$   
$$\max_{V(m,n)} P(A_m) P(B_n), \text{ if } Start(A_m) < Start(B_n),$$
  
where “m” and “n” indicate instances of events
- **Logical Relation “Or”** is evaluated by taking the maximum value, i.e.
  - $P(\text{“A or B”}) = \max_{V(m,n)} (P(A_m), P(B_n))$

32



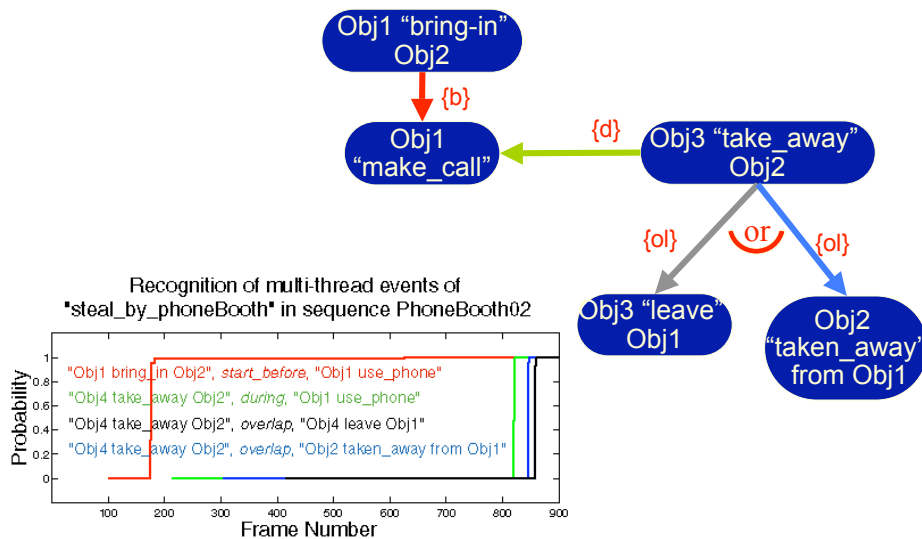
## Inference of a Multi-Thread Event



- Propagate temporal constraints and probabilities of events along forward temporal direction of event graph
  - We need to consider “bring\_in before use\_phone” before we evaluate “take\_away during use\_phone”
- $O(TP^{R+1})$  complexity if there are R event relations and P average number of event instances with different starting times

33

## Recognition of “Theft at PB”



34

## Annotated Videos

- **Needs standard interface for video content descriptions**
  - eXtended Markup Language (XML) interface can be defined for event descriptions
- **Event analysis results can be written in XML**
  - moving object and event descriptions
  - allows the search for content of videos
- **Information in XML files can be parsed and overlaid on the original videos for visualization**

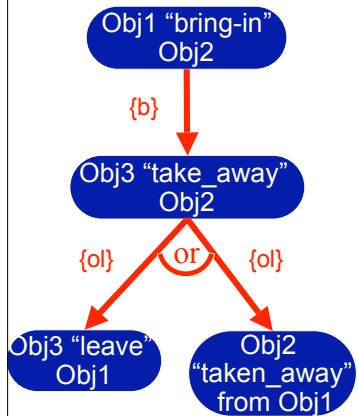
35

## Annotated “Theft at PB”



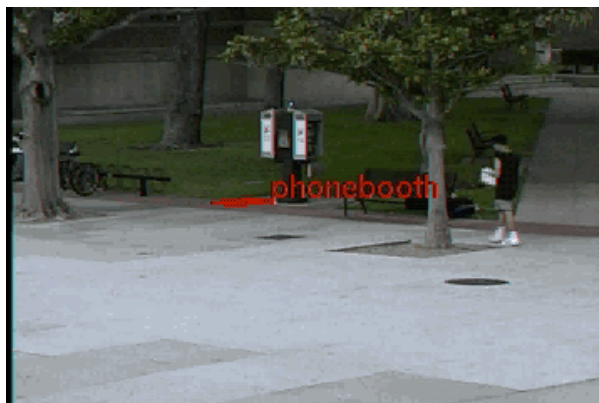
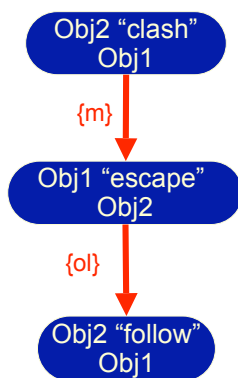
36

## Annotated "Object Transfer"



37

## Annotated "Assault"



m: meet  
ol: overlap

38

## Performance

- **96.7% accuracy on discriminating competing single-thread events of 30 objects (including human and vehicles)**
- **Small trajectory perturbation with Gauss noise**
  - Performance drops 5% on 40 simulated noisy sequences corrupted with  $N(\mu=0, \sigma=6.68\text{cm})$ , equivalent of human walking speed variance
- **Large variations simulating different execution styles (and some tracking blunders)**
  - 81% detection rate, 16% false alarms

39

## Computation Time

- **P2-333 MHz, 128 MB RAM (approximately 1/8<sup>th</sup> of today's processing power)**
- **Computation time excludes motion detection and tracking processes**

Sequence	No of objs	Frames	SE/CE/MT/Ctx	Time (sec)	fps
Chekpnt A	2	109	38/3/0/1	2.5	43.6
Chekpnt D	3	292	38/3/0/1	18	16.22
Assault	2	240	68/8/1/0	22.5	10.67
Object Transfer	3	640	83/11/3/1	453	0.71
Steal by Blocking	4	460	104/15/2/3	994	0.46

40

## Conclusion and Future Work

- Probabilistic event analysis is robust, but performance depends on tracking accuracy
- Closely coordinated actions (e.g. dancing) may require enhancements to the framework
- Object recognition remains a difficult problem
- A language formalism can be provided for defining events to ease human communication
- Needs to extend high level interpretation logic
- Extension to multi-camera systems
- Integrates with other types of information
  - Face, gestures, sounds, text, etc.