Video Indexing and Retrieval

CMSC828K

Kyongil Yoon

Contents

- **Part 1 : Survey**

  "Multimedia Database Management Systems"
  Guojun Lu

  Chapter 7. Video Indexing and Retrieval

- **Part 2 : Example**

  Automatic Video Indexing via Object Motion Analysis
  Jonathan D. Courtney
  Texas Instruments

Introduction

- ■ **Video**
  - – **A combination of text, audio, and images with a time dimension**
- ■ **Indexing and retrieval methods**
  - – **Metadata-based method**
  - – **Text-based method**
  - – **Audio-based method**
  - – **Content-based method**
    - • **Video : A collection of independent images or frames**
    - • **Video : A sequence of groups of similar frames (shot-based)**
  - – **Integrated approach**

Shot-Based Video …

- Video shot : logical unit or segment
  - Same scene
  - Single camera motion
  - A distinct event or an action
  - A single indexable event
- Query
  - Which video?
  - What part of video?
- Steps
  - Segment the video into shots
  - Index each shots
  - Apply a similarity measurement between queries and video shots Retrieve shots with high similarities

Shot Detections (Segmentation)

- **Segmentation**
  - A process of dividing a video sequence into shots
- **Key issue**
  - Establishing suitable difference metrics
  - Techniques for applying them
- **Transition**
  - Camera break
  - Dissolve, wipe, fade-in, fade-out

Basic Video Segment Techniques

- **Sum of pixel-to-pixel differences**
- **Color histogram difference**
  - **To be tolerant with object motion**
  - $SD_i = \Sigma_j |H_i(j) - H_{i+1}(j)|$ where i : frame number, j : gray level
- **Modification of color histogram**
  - $SD_i = \Sigma_j ((H_i(j) - H_{i+1}(j))^2 / H_{i+1}(j))$
  - χ^2 **test**
- **Selection of appropriate threshold - Critical**
  - **e.g.) The mean of the frame-to-frame difference**
    + small tolerance value

Detecting Gradual Change

- Fade-in, fade-out, dissolve, wipe, …
- Twin-comparison technique
  - T_b : Normal camera breaks
    T_s : Potential frames of gradual change
  - If T_b < diff — shot boundary
    T_s < diff < T_b — accumulate differences
    diff < T_s — nothing
  - If the accumulated value is greater than T_b, a gradual change is detected.
- Detection techniques based on wavelet transformation
- Very hard to detect!

False Shot Detection

- **Camera panning, tilting, and zooming**
  - **Motion analysis techniques**
  - **Camera movements**
    - **Optical flow computed by block matching method**
- **Illumination change**
  - **Normalization of color images before carrying out shot detection**

1. $R_i' = R_i / Sqrt(\Sigma^N R_i^2), G_i' = \ldots, B_i' = \ldots$
2. Chromaticity
   1) $r_i' = R_i' / (R_i' + G_i' + B_i')$
   2) $g_i' = R_i' / (R_i' + G_i' + B_i')$
3. A combined histogram for r and g : CHI (Chromaticity histogram image)
4. Reduce it to 16x16
5. 2D DCT
6. Pick only 36 significant DCT values
7. Distances are calculated based on these values

Other Shot Detection

- **Motion removal**
  - Ideally, frame-to-frame distance should be
    - Close to zero with very little variation within a shot
    - Significantly larger than within-values between shots
  - However, within a shot
    - Object motion, camera motion, other changes
    - Filter to remove the effects of camera/object motion
- **Based on edge detection**
- **Advanced cameras**
  - Recording extra information such as position, time, orientation, …

Segmentation of Compressed Video

- Based on MPEG compressed video
  - DCT coefficients
  - Motion information
    - E.g. # of bidirectional coded macro blocks in B frame, it is very likely shot boundary occurs around the B frame
- Based on VQ compressed video

Video Indexing and Retrieval

- Shot detection is preprocessing for indexing

- R (representative) frames
  - One or more key frames for each shot
  - Retrieval is based on these frames
- Other information
  - Motion, objects, metadata, annotation

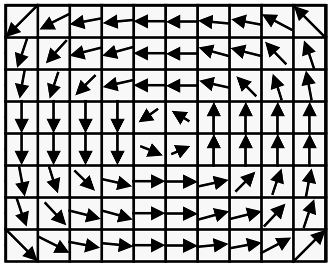
Based on R frames

- **An r frame captures the main content of the shot**
- **Image retrieval : color, shape, texture, …**
- **Choosing r frames**
  - **How many?**
    1. **One per shot**
    2. **The number of r frames according to their length**
    3. **One per subshot/scene**
  - **How to select?**
    1. **First frame of segment**
    2. **An average frame**
    3. **The frame whose histogram is closest to the average histogram**
    4. **Large background + all foregrounds superimposed**
  - **First frame + frame with large distance**

Based on Motion Information

- **R frame base ignores temporal or motion information**
- **Motion information is derived from optical flow or motion vectors**
- **Parameters for indexing**
  - **Content :** talking head vs car crash
  - **Uniformity :** smoothness as a function of time
  - **Panning :** horizontal camera movement
  - **Tilting :** vertical camera movement

- Camera motion
  - –Pan, tilt, zoom, swing, (horizontal/vertical) shift



Based on Objects

- Content based representation
- If one could find a way to distinguish individual objects throughout he sequence, …
- In a still image, object segmentation is difficult
  In a video sequence, we can group pixels that move together into an object.

- MPEG-4 object-based coding
  - How to represent
  - NOT how to segment and detect

Based on Others

- **Metadata**
  - **DVD-SI : DVD service information**
    Title, video type, directors
- **Annotation**
  1. **Manually**
  2. **Associated transcripts or subtitles**
  3. **Speech recognition on sound track**
- **Integrated method**

Effective Video Representation and Abstraction

- Useful to have effective representation and abstraction tool
- How to show contents in a limited space
- Applications
  - Video browsing
  - Presentation of video results
  - Reduce network bandwidth requirements and delay
- Then how?

Representation and Abstraction

- **Topical or subject classification**
  - **News : (local, international, finance, sport, weather)**
- **Motion icon (micon) or video icon**
  - **Easy shot boundary representation**
  - **Operations : browsing, slicing, extraction a subicon**
- **Video streamer**
- **Clipmap**
  - **A window containing a collection of 3D micons**
- **Hierarchical video browser**

Representation and Abstraction

- <u>Storyboard</u>
  - A collection of representative frames
- <u>Mosaicking</u>
  - An algorithm to combine information from a number of frames
- Scene transition graph
  - Node : image which represents one or more video shots
  - Edge : the content and temporal flow of video
- Video skimming
  - High-level video characterization, compaction, and abstraction

Automatic Video Indexing via Object Motion Analysis As an Object Tracking Example

- **Video indexing**
  - The process of identifying important frames or objects in the video data for efficient playback
  - Scene cut detection, camera motion, object motion
  - Hierarchical segmentation
- **Three steps**
  - Motion segmentation, object tracking, motion Analysis
- **Events**
  - Appearance/Disappearance
  - Deposit/Removal
  - Entrance/Exit
  - Motion/Rest

Motion Segmentation

- Segmented Image C_n
  - C_n = ccomps(T_h•k)
    Th : binary image resulting from thresholding $|I_n - I_0|$
    T_h•k : morphological close operation on T_h
  - Reference frame I_0

- Strong assumptions may fail when
  - Sudden lighting
  - Gradual lighting
  - Change of viewpoint
  - Objects in reference frame

Imperfectness of Segmentation

- **All the possible problems**
  - **True objects will disappear temporarily**
  - **False objects**
  - **Separate objects will temporarily join together**
  - **Single objects will split into multiple regions**

Object Tracking

- **Terminology**
  - **Sequence - ordered set of N frames**
    $S = \{F_0, F_1, \ldots F_{N-1}\}$ **:** F_i **is i-th frame**
  - **Clip** $C = (S, f, s, l)$ **:** F_f, F_l **- first and last valid frame,** F_s **- start frame**
  - **Frame** F **: image** I **annotated with a timestamp** t**,** $F_n = (I_n, t_n)$
  - **Image** I **:** r **x** c **array of pixel**
  - **Timestamp records the date and the time**
- **V-object**
  - **Extracted by motion segmentation comparing a frame to a reference frame**
  - **Label, centroid, bounding box, shape mask**
  - $V_n = \{ V_n^p ; p = 1, \ldots P \}$

Object Tracking

- **Tracking procedure**
  - **Iterate (forward) step 1-3 for frames 0, 1, …, N-2**

$$\mu_n^p = \mu_n^p + \nu_n^p(t_{n+1}-t_n)$$

1. For each V-object, predict its position in next frame

2. For each V-object, determine the V-object in the next frame with centroid nearest to the prediction

3. For every pair, estimate forward velocity

4. Do 1-3 in backward

- For all frames

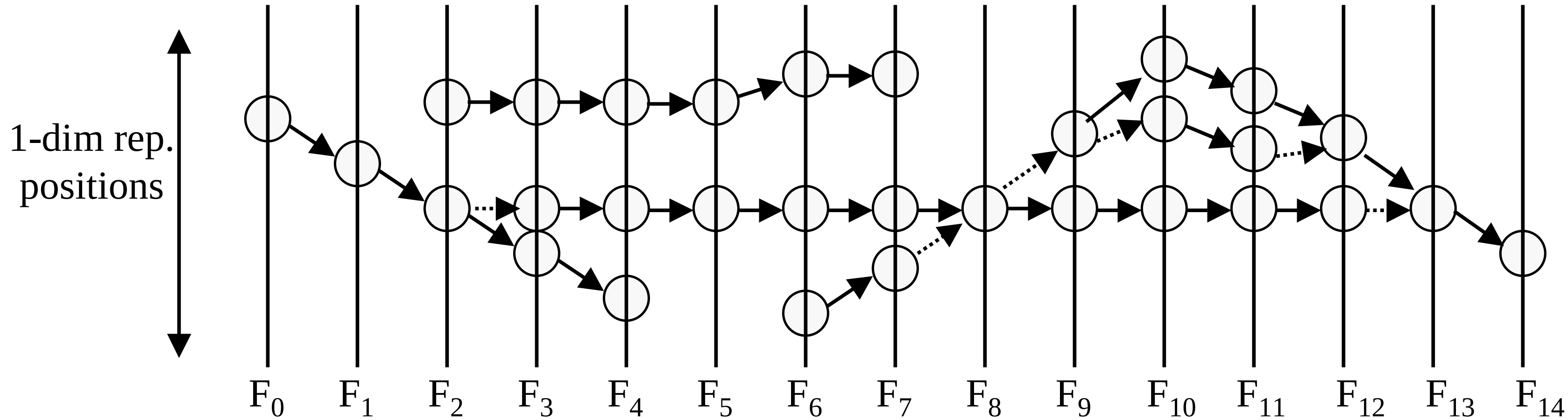
5. Determine primary links for mutual nearest neighbor

6. Determine secondary links from forward step

7. Determine secondary links from backward step

Object Tracking

- Following graph is produced
  - Node - V-objects
  - Primary links (mutually nearest)
  - Secondary links (others)



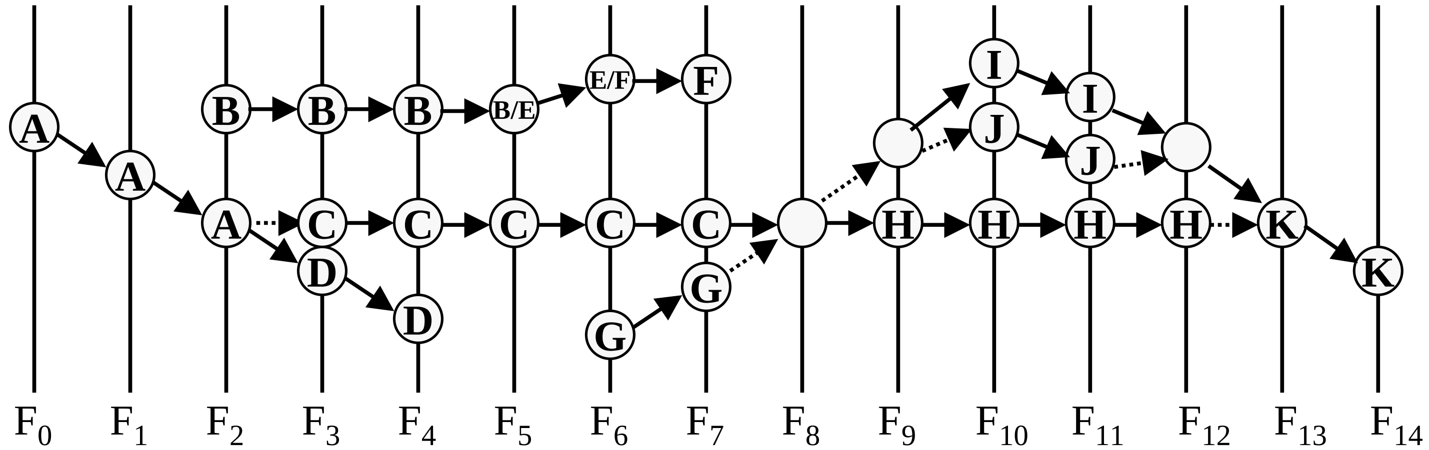
Motion Analysis
V-object Grouping

- **Group V-objects with difference levels**
  - **Stem, *M***
  - **Branch, *B***
  - **Trail, *L***
  - **Track, *K***
  - **Trace, *E***
- $E \supseteq K \supseteq L \supseteq B \supseteq M$
- **Each level implies a feature of the blob**

V-object Grouping - Stem

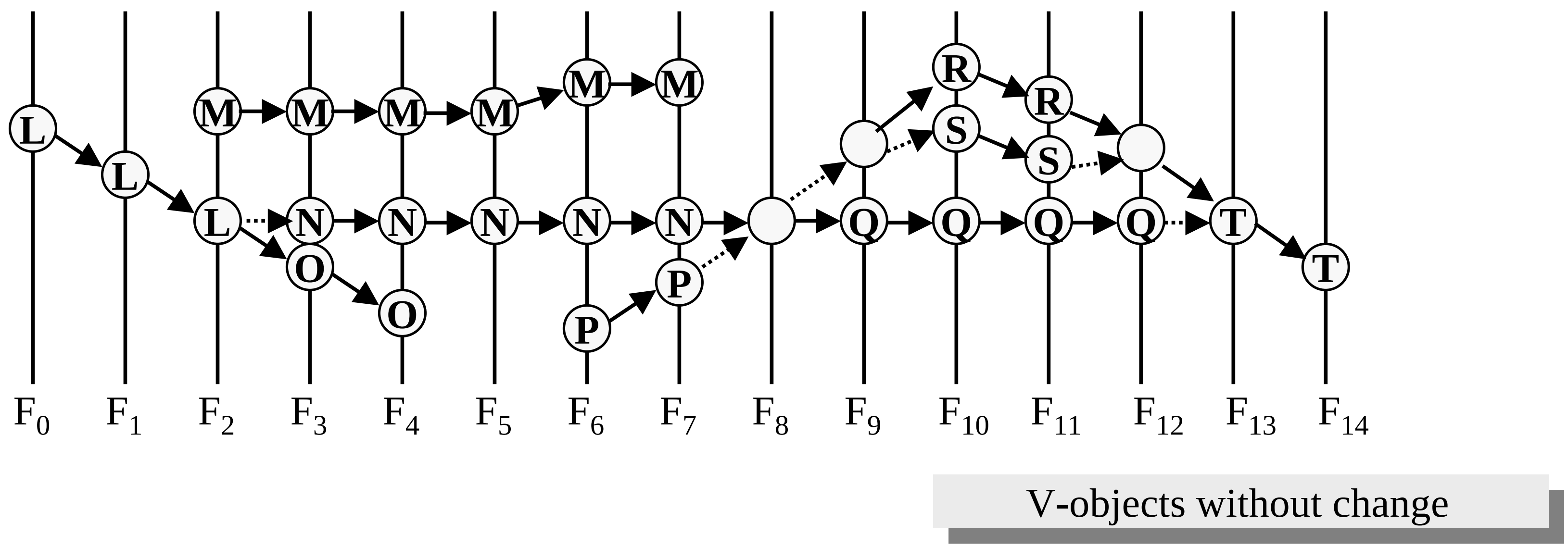
- **Maximal size path of two or more V-objects with no secondary links**
- $M = \{V_i : i = 1, 2, \ldots N_M\}$
  - outdegree(V_i) = 1 for $1 \le i < N_M$
  - indegree(V_i) = 1 for $1 < i \le N_M$
  - either $\mu 1 = \mu 2 = \ldots = \mu_{N_M}$
    or $\mu 1 \ne \mu 2 \ne \ldots \ne \mu_{N_M}$
- **Stationary/moving**

V-objects with constant/no movement without any change



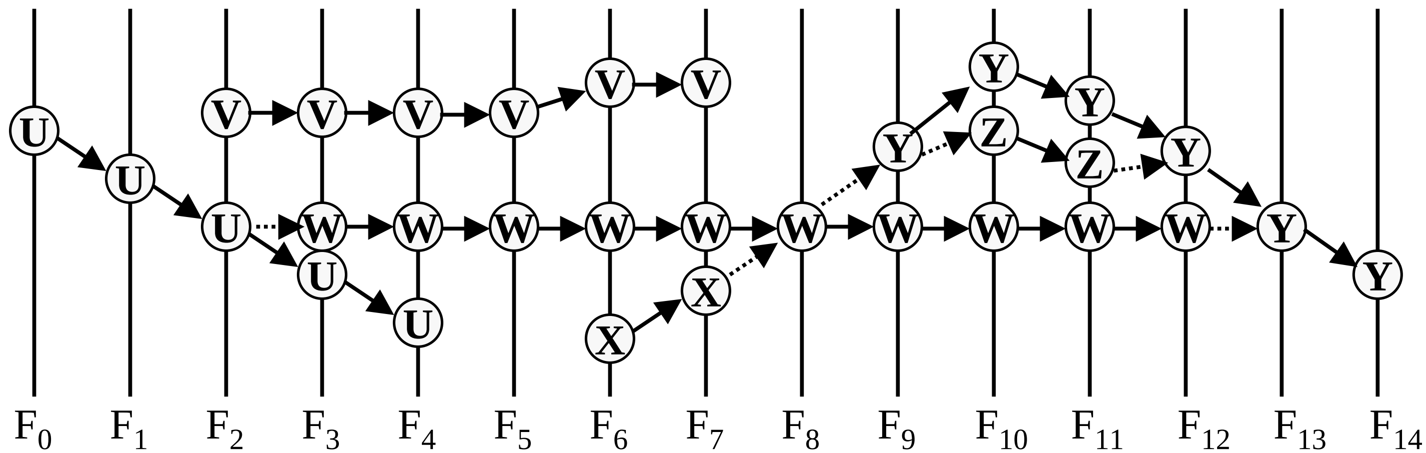
V-object Grouping - Branch

- **Maximal size path containing no secondary links and composed with only one path**
- $B = \{V_i : i = 1, 2, \ldots, N_B\}$
  - **outdegree(V_i) = 1 for $1 \le i < N_B$**
  - **indegree(V_i) = 1 for $1 < i \le N_B$**
- **Stationary(one stem) / moving(otherwise)**



V-object Grouping - Trail

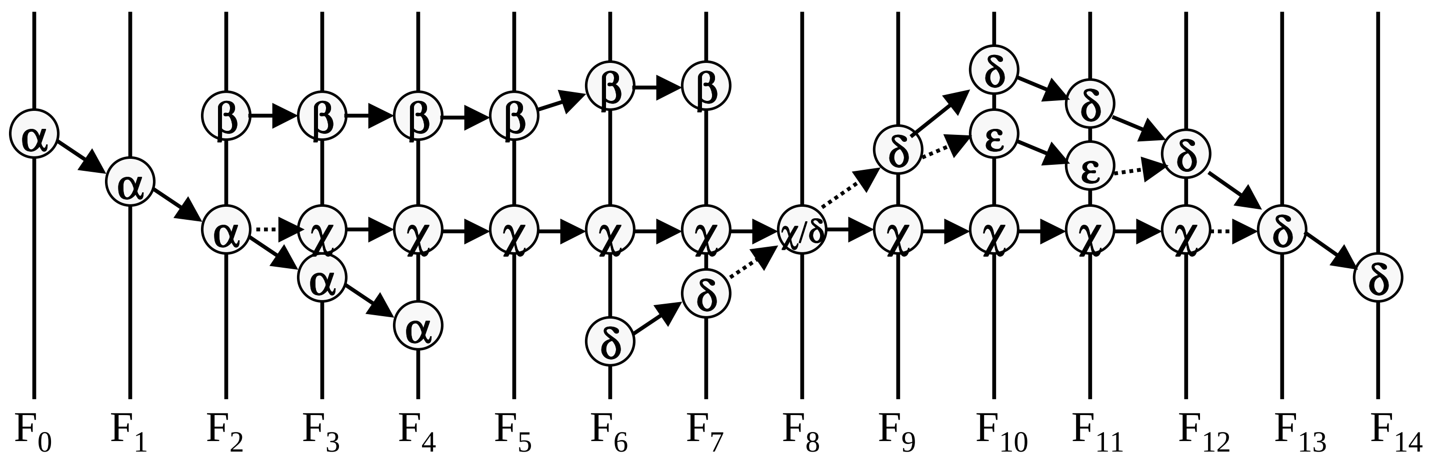
- ***L***
- **Maximal-size path without secondary links**
- **Stationary/moving/unknown**



V-objects with main movement

V-object Grouping - Track

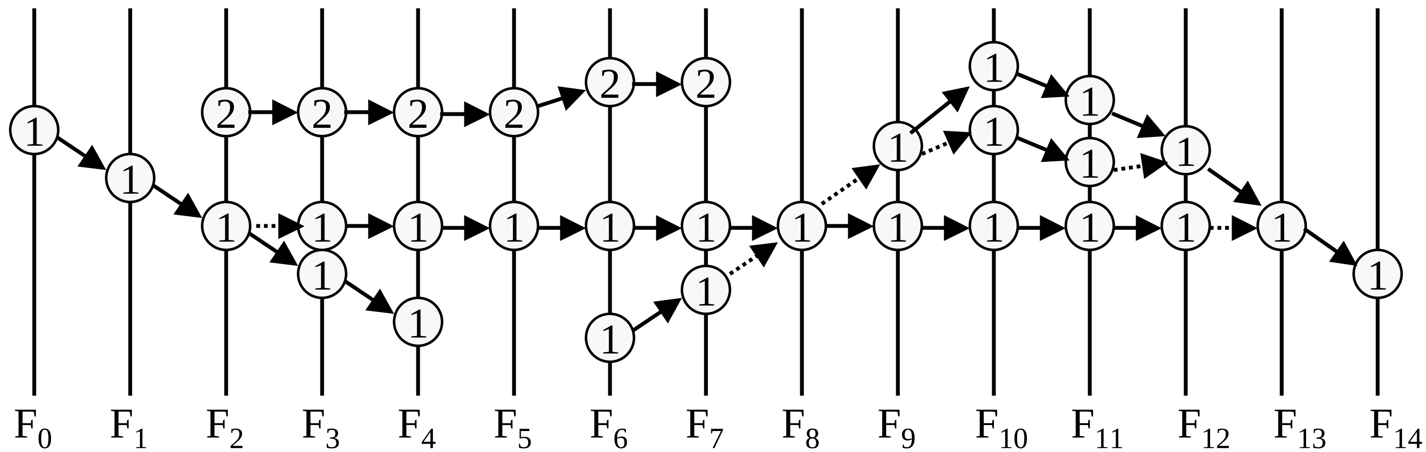
- $K = \{L_1, G_1, \ldots, L_{N_{K-1}}, G_{N_{K-1}}, L_{N_K}\}$
  - L_i : trail
  - G_i : connecting dipath with constant velocity through $H = \{V^l_i, G_i, V^1_{i+1}\}$ where V^l_i is the last object of L_i and V^1_{i+1} is the first object of L_{i+1}
- Stationary/moving/unknown



V-objects with movement guessing occluded part

V-object Grouping - Trace

- ***E***
- **Maximal size connected digraph of V-objects**



Group of V-objects overlapped

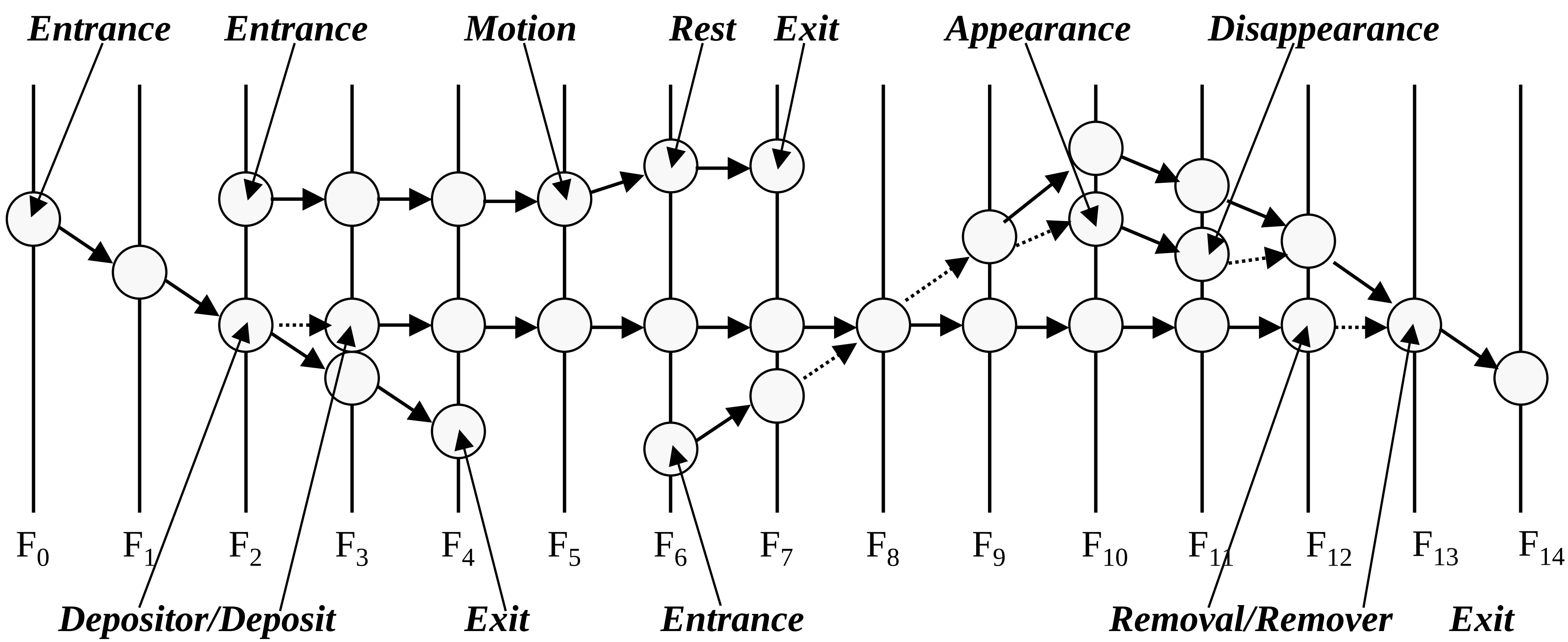
Events

- **Appearance** - an object emerges in the scene
- **Disappearance** - an object disappears from the scene
- **Entrance** - moving object enters the scene
- **Exit** - moving objects exits from the scene
- **Deposit** - an inanimate object is added to the scene
- **Removal** - an inanimate object is removed from the scene
- **Motion** - an object at rest begins to move
- **Rest** - a moving object comes to a stop
- **(Depositor)** - a moving object adds an inanimate object to the scene
- **(Remover)** - a moving object removes an inanimate object from the scene

Annotating V-objects

	V-object motion state		
	Moving	Stationary	Unknown
Appearance	1. Head of track 2. Indegree(*V*) > 0	1. Head of track 2. Indegree(*V*) = 0	
Disappearance	1. Tail of track 2. Outdegree(*V*) > 0	1. Tail of track 2. Outdegree(*V*) = 0	
Entrance	1. Head of track 2. Indegree(*V*) = 0		1. Head of track 2. Indegree(*V*) = 0
Exit	1. Tail of track 2. Outdegree(*V*) = 0		1. Tail of track 2. Outdegree(*V*) = 0
Deposit		1. Head of track 2. Indegree(*V*) = 1	
Removal		1. Tail of track 2. Outdegree(*V*) = 1	
(Depositor)	Adjacent to V-object with deposit tag		
(Remover)	Adjacent from V-object with removal tag		
Motion	1. Tail of stationary stem 2. Head of moving stem		
Rest	1. Tail of moving stem 2. Head of stationary stem		

Example of Annotation



Query

- **Y = (*C*, *T*, *V*, *R*, *E*)**
  - *C* **: a video clip**
  - ***T* = (t_i, t_j) : a time interval within the clip**
  - ***V* : V-object in the clip**
  - ***R* : a spatial region in the field of view**
  - ***E* : an object motion event**
- **Processing a query**
  - **Keeps truncating domain with query parameters**

Experimental Result

- 3 videos, 900 frames, 18 objects, 44 events
-

Video 1	Video 2	Video 3
Inventory or Security monitoring 300 frs, 10fr/sec 5 objects, 10 events entrance/exit, deposit/removal	retail customer monitoring 285 frames, 10 fr/sec 4 objects, 14 events all eight events 3 foreground objects in ref. frame Most complicated	parking lot traffic monitoring 315 frames, 3fr/sec 9 objects, 20 events most noisy

- 1 false negative, 10 false positive
- Conservative

Errors come from

- Noise in the sequence
- Assumption of constant trajectories of occluded objects
- No means to track objects through occlusion by fixed scene objects

Mosaicking

Story board, Video Multiplexing

- Show 20 minutes of video in 6 seconds
- Loop all shots as thumbnails at same time
- Let the user focus on the interesting shots

Micon