ECE624

Week 4-a

Querying Video Database

- Query involving spatio events:
  - "Show that segment of the sonogram where the main artery is 90% blocked".
- Query involving spatio-temporal events:
  - "Get the information conatining video clip where assembly of Honda Accord's gear-box is described"
- Query involving searching of audio data associated with video data.
  - "Get all the documents containing video clips where Prime Minister Tanaka, during one of his Cabinet meetings, has mentioned trades with US".

Video Databases: Indexing Issues

Indexing based on:

- **Objects and Faces (fine-grained/ frame-level indexing)**
- **Segments and scene changes (medium grained indexing)**
- **Complex Spatio-Temporal Events (coarse-grained indexing)**

Multilevel Spatio-Temporal Abstraction of Image/Video Data

- **Spatial:**

  Video Frame/Image -----> Object Descriptors ------> Image Features -----> Spatial-Semantics of Objects (human) ------> Semantic (President)

- **Temporal:**

  Sequence of Frames ----------> Object identification (Bounding Box) ----------> Inter- /Intra Frame Analysis ----------> (Motion Tracking) ----------> Inter-Object Movement Analysis --------> Semantic (Description of Event or Activity)

Event Specification

- **Formalism I:**
  - **Spatial events in an image/video frame**
  - **Temporally related spatial events ----> Temporal events**
  - **Aggregation/linking of temporal events ----> Complex Temporal Events**
  - **Useful existing Object-Oriented and AI tools: Abstractions, temporal logic, predicate calculus,**
- **Formalism II: Event-by-Example**

Purdue University

School of Electrical Engineering

Semantic Modeling and Knowledge Representation in Video Databases

- **The key characteristic of video data that makes it different from a-isochronous data such as text, image, and maps is its spatio-temporal dimension.**

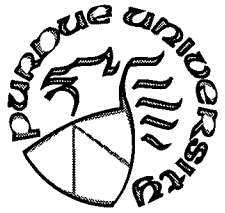
- **An example of a query in video events is:** ***Find video clips in which the dissection of liver is described***

- **Considerations in video data modeling**
  - how to specify spatio-temporal semantics and how to develop efficient indexing mechanisms, based on events and episodes in video data.
  - how to deal with the heterogeneity that may exist among semantics of such data due to differences in the interpretations of information in a video clip by different sets of users.

- **In general, most of the semantics and events in video data can be expressed by describing the interplay among physical objects in time along with spatial relationships between these objects.**

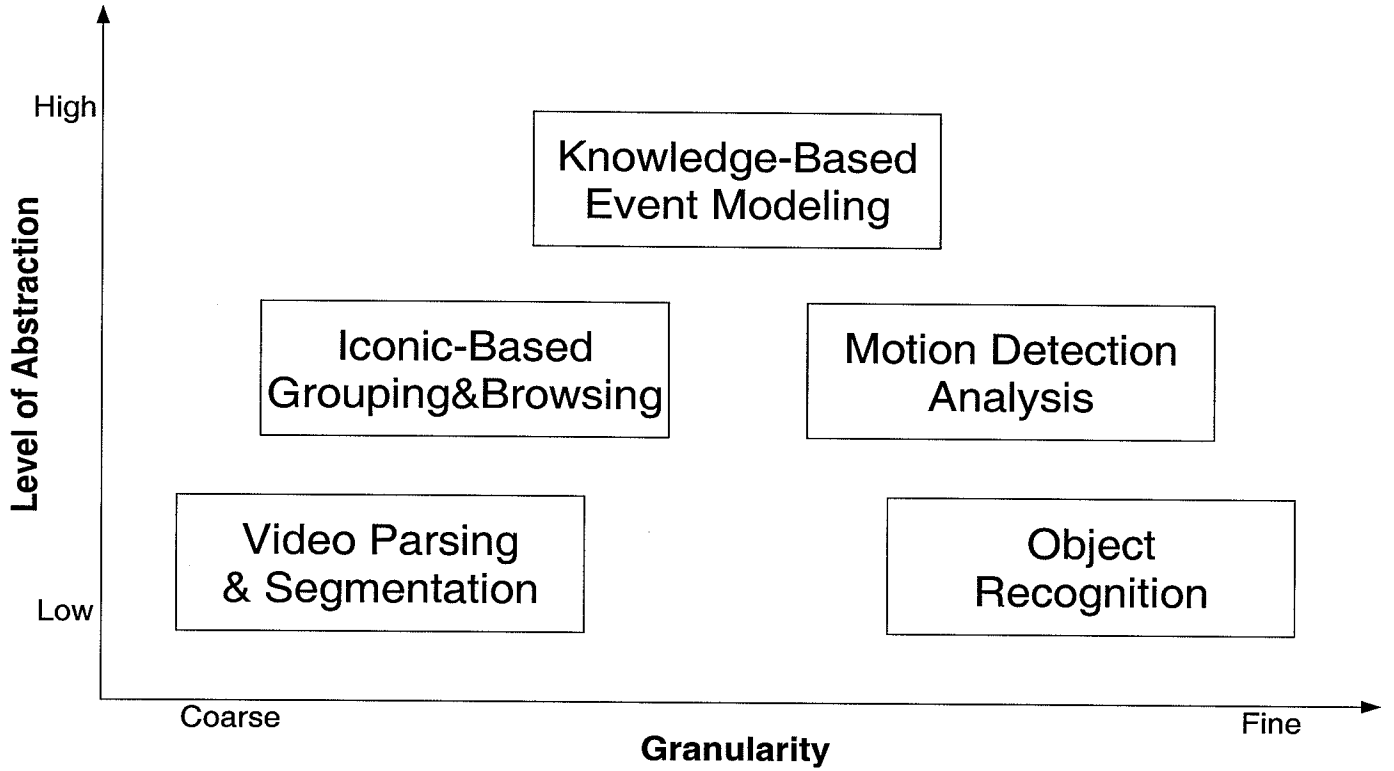
Purdue University

School of Electrical Engineering



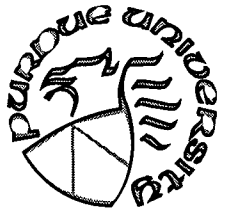
Semantic Modeling of Video Data

- **Two criteria for classifying existing approaches of modeling video data.**
  - Level of Abstraction
  - Granularity



Purdue University

School of Electrical Engineering



Video Parsing and Segmentation

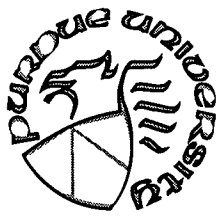
- **Employs image processing techniques to extract certain global features from individual video frames.**

  - Any significant change in the feature value in a sequence of frames is used to mark a change in the scene. The process allows a high level segmentation of video data into several shots.

  - Several scene change detection methods have been proposed in the literature.
    - pixel-level comparison
    - likelihood ratio
    - color histogram
    - chi-square histogram
    - DCT-based approaches
  - Detection of camera motion using optical flow techniques

Purdue University

School of Electrical Engineering



Pixel-Level Comparison

- The Number of pixels in frame F_{i+1} that changed its intensity from the corresponding pixels in frame F_i are counted.
- The percentage difference between the pixels in the two frames can be calculated as:

$$\Delta_i(x, y) = \begin{cases} 1 & \text{if } (|F_i(x, y) - F_{i+1}(x, y)| > t) \\ 0 & \text{Otherwise} \end{cases}$$

$$\frac{\sum_{x, y = 1}^{X, Y} \Delta_i(x, y)}{X \times Y} \times 100 > T$$

- The method is sensitive to noise, change in illumination, and camera movement.

Likelihood Ratio

- A likelihood ratio approach assumes uniform second order statistics over a region.
- The frames frames are divided into blocks and then blocks compared based on their statistical characteristics of their intensity levels.

$$\lambda = \frac{\left[\frac{\sigma_i + \sigma_{i+1}}{2} + \left(\frac{\mu_i - \mu_{i+1}}{2}\right)^2\right]^2}{\sigma_i \times \sigma_{i+1}}$$

$$\Delta_i(k,l) = \begin{cases} 1 & \text{if} \quad \lambda > t \\ 0 & \text{Otherwise} \end{cases}$$

- The method is sensitive to change in illumination and camera movement

Histogram Comparison

- The histogram of each frame is computed and compared.

$$\sum_{j=1}^{G} \left| H_i(j) - H_{i+1}(j) \right| > t$$

Purdue University

School of Electrical Engineering

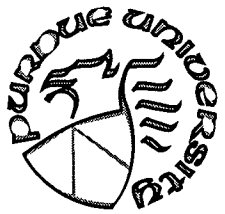
Chi-Square Color Histogram

- This function uses the square of the difference between the two histograms so as to strongly reflect the difference

$$\sum_{j=1}^{G} \frac{\left|H_i(j) - H_{i+1}(j)\right|^2}{H_{i+1}(j)} > t$$

Purdue University

School of Electrical Engineering



DCT Coefficients-Based Approaches

- Compressed video data is used to detect camera breaks
- Compression is carried out by:
  - dividing the image into a set of 8x8 pixel blocks
  - Pixels in the blocks are transformed into 64 coefficients using Discrete Cosine Transform, which are quantized and Huffman entropy encoded
- The coefficients in the frequency domain are mathematically related to the spatial domain, therefore can be used to detect changes in video sequences
  - Given 8x8 blocks of a single DCT-based encoded video frame f, a subset of blocks is chosen from n connected regions in each frame apriori
  - For each block, a subset of the 64 coefficients is chosen
  - Taking coefficients from each frame, a vector is formed:

$$V_f = \{c_1, c_2, c_3, \ldots\}$$

  - The inner product is used to find the difference between the two frames

$$\Psi = \frac{V_f \bullet V_{f+1}}{|V_f||V_{f+1}|}, \qquad 1 - |\Psi| > t$$

Purdue University

School of Electrical Engineering

Iconic-Based Grouping and Browsing

- **Video segmentation techniques are also suitable for building iconic based browsing environments.**

- ***Representative Frame(s)* of each scene can be displayed to the user in order to provide the information about the objects and possible events present in that scene.**

- **Example 1 (Yeung et al)**
  - Use directed graphs to portray an overall "visual" summary of a video clip consisting of different scenes. Here, nodes represent representative frames and edges denote the temporal relationships between them.
- **Example 2 (Chen et al)**
  - A similarity pyramid is proposed giving a hierarchical clustering of all representative frames present in the video database.
  - Organization of this pyramid is based on extracted features and user feedback.

Purdue University

School of Electrical Engineering

Motion Detection Analysis

- **Capture of motion information (about salient objects and persons).**

- **One approach modifies known compression algorithms such as MPEG to identify objects and to track their motion.**
  - A motion tracking algorithm using forward and backward motion vectors of macro-blocks used by MPEG encoding algorithm to generate trajectories for objects.

- **Second approach uses a directed graph model to capture both spatial and temporal attributes of objects and persons. This is achieved by specifying the changes in the 3D projection parameters associated with the bounding volume of objects in a given sequence of frames.**

Purdue University

School of Electrical Engineering

Knowledge-Based Event Modeling

- **Based on the information available from the low layers, higher level events can be specified by the user to construct different views of the video data.**

- **Higher level semantics can be derived from either video parsing and segmentation or from temporal modeling and specification of events present in video data**

- **Video parsing and segmentation-based higher level semantic modeling**
  - Identify key objects and other features within each scene or using textual annotation available from captions.
    - For example, Chen et al. use a set of features of video frames along with the motion vectors and shot duration classifying the data into pseudo-semantic classes corresponding to head and shoulders, indoor versus outdoor, high action, and man-made versus natural.
    - Smoliar et al. take advantage of the well-structured domain of news broadcasting to build an *apriori model* consisting of reference frames as a knowledge base to semantically classify the video segments of a news broadcast.