CISC 7610 Lecture 11
Video retrieval

Topics:

Structure analysis

Feature extraction

Video analysis

Query and retrieval

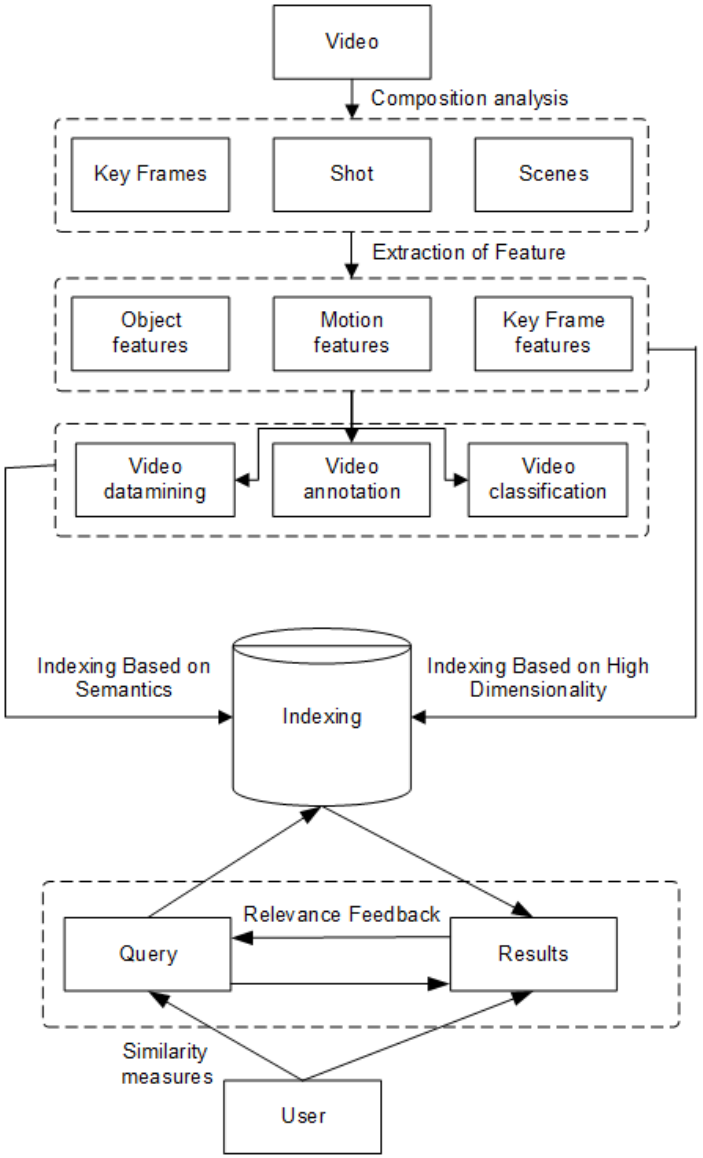
Summarization and browsing

What types of video are we indexing?

What types of video are we indexing?

- Web video: huge repositories of relatively short videos
- Remote instruction: video lectures, presentations, interaction
- TV shows: Sit-coms, sports, news
- Films: Action, comedy, drama, animation
- Surveillance
- Each type has its own set of conventions / properties

Video retrieval process overview



- Video retrieval is like image retrieval plus the aspect of time
- Temporal continuity can help object identification and segmentation
- Motion features can help identify objects and actions
- Good overview: Hu et al (Trans SMC-C, 2011)

Asghar, Hussain, & Manton, (2014). Video indexing: a survey. Int. J. Comp. & Info. Tech., 3(01).

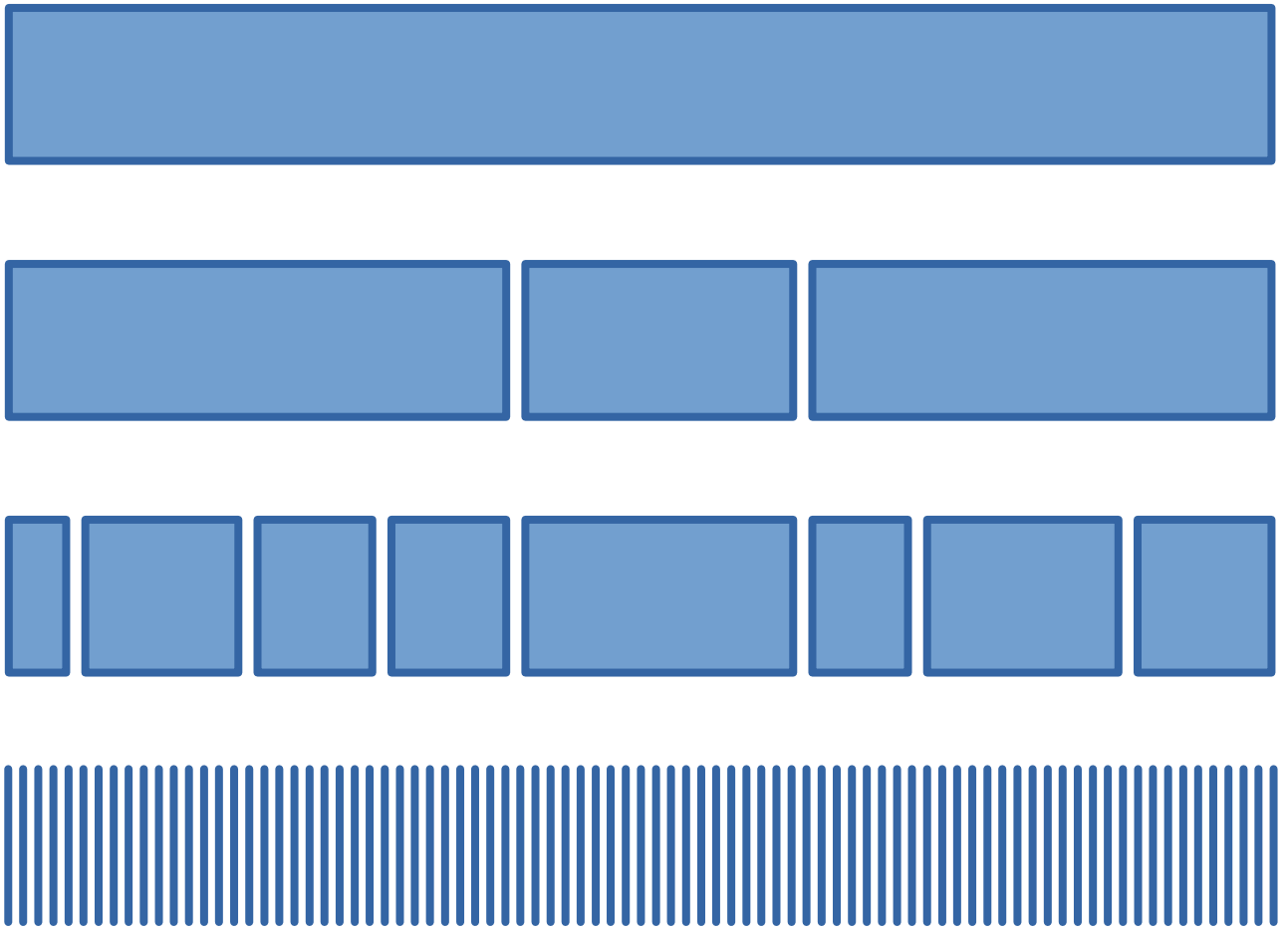
Videos can be decomposed into

Videos can be decomposed into scenes, shots, and frames

- Clip
- Scenes
- Shots
- Frames

Videos can be decomposed into scenes, shots, and frames

- Clip
- Scenes
- Shots
- Frames



Pre-processing: Predicting video structure

- Greatly improves subsequent processing
  - Allowing analysis of homogeneous video regions
- Predict shot boundaries and scene boundaries
- Identify key frames for shots and scenes

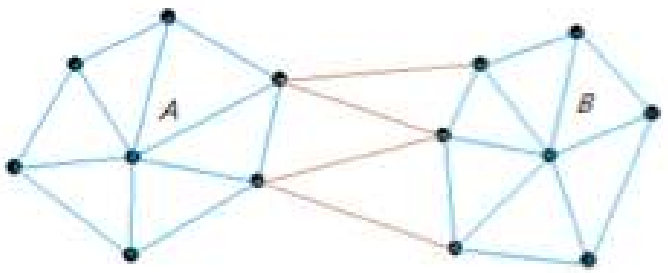
Predicting video structure: Scene segmentation

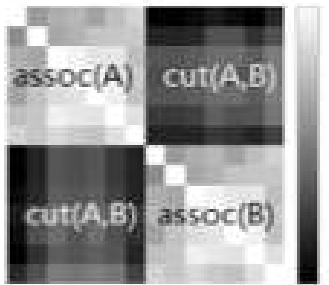
- A scene is a group of contiguous shots that occur at a single place in continuous time
- Three types of features: key frames, audio-visual, background color/texture
- Four types of approaches: merging, splitting, model-based, shot boundary classification

Predicting video structure: Shot boundary detection

- Two types of shot boundaries: cut (easier to find) and gradual (harder to find)
- General approach
  - Extract features from frames
  - Compute similarities between frames
  - Look for transitions

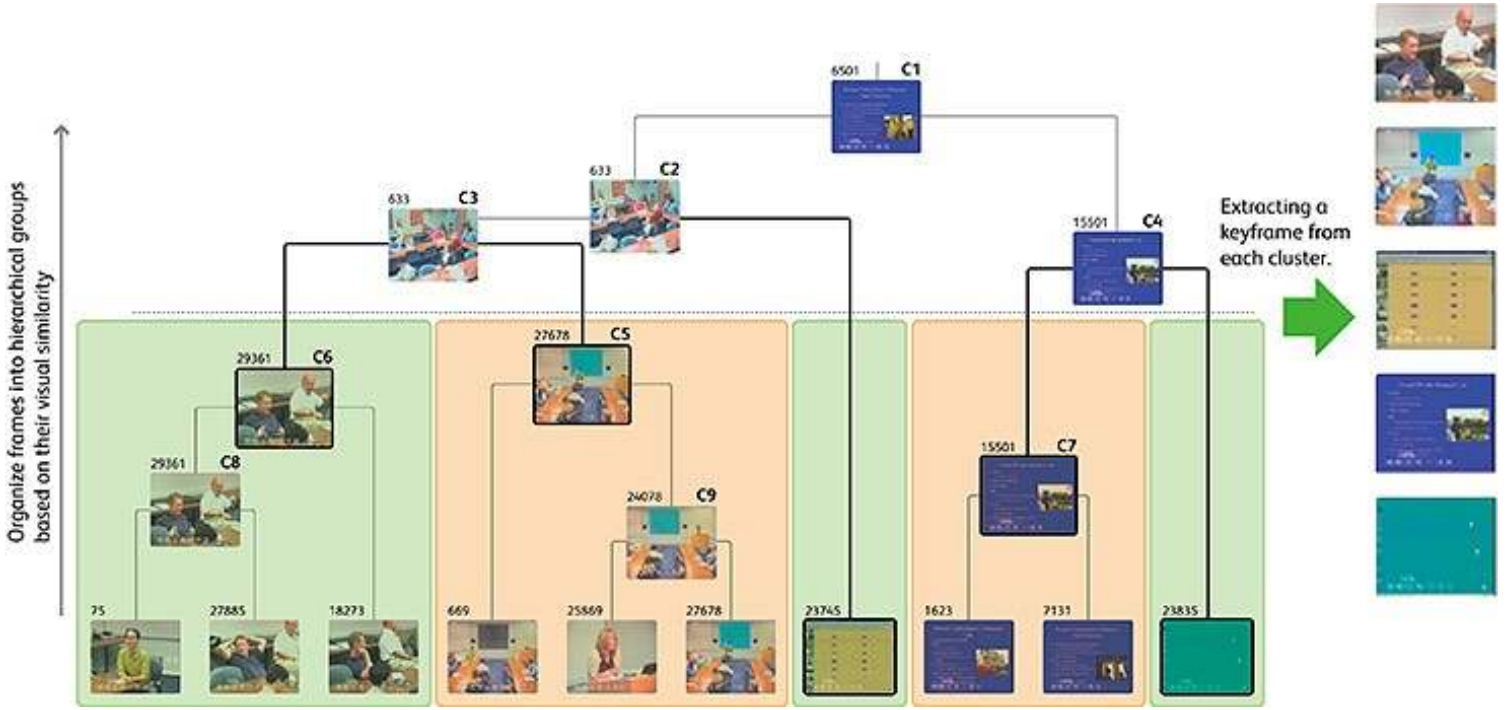
K frames in a window (K = 14)





Predicting video structure: Key frame extraction

- For various interface and analysis purposes, we usually want a still image to summarize a clip, scene, or shot
- Called “key frame extraction”, can be based on
  - Visual similarity, clustering, objects/events, motion curves
- Hard to evaluate because there is no one right answer

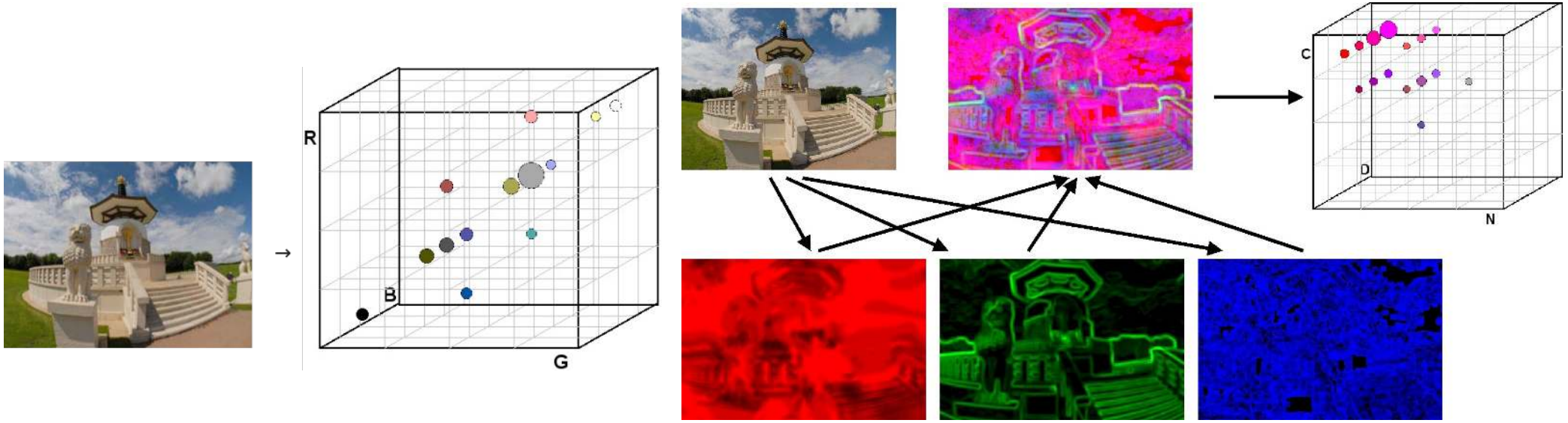


Video features

- Once videos are segmented, each segment is characterized by features
- Can use image features, but can also take advantage of temporal continuity and motion

Video features: (key) frame features

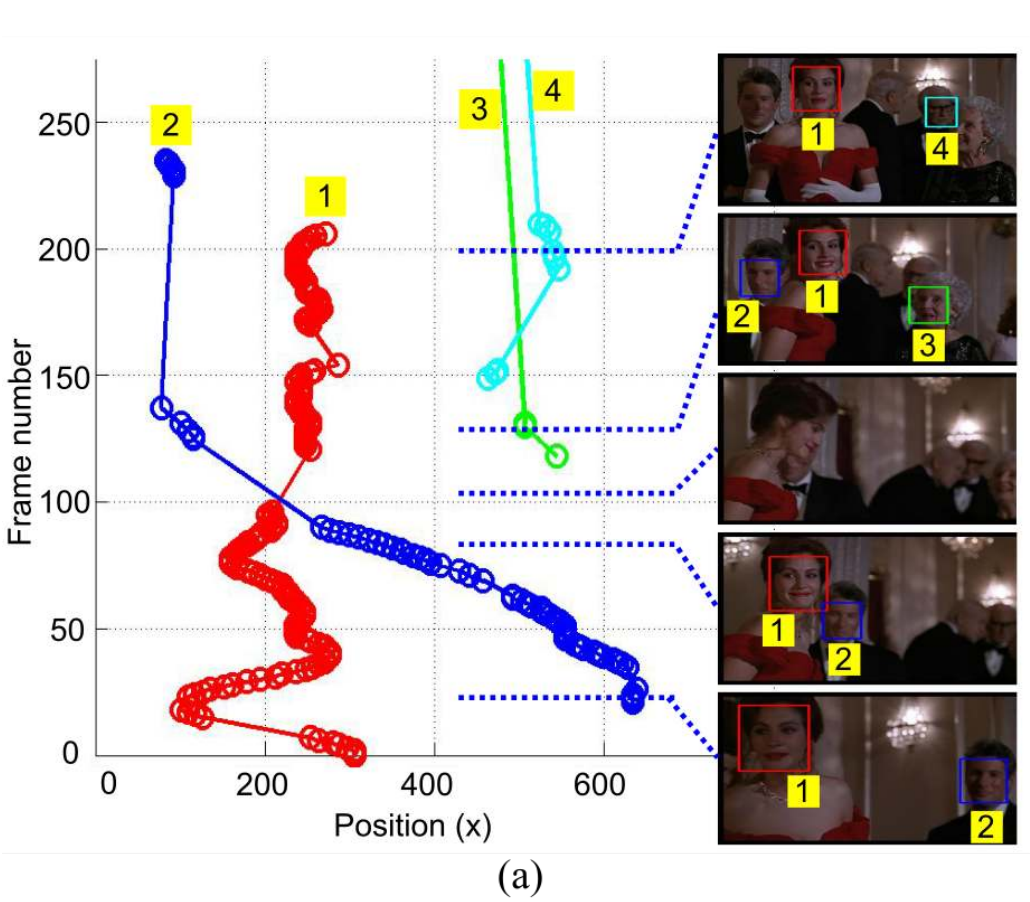
- Can analyze key frames as images
  - Color, texture, shape features
- Pro: easy and fast to compute, usually works ok
- Con: ignores motion or changes over time



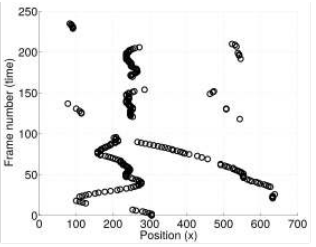
Video features: object features

- Goal: find videos with similar objects
- First, identify objects of interest in video
- Then compute features on those objects
- Most useful for specific kinds of objects (eg, faces)

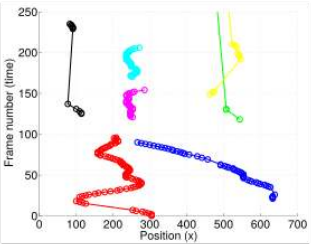
Video features: object features



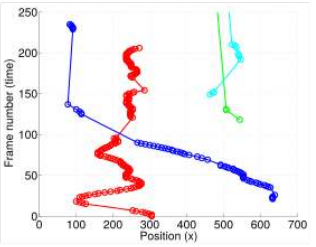
(a)



(b) – raw face detections



(c) – temporal association



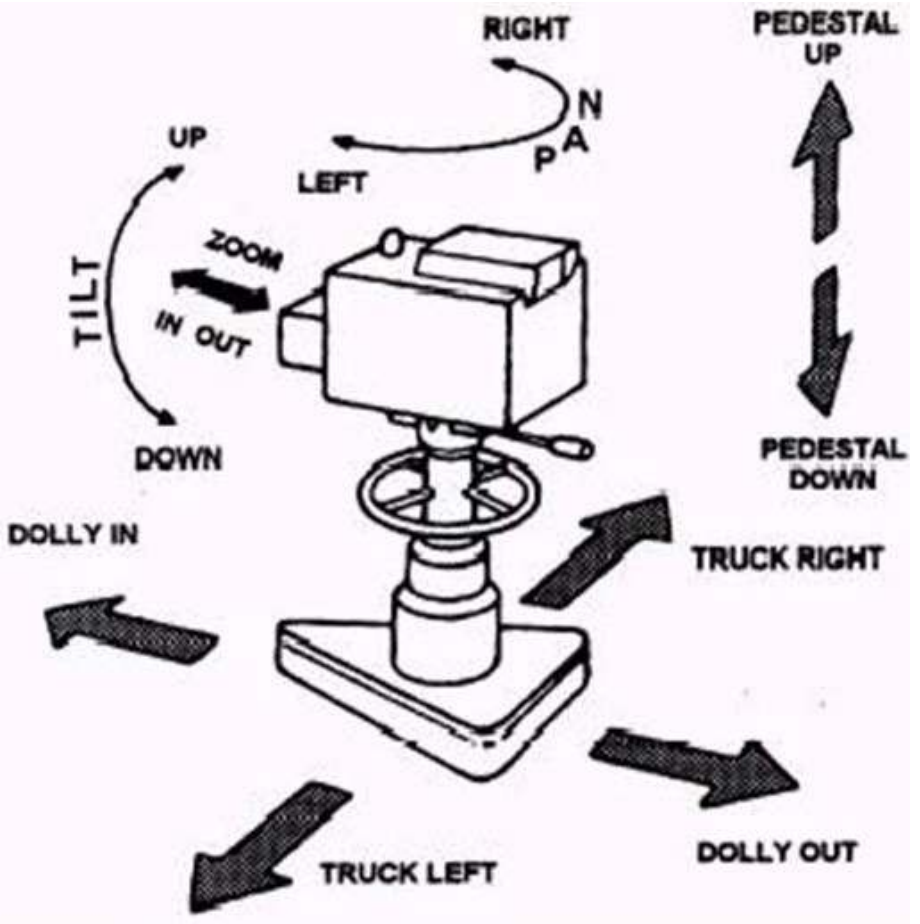
(d) – intra-shot matching

Video features: motion features

- Motion separates video indexing from still images
- Two kinds of motion: camera and object
- Camera:

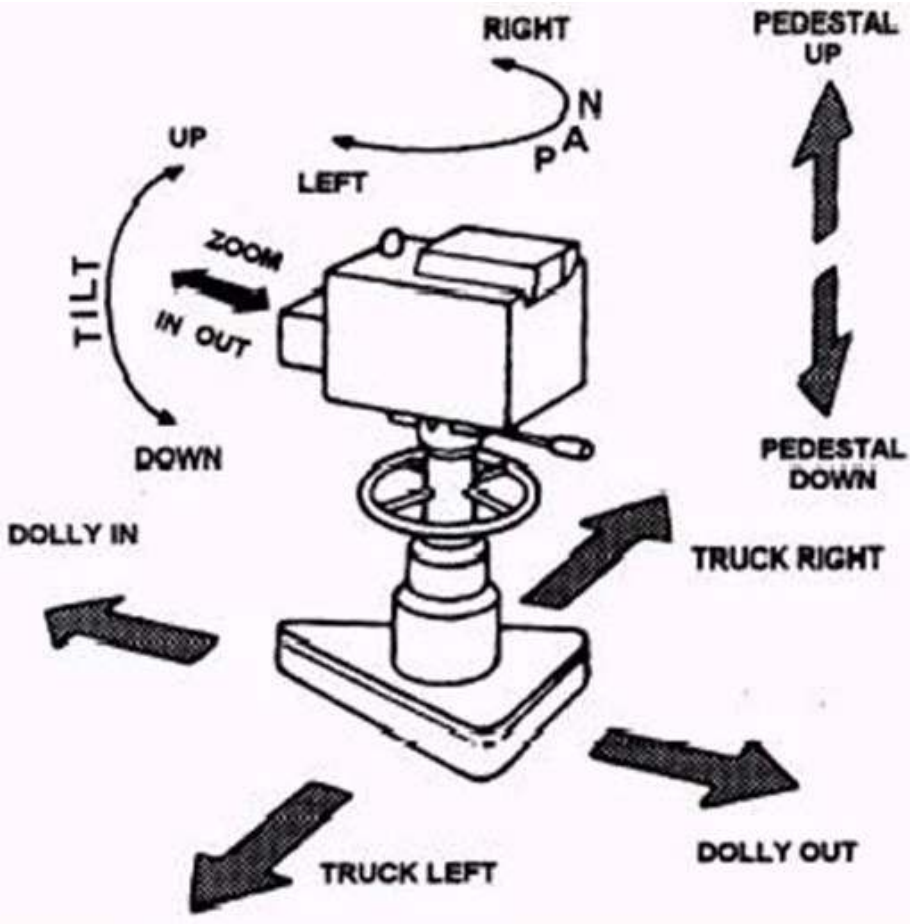
Video features: motion features

- Motion separates video indexing from still images
- Two kinds of motion: camera and object
- Camera: pan, tilt, zoom, pedestal, dolly, truck
  - Pro: easy to compute
  - Con: not useful for retrieval alone
-



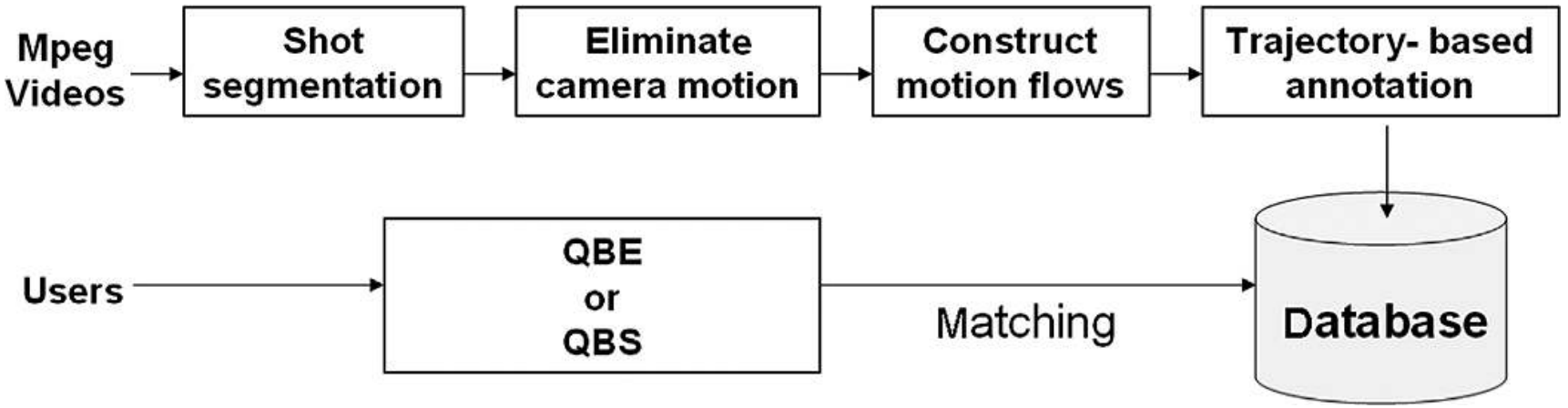
Video features: motion features

- Motion separates video indexing from still images
- Two kinds of motion: camera and object
- Camera: pan, tilt, zoom, pedestal, dolly, truck
  - Pro: easy to compute
  - Con: not useful for retrieval alone
- Object motion: Statistics, trajectories, relationships



Video features: object motion features

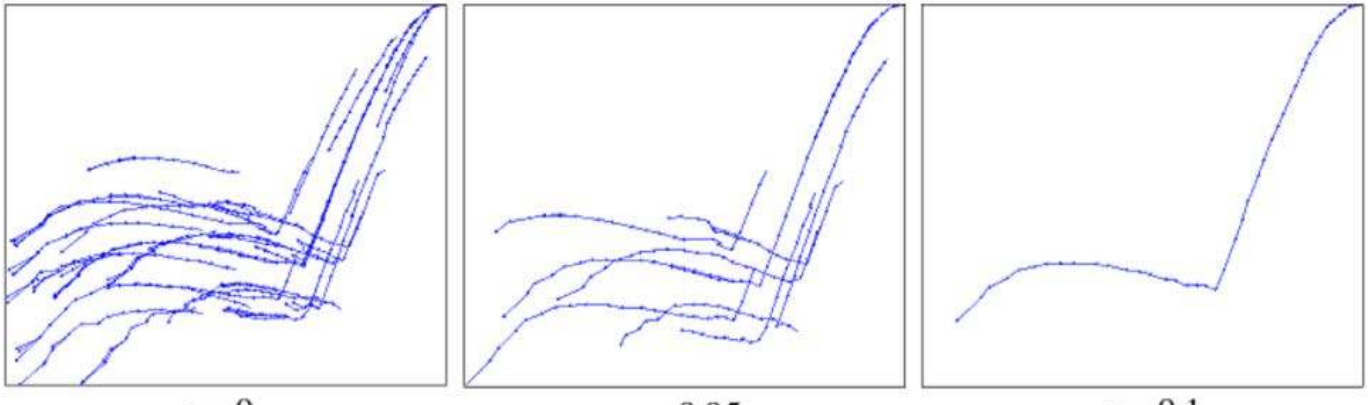
- Generally requires segmenting or identifying objects first, which is hard
- Possible without performing object segmentation from MPEG motion pointers (Su et al, 2007)



Video features: object motion features

- Generally requires segmenting or identifying objects first, which is hard
- Possible without performing object segmentation from MPEG motion pointers (Su et al, 2007)

(a)



(b)

Video analysis

- Data mining: unsupervised pattern discovery
- Classification: supervised concept recognition
- Annotation: supervised prediction of textual description of a video
- Generally difficult to draw hard boundaries between these activities

Video data mining

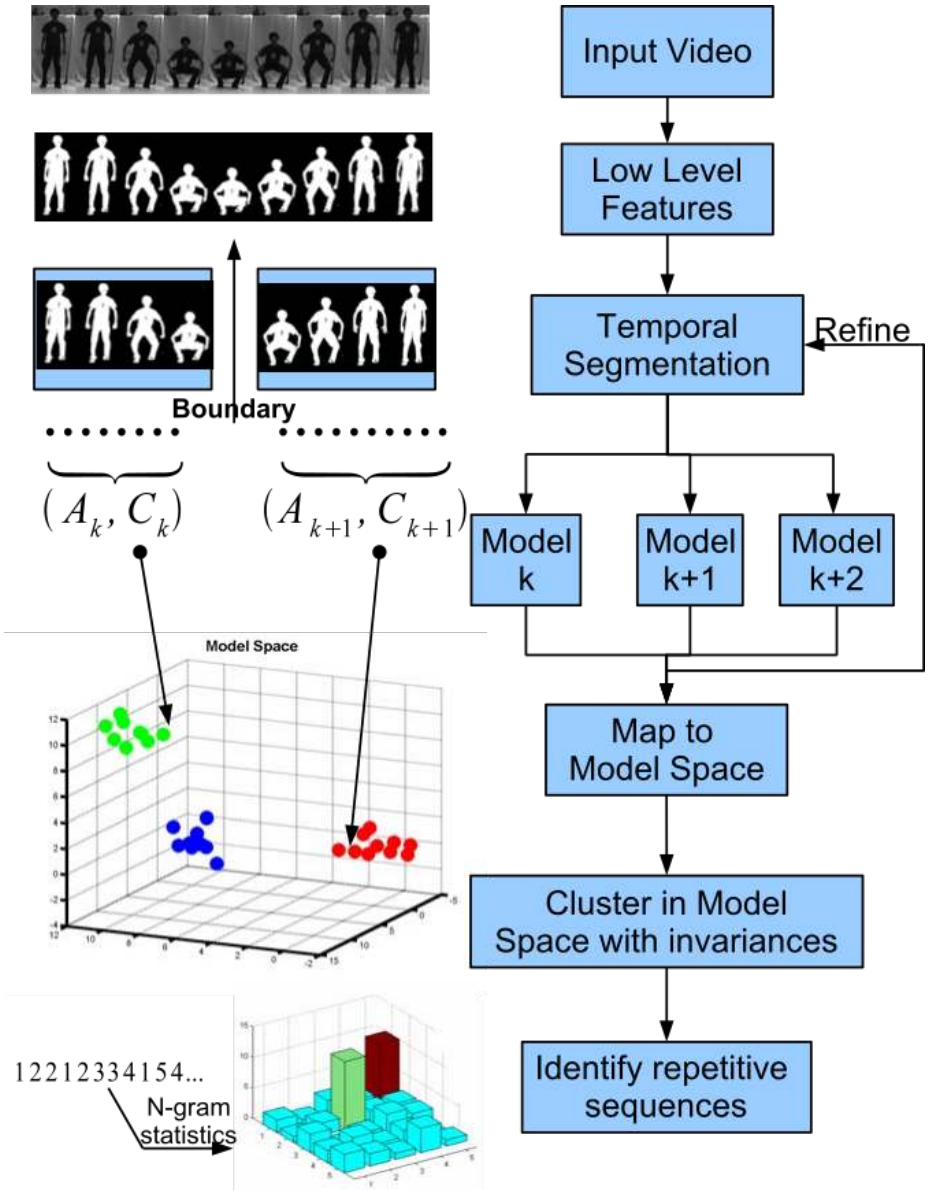
- Data mining is the unsupervised identification of patterns
- For video these patterns can include recurring objects, actions, and behaviors
- Or unusual objects, actions, and behaviors
- Useful for surveillance, video structure analysis, data exploration, subsequent classification training

Video data mining: Object mining

- Group different instances of the same object in a video
- Difficult because of changes in lighting, camera angle, deformation, etc
- Sivic and Zisserman (2004) use SIFT features

Video data mining: pattern discovery

- Recognize sequences of frames that recur
- Turaga et al (CVPR 2007) Use models that are invariant to camera angle, lighting, etc.
- Can find semantically meaningful clusters

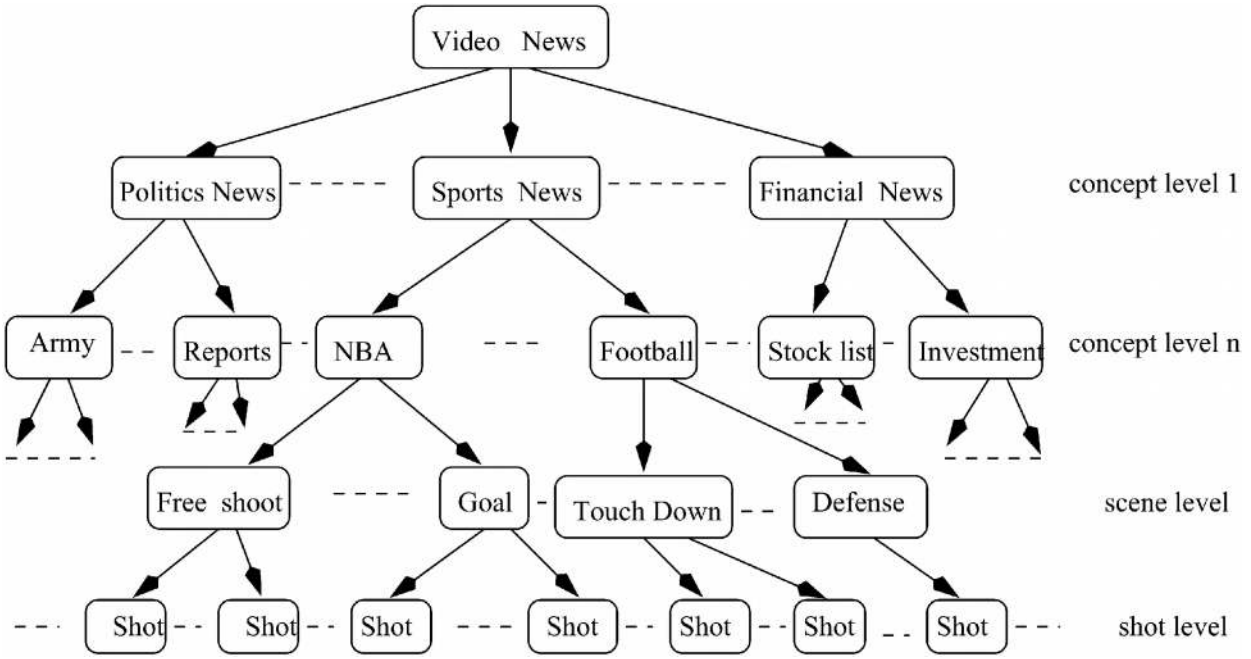


Video classification

- Difficult because of the gap between the input features (color, texture, shape) and a human's interpretation of this information
- Classification typically refers to full videos, annotation to parts
- Can classify editing effects, and semantic information, including
  - Video genre
  - Video shots
  - Video event recognition
  - Video object recognition

Classification: genre

- Various types of genres to predict
- Film genre: comedy, action, drama, horror, etc
  - Rasheed et al (TCSVT, 2004) separate using average shot length, color variance, motion content, and lighting key
- Genre is really hierarchical (Fan et al, TMM 2004)



Classification: shots

- Identify the type of shot from a finite set
- Straightforward in certain types of videos
  - E.g., sports
- Ekin et al (TIP, 2003) classify soccer shots into:

Classification: shots

- Identify the type of shot from a finite set
- Straightforward in certain types of videos
  - E.g., sports
- Ekin et al (TIP, 2003) classify soccer shots into: long, medium, closeup, and non-field

Classification: shots

- Identify the type of shot from a finite set
- Straightforward in certain types of videos
  - E.g., sports
- Ekin et al (TIP, 2003) classify soccer shots into: long, medium, closeup, and non-field

Long Shot

Long Shot

Medium Shot

Medium Shot

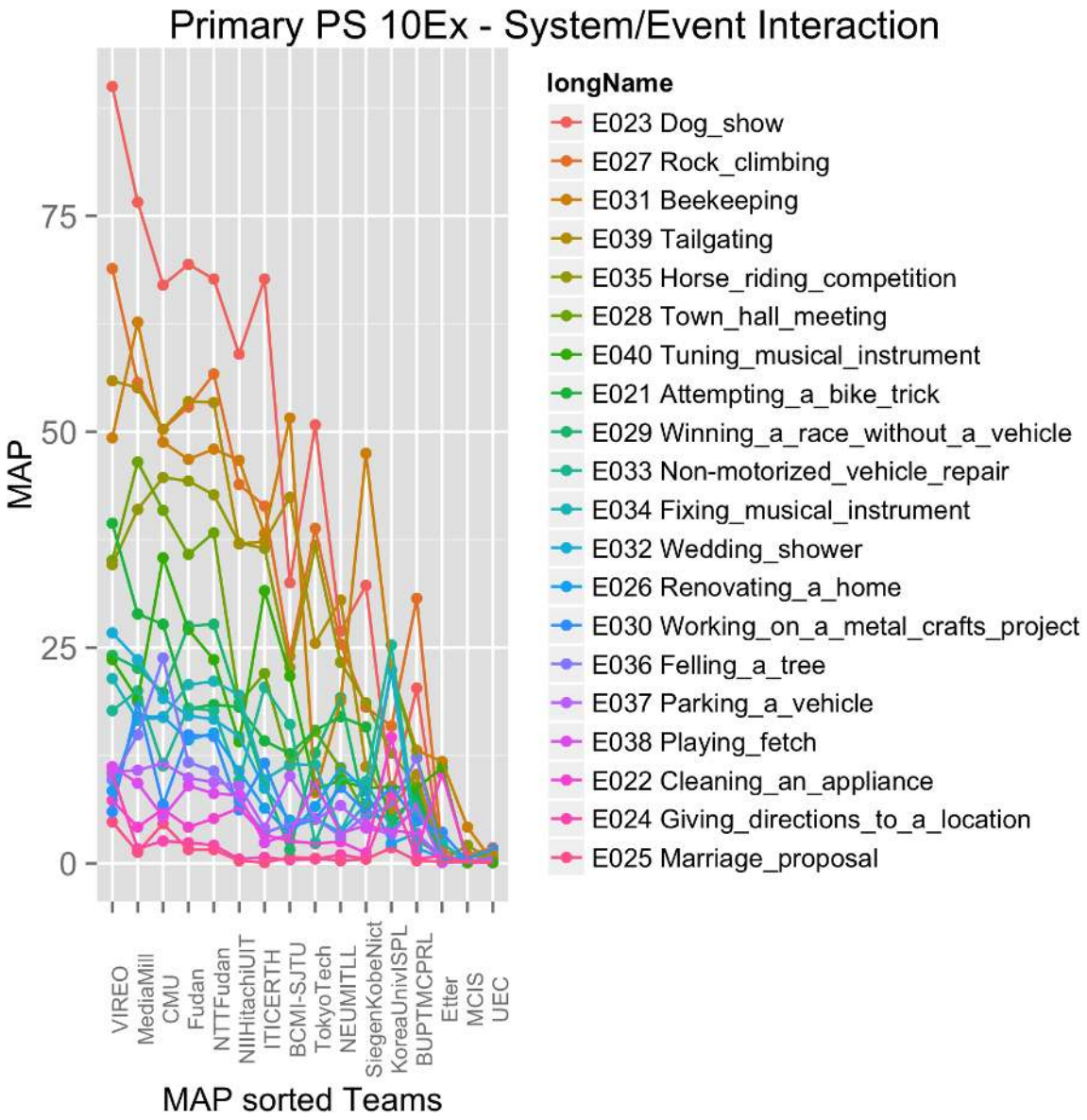
Close-up

Non-field

Classification: events

- TREC Video Retrieval Evaluation (TRECVID) is a yearly collection of competitions for video analysis and retrieval
- Events from 2015 are shown at right
- MAP = mean average precision, higher is better
- Events are specified by giving 0-100 example videos
- Other events: traffic patterns, sports, crime

Classification: events



Video annotation: object recognition

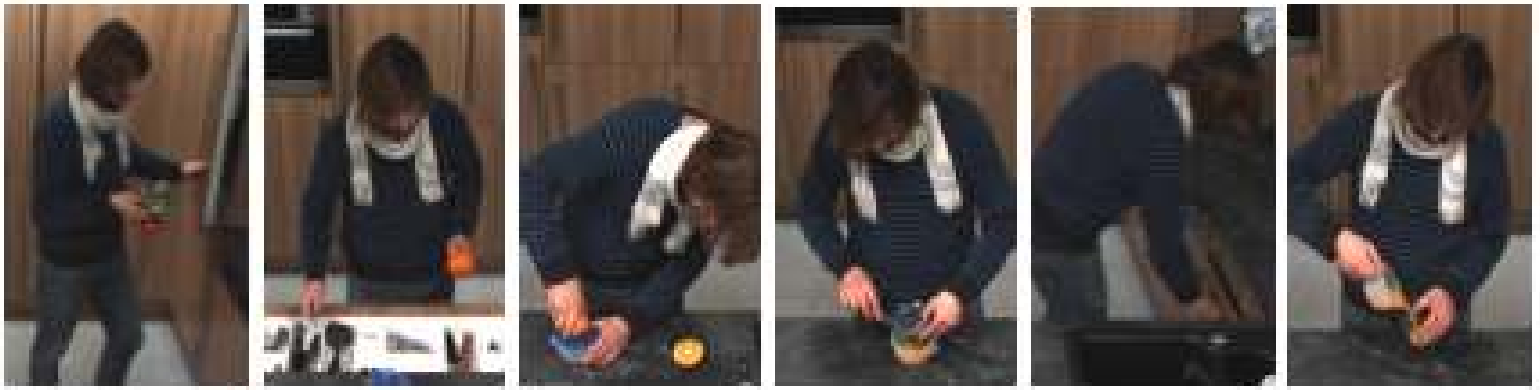
- Identify where, when, and what an object is
- Can use image-based classifier on video frames
  - With temporal continuity to reduce false positives
- Additional context from
  - Relationships between labels
  - Scene that objects are in
- ImageNet 2015 object detection from video task

Video annotation: captioning

- Can caption on several levels of detail
- Corpus of Textually Annotated Cooking Scenes (TACoS) multi-level version
- Example predictions from Donahue et al (CVPR 2015)

Detailed: A man took a cutting board and knife from the drawer. He took out an orange from the refrigerator. Then, he took a knife from the drawer. He juiced one half of the orange. Next, he opened the refrigerator. He cut the orange with the knife. The man threw away the skin. He got a glass from the cabinet. Then, he poured the juice into the glass. Finally, he placed the orange in the sink.
Short: A man juiced the orange. Next, he cut the orange in half. Finally, he poured the juice into a glass.
One sentence: A man juiced the orange.

Video annotation: captioning



Detailed: A man took a cutting board and knife from the drawer. He took out an orange from the refrigerator. Then, he took a knife from the drawer. He juiced one half of the orange. Next, he opened the refrigerator. He cut the orange with the knife. The man threw away the skin. He got a glass from the cabinet. Then, he poured the juice into the glass. Finally, he placed the orange in the sink.
Short: A man juiced the orange. Next, he cut the orange in half. Finally, he poured the juice into a glass.
One sentence: A man juiced the orange.

Video query and retrieval

- Many possible query formulations
- Different similarity measures work best for different kinds of queries
- Relevance feedback allows the user and system to define the task together

Query and retrieval: query types

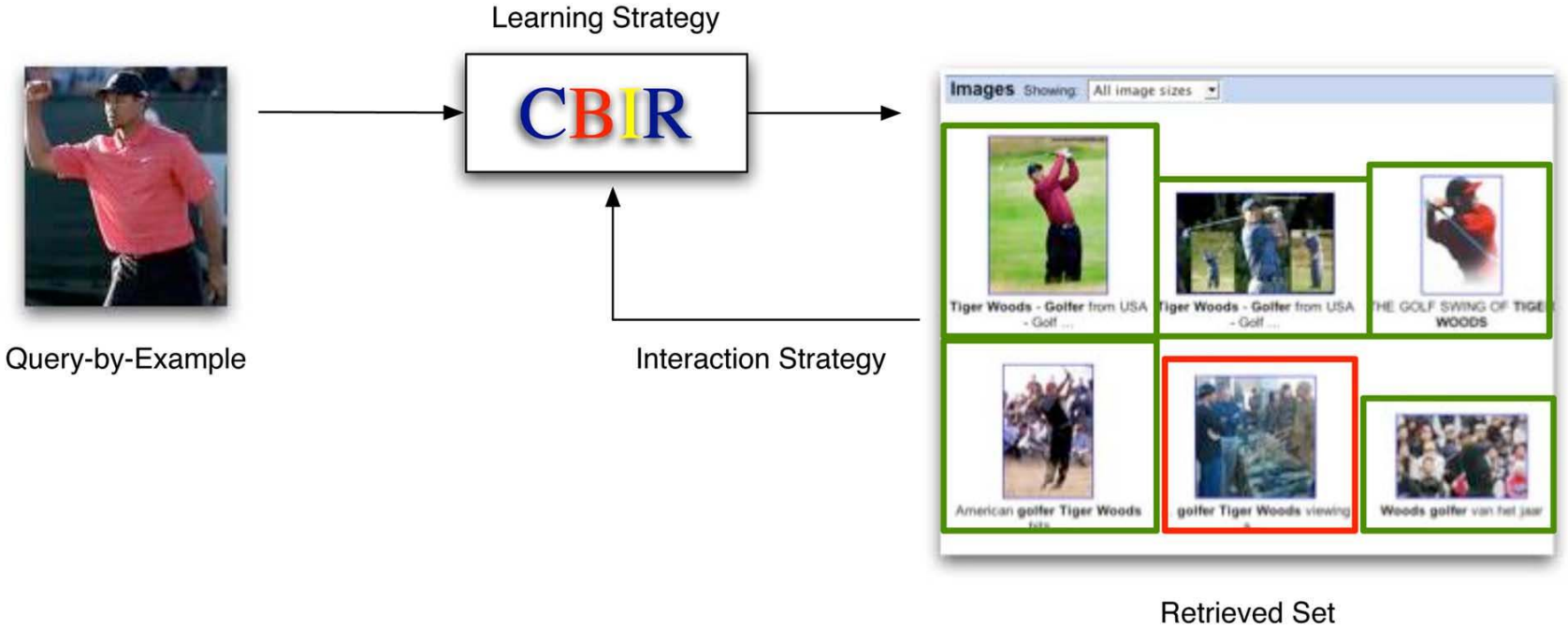
- Query by example
- Query by sketch
- Query by object(s)
- Query by keywords
- Query by natural language

Query and retrieval: Similarity measures

- What does the user mean by “similar”?
- Similar color, texture, motion
- Similar description
- Some combination
- Often the user doesn’t know exactly what they want ahead of time

Query and retrieval: Relevance feedback (Active learning)

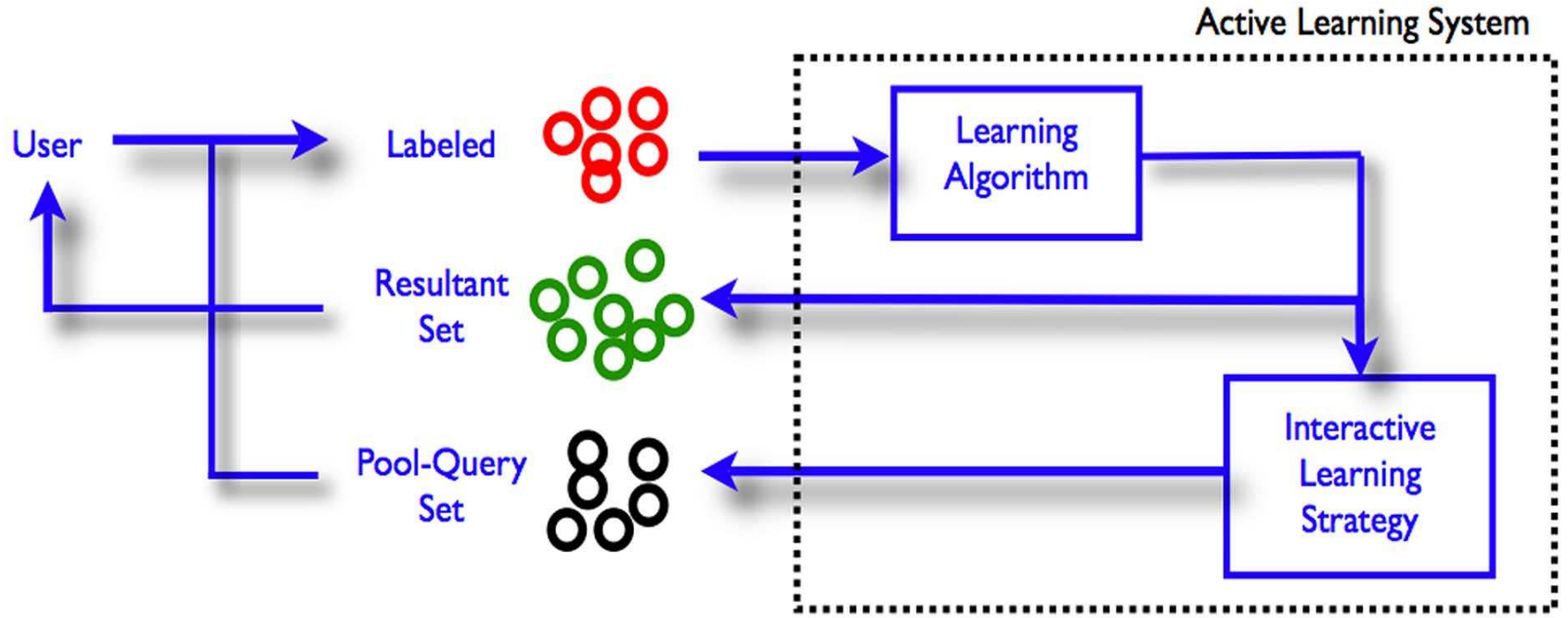
- System and user work together to define query and similarity metric on the fly
- Good when user knows what they want, but have difficulty formulating a specific query



Huang et al (2008), "Active learning for interactive multimedia retrieval" Proceedings of the IEEE.

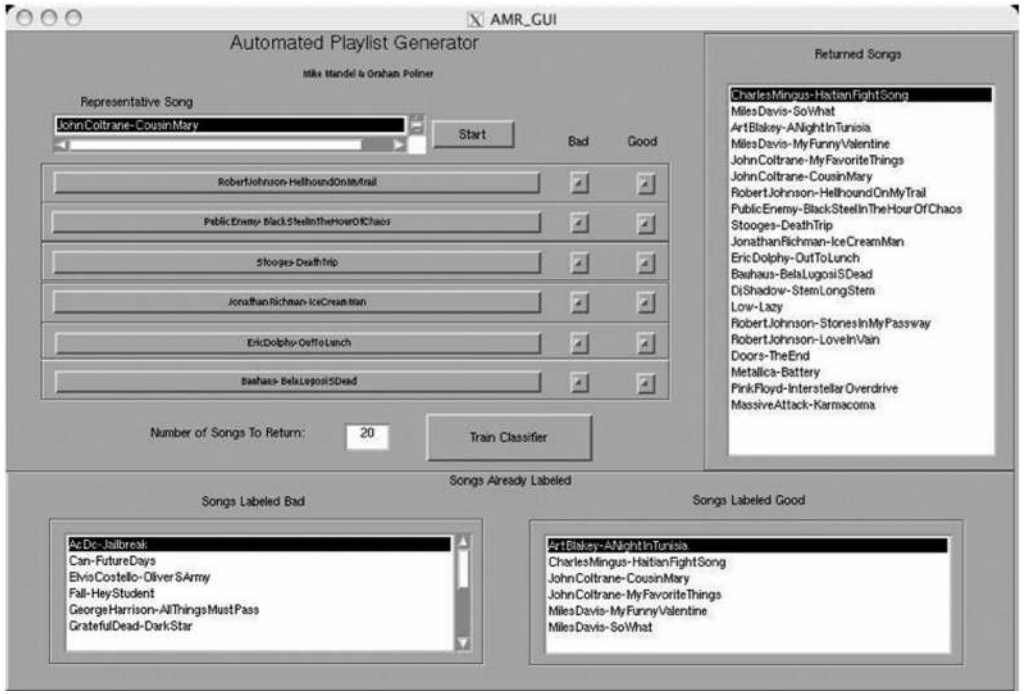
Query and retrieval: Relevance feedback (Active learning)

- System and user work together to define query and similarity metric on the fly
- Good when user knows what they want, but have difficulty formulating a specific query



Huang et al (2008), "Active learning for interactive multimedia retrieval" Proceedings of the IEEE.

Relevance feedback also applicable to other media like music

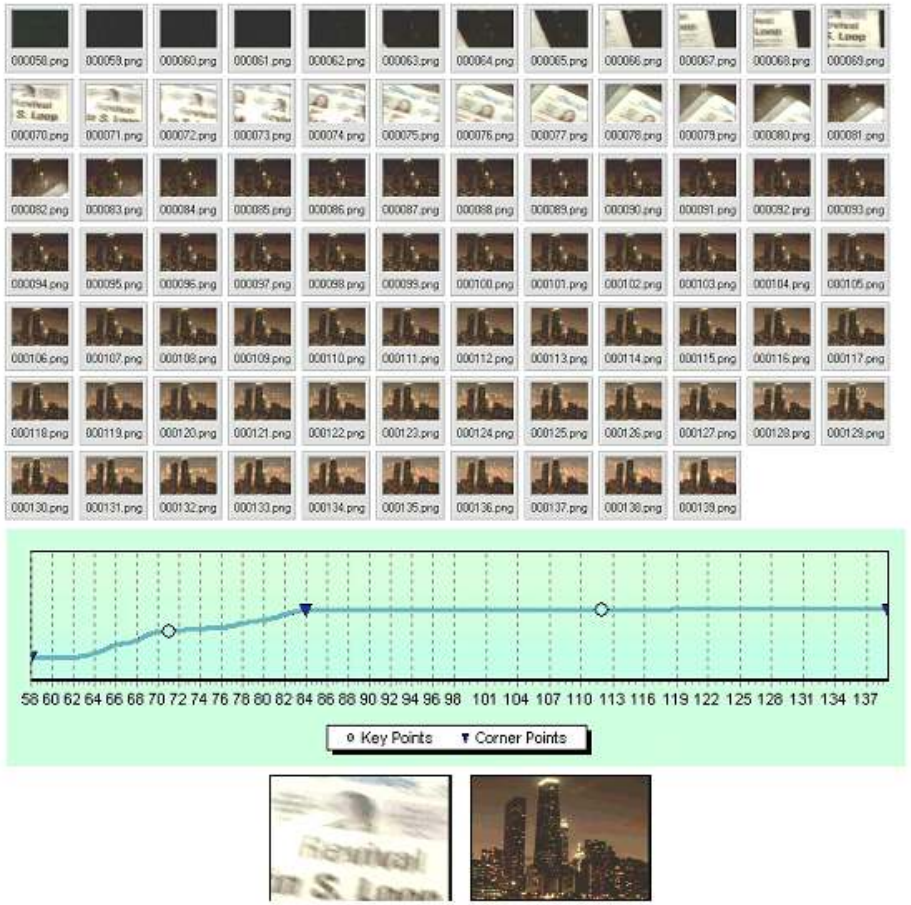


Huang et al (2008), "Active learning for interactive multimedia retrieval" Proceedings of the IEEE.

Video summarization and browsing

- Videos are long, want a summary of each one for exploring a large collection or result set
- Static summaries are (collections of) still images that summarize a video
- Dynamic skims are collections of excerpts from a longer video concatenated together

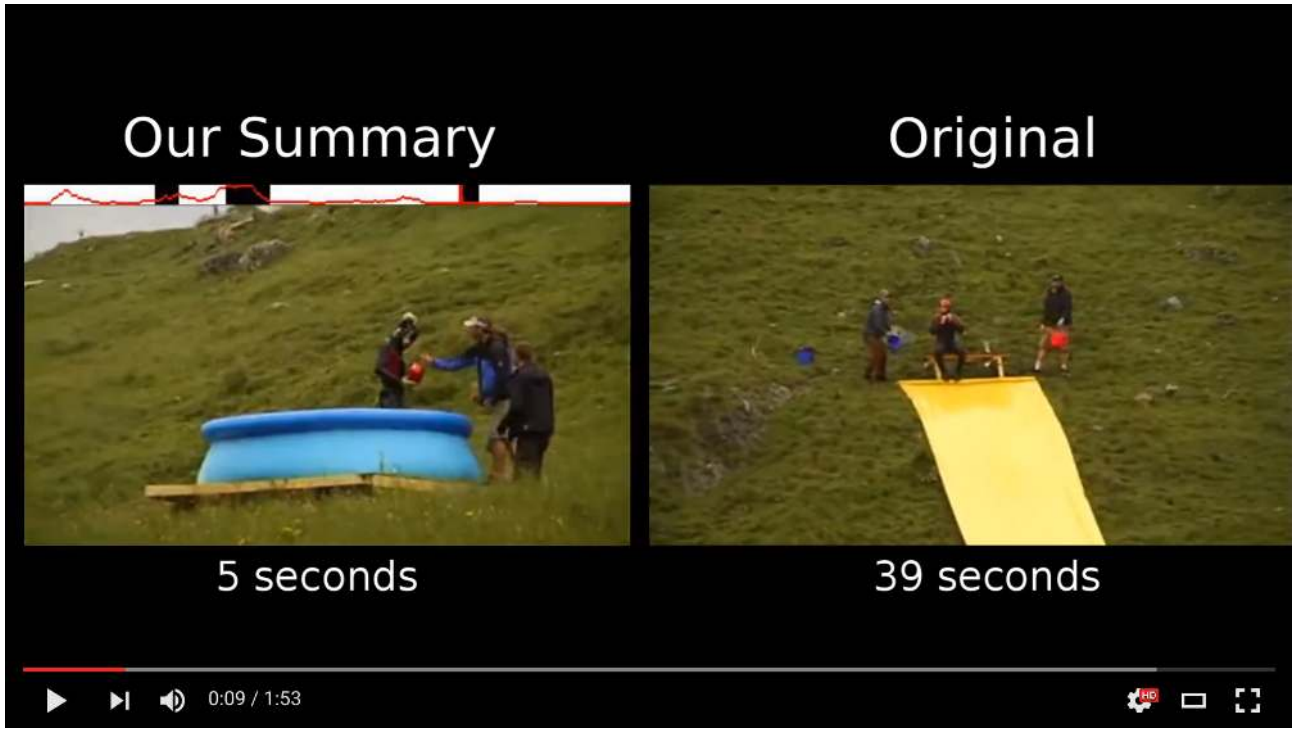
Summarization and browsing: Key frame-based static abstracts



- Select key frames to display as a "storyboard"
- For example, Ciocca and Schettini (JRTIP, 2006)
- Pros: fast, compact, non-linear, complete
- Cons: no audio, no motion, hard to understand for complex videos

Summarization and browsing: Skimming

- Stitch together important sections of a video
- Pros: preserves time, action, audio
- Cons: still linear, still takes time to watch



Gygli, M., Grabner, H., Riemenschneider, H., & Van Gool, L. (2014). Creating summaries from user videos. In ECCV

Summary

- Video retrieval is like image retrieval, but with temporal coherence, context, and motion
- Video segmentation identifies more homogeneous sequences of frames to further analyze
- Video features capture image characteristics and motion
- Unsupervised data mining can find patterns
- Supervised classification can identify specific objects, actions, genres, and shots