# Video Summarization

Ben Wing

CS 395T, Spring 2008

April 11, 2008

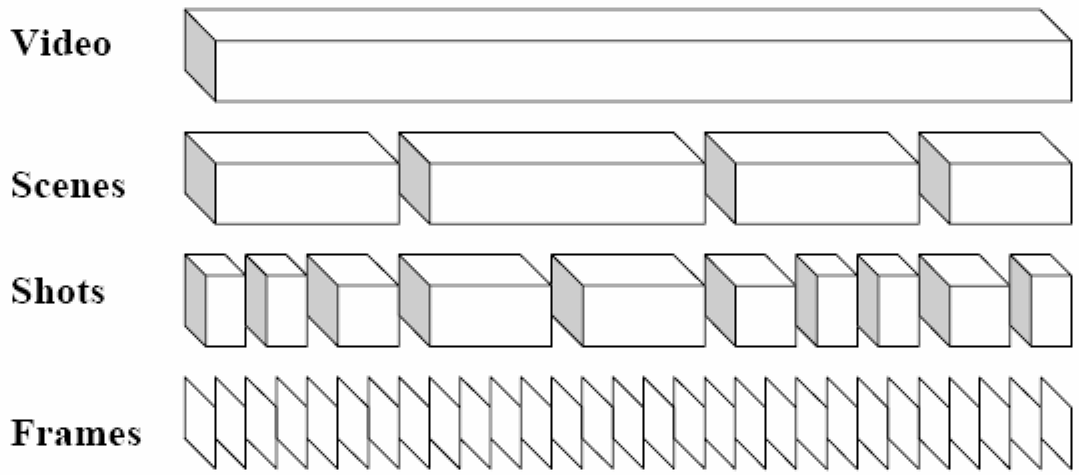
# Overview

- *"Video summarization methods attempt to abstract the main occurrences, scenes, or objects in a clip in order to provide an easily interpreted synopsis"*
  - Video is time-consuming to watch
  - Much low-quality video
  - Huge increase in video generation in recent years

# Overview

- Specific situations:
  - Previews of movies, TV episodes, etc.
  - Summaries of documentaries, home videos, etc.
  - Highlights of football games, etc.
  - Interesting events in surveillance videos (major commercial application)

# Anatomy of a Video




- **frame**: a single still image from a video
    - 24 to 30 frames/second
- **shot**: sequence of frames recorded in a single camera operation
- **scene**: collection of shots forming a semantic unity
    - conceptually, a single time and place

# Outline

- **Series of still images (*key frames*)**
  - Shot boundary based
  - Perceptual feature based
    - color-based (Zhang 1997)
    - motion-based (Wolf 1996; Zhang 1997)
    - object-based (Kim and Huang 2001)
  - Feature vector space based (DeMenthon et al. 1998; Zhao et al. 2000)
  - Scene-change detection (Ngo et al. 2001)
- Montage of still images
  - Synopsis mosaics (Aner and Kender 2002; Irani et al. 1996)
  - Dynamic stills (Caspi et al. 2006)
- Collection of short clips (*video skimming*)
    - Highlight sequence
      - Movie previews: VAbstract (Pfeiffer et al. 1996)
      - Model-based summarization (Li and Sezan 2002)
    - Summary sequence: full content of video
      - Time-compression based ("fast forward")
      - Adaptive fast forward (Petrovic, Jojic and Huang 2005)
      - Text- and speech-recognition based
- Montage of moving images
  - Webcam synopsis (Pritch et al. 2007)

# Shot Boundary-Based Key Frame Selection

- segment video into shots
  - typically, difference of one or more features greater than threshold
    - pixels (Ardizzone and Cascia, 1997; …)
    - color/grayscale histograms (Abdel-Modttaleb and Dimitrova, 1996; …)
    - edge changes (Zabih, Miller and Mai, 1995)
- select key frame(s) for each shot
  - first, middle, last frame (Hammoud and Mohr, 2000)
  - look for significant change within shot (Dufaux, 2000)

# Color-Based Selection (Zhang 1997)

- quantize color space into N cells (e.g. 64)
- compute histogram: number of pixels in each cell
- compute distance between histograms

$$D_{his}(I,Q)=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}a_{ij}(I_i-Q_i)(I_j-Q_j)$$

- $a_{ij}$ is perceptual similarity between color bins

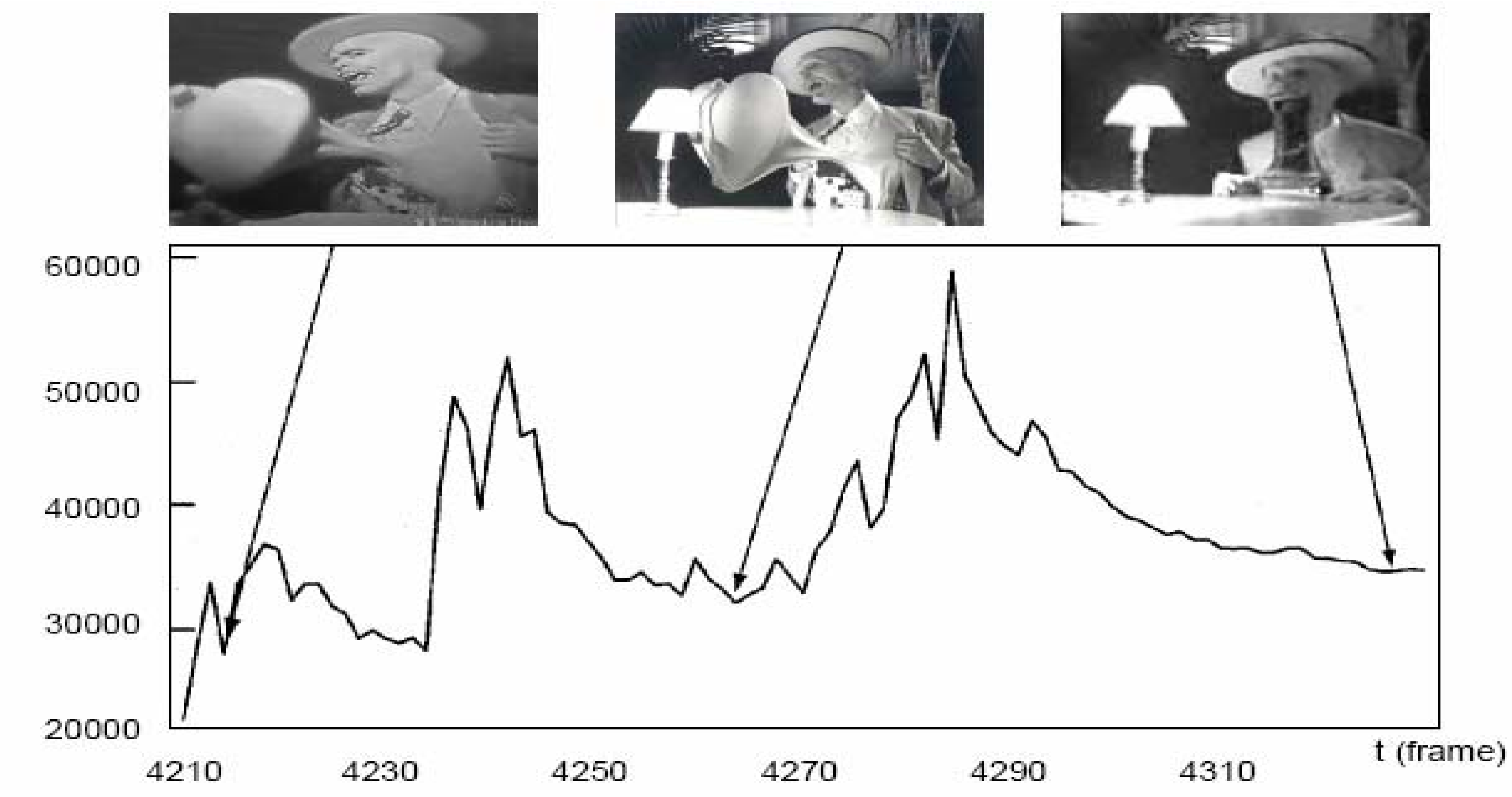
# Motion-Based Selection (Wolf 1996; Zhang 1997)

- color-based selection may not be enough given significant motion
- motion metric based on optical flow

$$M(t) = \sum_{i=1}^{r} \sum_{j=1}^{c} |o_x(i,j,t)| + |o_y(i,j,t)|$$

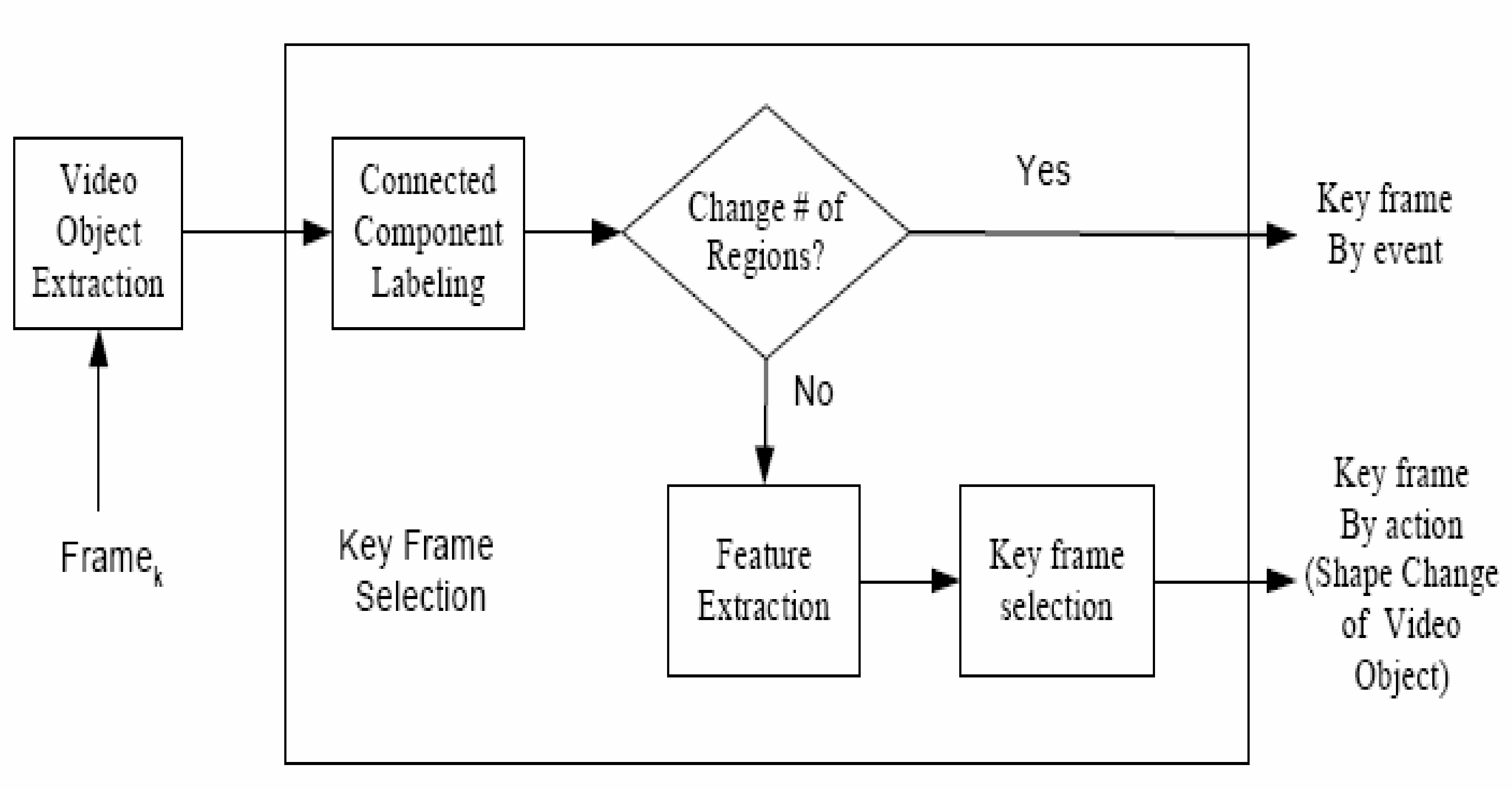
- $o_x(i,j,t)$, $o_y(i,j,t)$ are $x/y$ components of optical flow of pixel $(i,j)$, frame $t$
- identify two local maxima $m_1$ and $m_2$ where difference exceeds threshold
- select minimum point between $m_1$ and $m_2$ as key frame
- repeat for maxima $m_2$ and $m_3$, etc.

# Motion-Based Selection (Wolf 1996; Zhang 1997)



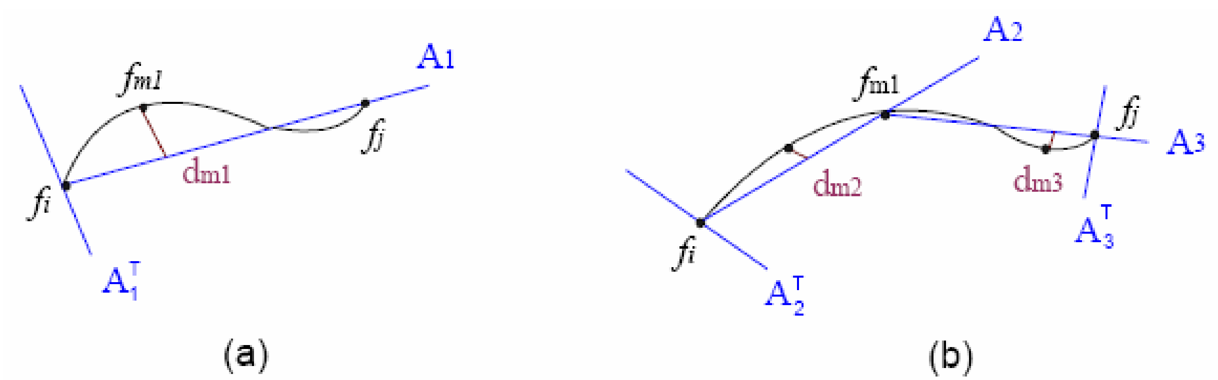

Values of M(t) and sample key frames from *The Mask*

# Object-based Selection (Kim and Huang, 2001)

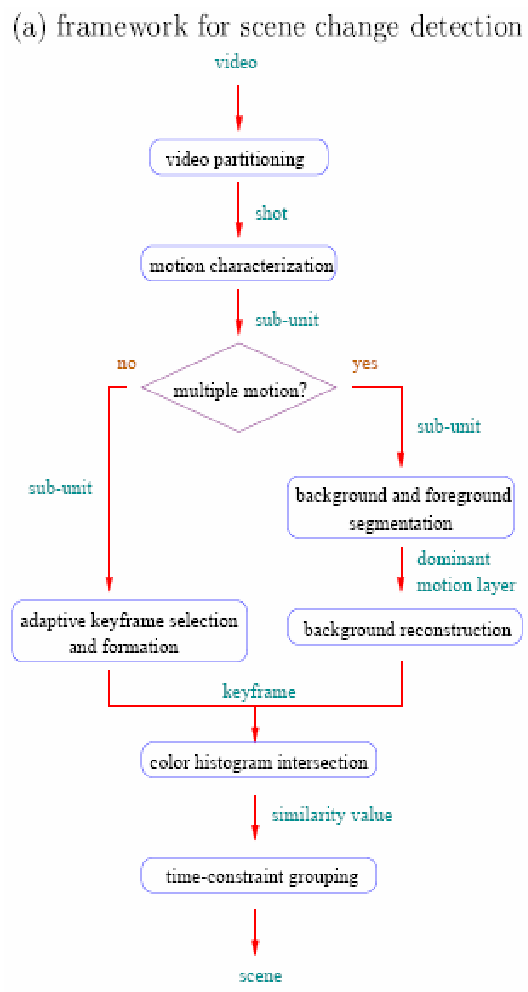


# Feature Vector Space-Based Key Frame Detection

- DeMenthon, Kobla and Doermann (1998)
- Zhao, Qi, Li, Yang and Zhang (2000)
  - Represent frame as point in multi-dimensional feature space
  - Entire clip is curve in same space
  - Select key frames based on curve properties (sharp corners, direction change, etc.)
  - Curve-splitting algorithm can successively add new frames

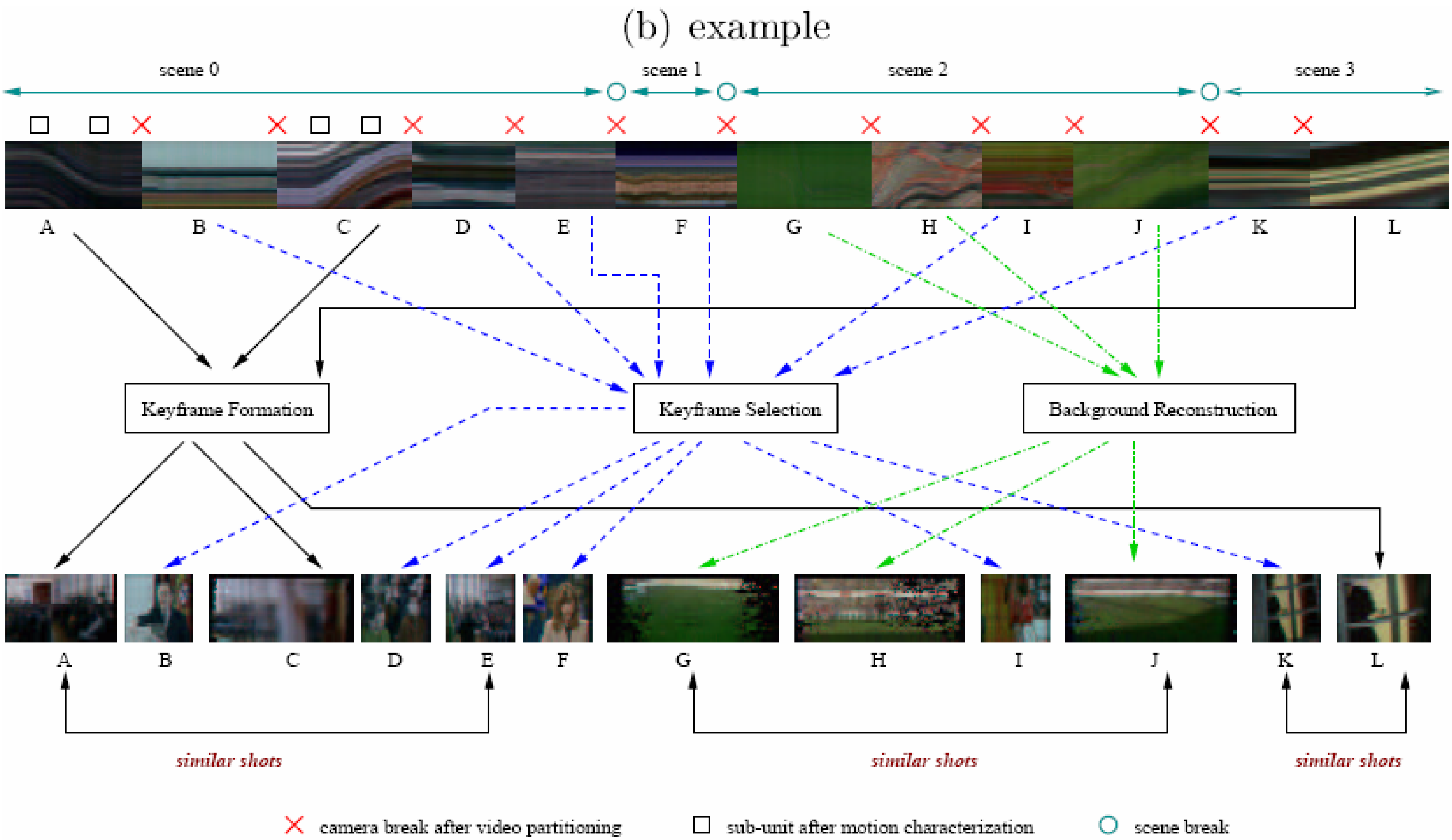


# Scene-Change Detection




- Ngo, Zhang and Pong (2001)

# Scene-Change Detection



# Outline

- Series of still images (*key frames*)
  - Shot boundary based
  - Perceptual feature based
    - color-based (Zhang 1997)
    - motion-based (Wolf 1996; Zhang 1997)
    - object-based (Kim and Huang 2001)
  - Feature vector space based (DeMenthon et al. 1998; Zhao et al. 2000)
  - Scene-change detection (Ngo et al. 2001)
- **Montage of still images**
  - Synopsis mosaics (Aner and Kender 2002; Irani et al. 1996)
  - Dynamic stills (Caspi et al. 2006)
- Collection of short clips (*video skimming*)
    - Highlight sequence
      - Movie previews: VAbstract (Pfeiffer et al. 1996)
      - Model-based summarization (Li and Sezan 2002)
    - Summary sequence: full content of video
      - Time-compression based ("fast forward")
      - Adaptive fast forward (Petrovic, Jojic and Huang 2005)
      - Text- and speech-recognition based
- Montage of moving images
  - Webcam synopsis (Pritch et al. 2007)

# Synopsis Mosaics

- Aner and Kender (2002)
- Irani et al. (1996)

(a)

(b)

**Fig. 1.** (a) Hand-chosen key-frames. (Automatic key-frames generation often does not give complete spatial information). (b) Mosaic representation. Note that the whole background is visible, no occlusion by the foreground objects.

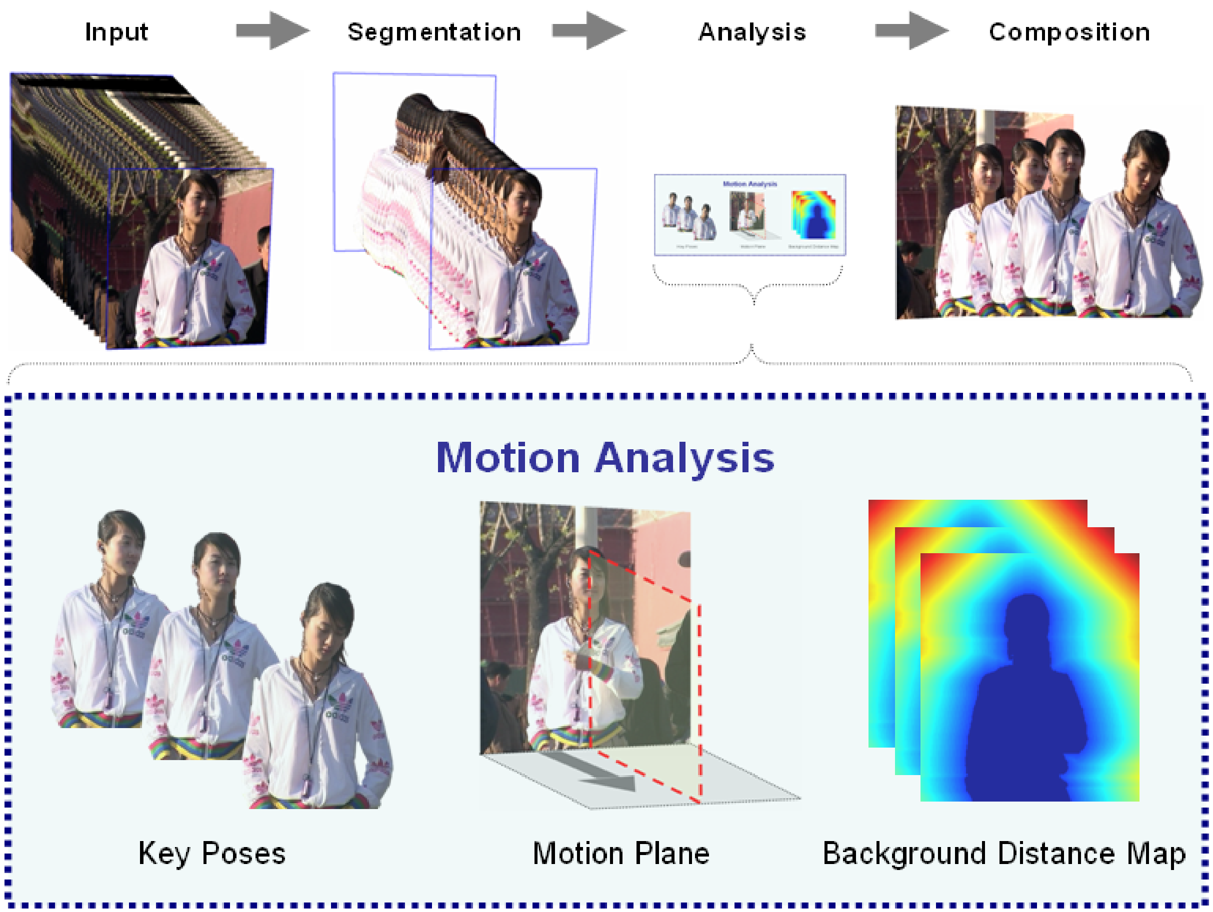
# Synopsis Mosaics

- Select or sample key frames
- Compute affine transformations between successive frames
- Choose one frame as reference frame
- Project other frames into plane of reference coordinate system
- Use median of all pixels mapped to same location
- Optionally, use outlier detection to remove moving objects

# Synopsis Mosaics

- Advantages
  - Combine key frames into single shot
  - Can recreate full background when occluded by moving objects
- Disadvantages
  - May require manual key-frame selection to get complete background
  - Moving objects may not display well – need to segment out and recombine through other means

# Dynamic Stills (Caspi et al. 2006)



# Dynamic Stills (Caspi et al. 2006)

(a) Key frames (b) Synopsis mosaic (c) Our representation

# Dynamic Stills (Caspi et al. 2006)

- Advantages
  - Better sense of motion than key frames
  - Better screen usage
  - Can handle self-occluding sequences (vs. synopsis mosaics)
- Disadvantages
  - Single image is limited in complexity (max number of poses representable is about 12)
  - Rotation of multiple objects may lead to occlusion
  - Exact spatial information is lost (cf. running in place)

# Outline

- Series of still images (*key frames*)
  - Shot boundary based
  - Perceptual feature based
    - color-based (Zhang 1997)
    - motion-based (Wolf 1996; Zhang 1997)
    - object-based (Kim and Huang 2001)
  - Feature vector space based (DeMenthon et al. 1998; Zhao et al. 2000)
  - Scene-change detection (Ngo et al. 2001)
- Montage of still images
  - Synopsis mosaics (Aner and Kender 2002; Irani et al. 1996)
  - Dynamic stills (Caspi et al. 2006)
- **Collection of short clips (*video skimming*)**
    - Highlight sequence
      - Movie previews: VAbstract (Pfeiffer et al. 1996)
      - Model-based summarization (Li and Sezan 2002)
    - Summary sequence: full content of video
      - Time-compression based ("fast forward")
      - Adaptive fast forward (Petrovic, Jojic and Huang 2005)
      - Text- and speech-recognition based
- Montage of moving images
  - Webcam synopsis (Pritch et al. 2007)

# VAbstract (Pfeiffer et al 1996)

1. Important objects/people
   - Scene-boundary detection (Kang 2001; Sundaram and Chang 2002; etc.)
   - Find high-contrast scenes
2. Action
   - Find high-motion scenes
3. Mood
   - Find scenes of average color composition
4. Dialog
   - Find scenes with dialog
5. Disguised ending
   - Delete final scenes

# Model-Based Summarization: Li and Sezan (2002)

- Summarization of football broadcasts
- Model video as sequence of plays
  - Remove non-play footage
  - Select most important/exciting plays
    - Use waveform of audio
- Start-of-play detection:
  - Field color, field lines
  - Camera motions
  - Team jersey colors
  - Player line-ups
- End-of-play detection:
  - Camera breaks after start of play
- Also applied to baseball and sumo wrestling

# Summary Sequence

- Time-compression based ("fast forward")
  - Drop some fixed proportion of frames
  - Extreme case: time-lapse photography
- Adaptive fast forward
  - Petrovic, Jojic and Huang (2005)
  - Create graphical model of video scenes (occlusion, appearance change, motion)
  - Maximize likelihood of similarity to target video
- Text- and speech-recognition based
  - Use dialog (from speech recognition, closed captions, subtitles) to guide scene selection

# Outline

- Series of still images (*key frames*)
  - Shot boundary based
  - Perceptual feature based
    - color-based (Zhang 1997)
    - motion-based (Wolf 1996; Zhang 1997)
    - object-based (Kim and Huang 2001)
  - Feature vector space based (DeMenthon et al. 1998; Zhao et al. 2000)
  - Scene-change detection (Ngo et al. 2001)
- Montage of still images
  - Synopsis mosaics (Aner and Kender 2002; Irani et al. 1996)
  - Dynamic stills (Caspi et al. 2006)
- Collection of short clips (*video skimming*)
    - Highlight sequence
      - Movie previews: VAbstract (Pfeiffer et al. 1996)
      - Model-based summarization (Li and Sezan 2002)
    - Summary sequence: full content of video
      - Time-compression based ("fast forward")
      - Adaptive fast forward (Petrovic, Jojic and Huang 2005)
      - Text- and speech-recognition based
- **Montage of moving images**
  - Webcam synopsis (Pritch et al. 2007)

# Webcam Synopsis
# (Pritch, Rav-Acha, Gutman, Peleg 2007)

- Webcams and security cameras collect endless footage, most of which is thrown away without being viewed
- > 1,000,000 security cameras in London alone!
- Idea: *"Show me in one minute the synopsis of this camera broadcast during the past day"*
  - Issue: Security companies want to select by importance of event rather than by a fixed time

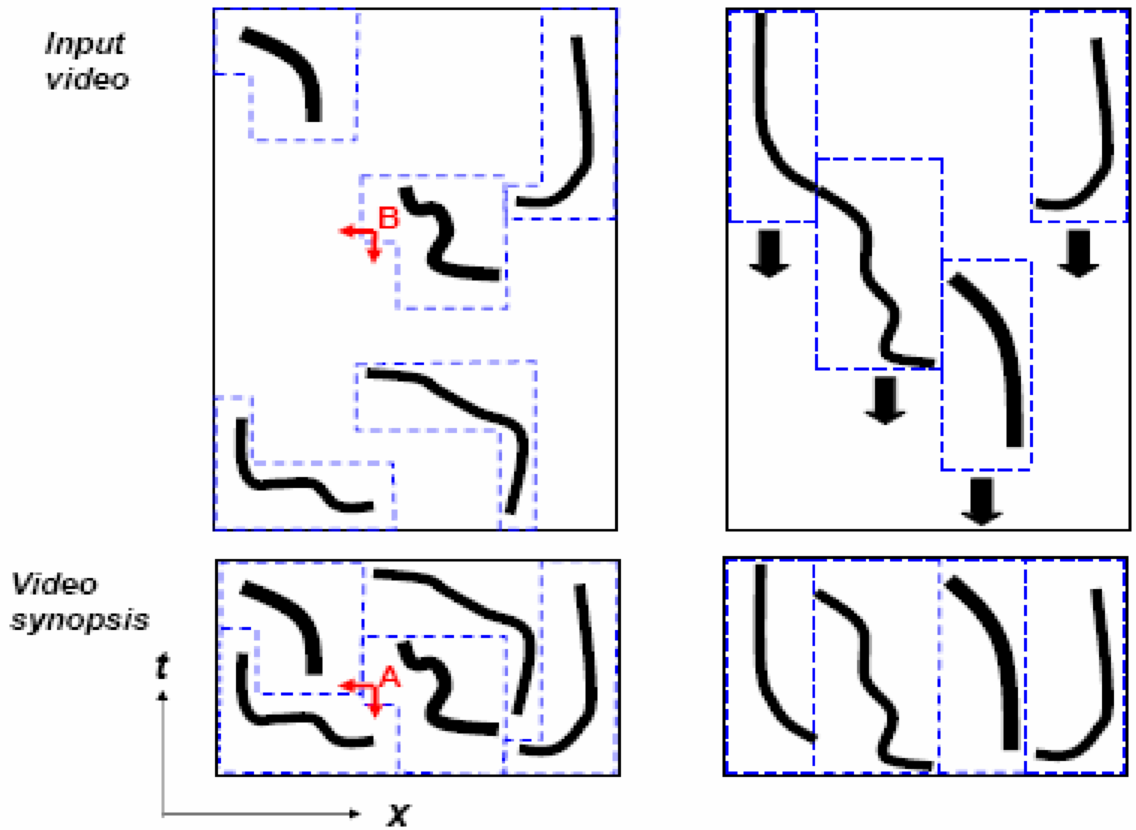
# Webcam Synopsis (Pritch, Rav-Acha, Gutman, Peleg 2007)

Example synopsis (from website):

- Note stroboscopic effect (duplicated instances of same person)

# Webcam Synopsis
# (Pritch, Rav-Acha, Gutman, Peleg 2007)

- Identify *tubes* of activity
- Find a lowest-cost synopsis:
  1. Maximize activity (pack as close as possible)
  2. Minimize overlap (“collision”)
  3. Maximize temporal consistency
- Pack tubes according to identified synopsis
- Place over a time-lapse background



(a)
(b)

# Webcam Synopsis: Object Detection and Segmentation

- For each frame, compute median background image over surrounding four-minute stretch
- Find moving objects using background subtraction + min-cut (for smoothness)
- Find connected components to get the object tubes
- More sophisticated object-detection algorithms are possible

# Webcam Synopsis:
# Object Detection and Segmentation

Examples of four computed tubes from an airport surveillance camera

# Webcam Synopsis: Finding Best Synopsis

- We seek to find the best synopsis, optimizing the *activity*, *background consistency*, *collision*, and *temporal consistency* costs.
- A synopsis is a mapping, for each tube $b$, from its original time extent $[t_s, t_e]$ to a shifted extent $[\hat{t}_s, \hat{t}_e]$. The tube in its shifted extent is notated as $\hat{b}$.
- The energy cost of a synopsis is defined as

$$\begin{aligned} E(M) &= \sum_{b\in B}(E_a(\hat{b}) + \gamma E_s(\hat{b})) + \\ &+ \sum_{b,b'\in B}(\alpha E_t(\hat{b},\hat{b}') + \beta E_c(\hat{b},\hat{b}')) \end{aligned}$$

- Where
  - $E_a$ is the activity cost of a tube
  - $E_s$ is the background consistency of a tube
  - $E_c$ is the collision cost between two tubes
  - $E_t$ is the temporal consistency cost between two tubes.

# Webcam Synopsis: Finding Best Synopsis (1)

The activity cost is 0 for tubes in the synopsis. For tubes not included, it is the sum over the "activity" of each pixel (difference from background).

$$E_a(\hat{b}) = \sum_{x,y,t} \chi_{\hat{b}}(x,y,t)$$

$$\chi_b(x,y,t) = \begin{cases} ||I(x,y,t) - B(x,y,t)|| & t \in t_b \\ 0 & otherwise \end{cases}$$

The background consistency cost is defined as the sum over the per-pixel difference between mapped tube and time-lapsed background.

$$E_s(\hat{b}) = \sum_{x,y \in \sigma(\hat{b}), t \in \hat{t_b} \cap t_{out}} \|I_{\hat{b}}(x,y,t) - B_{out}(x,y,t)\|$$

# Webcam Synopsis: Finding Best Synopsis (2): Collision Cost

- The collision cost is defined over pairs of tubes.
- It sums over each pixel in each frame where the tubes overlap.
- For such pixels, the cost is the product of their "activities" (differences from background).

$$E_c(\hat{b}, \hat{b'}) = \sum_{x,y,t \in \hat{t_b} \cap \hat{t_{b'}}} \chi_{\hat{b}}(x,y,t) \chi_{\hat{b'}}(x,y,t)$$

$$\chi_b(x,y,t) = \begin{cases} ||I(x,y,t) - B(x,y,t)|| & t \in t_b \\ 0 & otherwise \end{cases}$$

# Webcam Synopsis: Finding Best Synopsis (3): Temporal Consistency Cost

- The temporal consistency cost tries to ensure that each pair of tubes is temporally consistent in their mapped time stretches.
- We'd like to weight the cost per pair of tubes by the *interaction strength* between tubes. But it's too hard (impossible?) to compute, so approximate as how close the tubes ever got:

$$\text{if} \quad \hat{t_b} \cap \hat{t_{b'}} \neq \emptyset \quad \text{then}$$
$$d(b, b') = \exp(-\min_{t \in \hat{t_b} \cap \hat{t_{b'}}} \{d(b, b', t)\} / \sigma_{space})$$

- where d(b,b',t) = Euclidean distance between closest pixels in b and b' in *mapped* frame t.
- If, however, b and b' have no frames in common (one is *mapped* completely before the other, assume b), then weight is how close the tubes ever got in time space:

$$d(b, b') = \exp(-(\hat{t^s_{b'}} - \hat{t^e_b}) / \sigma_{time})$$

# Webcam Synopsis: Finding Best Synopsis (3): Temporal Consistency Cost

- Remember, d(b,b'):
  - Measures closeness between tubes at their closest point in time or space
  - Value drops off exponentially, so *only very "bad" tubes matter* (nearly touching when time overlaps, nearly time-overlapping otherwise)
- Finally, define temporal consistency cost: 0 if exact same relative timing applies between original and mapped pair of tubes; otherwise, constant-scaled version of d(b,b')
- Intuition: Keep tubes from getting too close in time or space

$$E_t(\hat{b}, \hat{b'}) = d(b, b') \cdot \begin{cases} 0 & t^s_{b'} - t^s_b = t^{\hat{s}}_{b'} - t^{\hat{s}}_b \\ C & \text{otherwise} \end{cases}$$

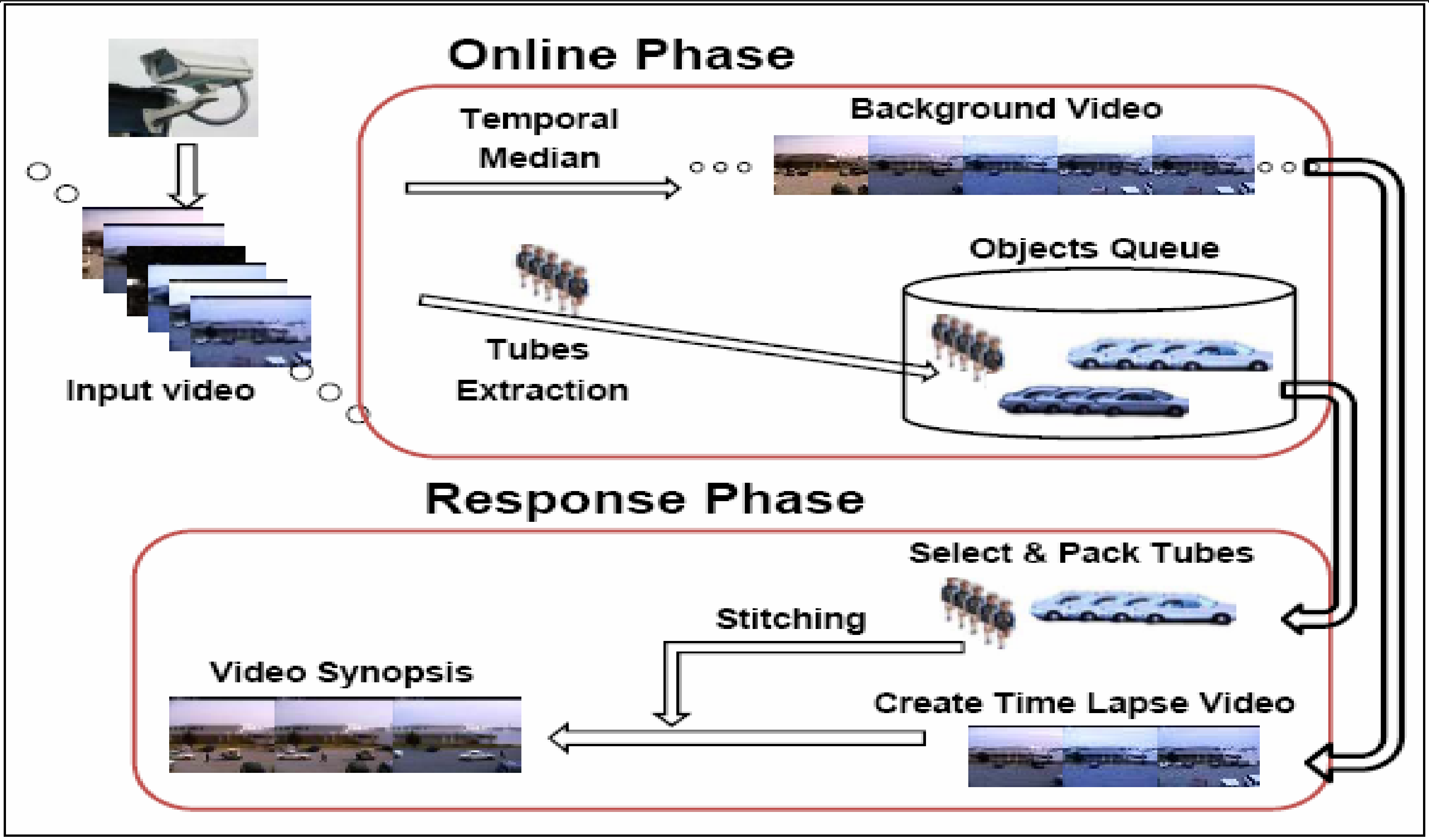
# Webcam Synopsis: Finding Best Synopsis (4)

- How do you optimize?

$$E(M) = \sum_{b \in B} (E_a(\hat{b}) + \gamma E_s(\hat{b})) + \\ + \sum_{b,b' \in B} (\alpha E_t(\hat{b}, \hat{b}') + \beta E_c(\hat{b}, \hat{b}'))$$

- The form of E(M) makes it amenable to MRF’s (Markov Random Fields), a generalization of HMM’s (Hidden Markov Models).
- But the authors just used a simple greedy optimization (with simulated annealing?) and got good results.

# Webcam Synopsis: Handling Endless Video




**Online phase:** computed in parallel with original streaming
**Response phase:** computed afterwards, in response to a user request

# Webcam Synopsis: Issues

- Advantages
  - Efficient compression of very lengthy surveillance videos
  - User-controllable compression threshold
  - Scheme for handling endless video
  - User can select for specific types of objects (cars vs. people) or motion (motion through frame or background/foreground transition)
- Disadvantages
  - Non-optimal user controls for compression
    - Security companies want an event importance threshold, not a time threshold
  - Limited applicability: Cannot handle videos with unpredictable background shift
  - May be compute-intensive

# Webcam Synopsis: Other Thoughts

- Combining speech/audio/dialog/voice
  - Use various techniques (cf. “Buffy”, Everingham, Sivic and Zisserman; 2006) to link audio/dialog with video
    - create combined audio/video tubes
    - Augment energy function with audio overlap term: audio information at same frequencies, and dialog in general, should not overlap
    - Generate mixed audio channel along with video
- Privacy concerns!  **Huge** can of worms.

# References


- Abdel-Mottaleb, M., & Dimitrova, N. (1996). CONIVAS: CONtent-based image and video access system. *Proceedings of ACM International Conference on Multimedia*, Boston, MA, 427-428.
- Aner, A. and J. Kender (2002). Video Summaries through Mosaic-Based Shot and Scene Clustering. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2002.
- Ardizzone, E., & Cascia, M. (1997). Automatic video database indexing and retrieval. *Multimedia Tools and Applications*, 4, 29-56.
- Everingham,M., J. Sivic and A. Zisserman (2006). “Hello! My name is... Buffy” – Automatic Naming of Characters in TV Video. *British Machine Vision Conference* (BMVC), 2006.
- DeMenthon, D., Kobla, V., & Doermann, D. (1998). Video summarization by curve simplification. *Proceedings of ACM Multimedia 1998*, 211-218.
- Dufaux, F. (2000). Key frame selection to represent a video. *Proceedings of IEEE 2000 International Conference on Image Processing*, Vancouver, BC, Canada, 275-278.
- Hammoud, R., & Mohr, R. (2000, Aug.). A probabilistic framework of selecting effective key frames from video browsing and indexing. *Proceedings of International Workshop on Real-Time Image Sequence Analysis*, Oulu, Finland, 79-88.
- Irani, M., P. Anandan, J. Bergenand R. Kumar, and S. Hsu (1996). Efficient representation of video sequences and their applications. In *Signal processing: Image Communication*, volume 8, 1996.
- Kang, H. (2001). A hierarchical approach to scene segmentation. *IEEE Workshop on Content-Based Access of Image and Video Libraries* (CBAIVL 2001), 65-71.
- Kim, C., & Hwang, J. (2001). An integrated scheme for object-based video abstraction. *Proceedings of ACM Multimedia 2001*, Los Angeles, CA, 303-309.
- Li, B., & Sezan, I. (2002). Event detection and summarization in American football broadcast video. *Proceedings of SPIE, Storage ad Retrieval for Media Databases*, 202-213.

# References


- Nagasaka, A., & Tanaka, Y. (1991). Automatic video indexing and full-video search for object appearance. *Proceedings of the IFIP TC2/WG2.6, Second Working Conference on Visual Database Systems*, North-Holland, 113-127.
- Ngo, C., H. Zhang, and T. Pong (2001). Recent Advances in Content-based Video Analysis. *International Journal of Image and Graphics*, 2001.
- Oh, J., Q. Wen, J. lee, and S. Hwang (2004). Video Abstraction. In S. Deb, editor, *Video Data Management and Information Retrieval*, Idea Group Inc. and IRM Press, 2004.
- Petrovic, N., N. Jojic, and T. Huang (2005). Adaptive video fast forward. *Multimedia Tools and Applications*, 26(3):327–344, August 2005.
- Pfeiffer, S., Lienhart, R., Fischer, S., & Effelsberg, W. (1996). Abstracting digital movies automatically. *Journal of Visual Communication and Image Representation*, 7(4), 345-353.
- Pritch, Y., A. Rav-Acha, A. Gutman, and S. Peleg (2007). Webcam Synopsis: Peeking Around the World. In *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), 2007.
- Pritch, Y., A. Rav-Acha, and S. Peleg (2008). Non-Chronological Video Synopsis and Indexing. *IEEE Trans. PAMI*, to appear Nov. 2008. 15p.
- Sundaram, H., & Chang, S. (2000). Video scene segmentation using video and audio Features. *ICME2000*, 1145-1148.
- Wolf, W. (1996). Key frame selection by motion analysis. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA,1228-1231.
- Zabih, R., Miller, J., & Mai, K. (1995). A feature-based algorithm for detecting and classifying scene breaks. *Proceedings of the Third ACM International Conference on Multimedia*, San Francisco, CA, 189-200.
- Zhang, H.J. (1997). An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4), 643-658.
- Zhao, L., Qi, W., Li, S., Yang, S., & Zhang, H. (2000). Key-frame extraction and shot retrieval using nearest feature line (NFL). *Proceedings of ACM Multimedia Workshop* 2000, Los Angeles, CA, 217-220.