Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution

H.J. Zhang
C. Y. Low
S. W. Smoliar
J. H. Wu

Video Parsing: Segmentation and Extraction

- Unstructured content
- If not subject to parsing, can't be indexed or accessed on basis of structural properties
- **Temporal segmentation of video into elemental units**
- **Content extraction from the units**
- Fully-automated content extraction is HARD
- **Build indexes using low-level visual features**
- Divide into low-level image features and semantic features based on objects and events (hard)
- **Retrieval and browsing based on indexes**

Video Parsing: Segmentation and Abstraction

- Basic unit: shot
- One or more frames generated and recorded contiguously
- Represent continuous action in time and space
- Detect not only simple camera breaks
- Also special effects like fades, wipes, dissolves
- Segment compressed video directly--no need to decompress
- Use DCT coefficients, motion vectors for content comparison
- Motion features: panning, zooming, tilting: false +

Key-frame extraction

- Map entire segment to some small number of representative images (key frames)
  - Index constructed from key frames
  - Retrieval queries directed at key frames
  - Display key frames for browsing
- Still images that best represent video content
  - Extract or reconstruct from data
- Needs to be automatic and content-based
  - Maintain important content of video
  - Semantic analysis of content too hard
- Use image features
- Heuristics: first frame of every shot is a key-frame

Image Features

- Color features:
  - Histogram of distribution of color components
  - Average brightness
  - Color moments (including mean, variance, etc.)
  - Dominant color
  - For compressed: use DCT coefficients
- Motion:
  - Camera operations
  - Large moving objects
  - MPEG: motion vectors from B and P frames directly extracted
- Shape and Edge features

Experimental Results

- Do segmentation and keyframe extraction together
- Segmentation accuracy over 95%
  - Includes missing shots and false detection of shot boundaries
  - Camera operations properly detected (panning and zooming)
- Average of 2 to 3 keyframes per shot
  - Accuracy of keyframe extraction is subjective
  - Reasonable comparison with librarians

Shot Content Representation and Similarity Measures

- After partitioning and abstraction
- Compute representation primitives
- Based on these, can index, compare, classify
- Representation primitives based on:
  - key-frames
  - temporal variation and motion information

Representation of Shot Content Based on KeyFrame Features

- **Color**
- Rare that two images of totally different objects have similar colors
- **Color histogram: distribution of color intensities**
- Especially useful for textured images
- Invariant under translation and rotation about view axis
- Change slowly under angle of view, change in scale, occlusion
- Like QBIC: quantized color space into 64 super-cells using sum-of-sqares

- 64-bin color histogram
- Equation in paper: distance between two histograms
- Use matrix aij to represent similarity between colors in bins i and j
- Determined from human visual perception studies

- **Dominant color:**
- Most images: small number of color ranges capture majority of pixels
- Use these dominant colors to construct approximate representation of color distribution
- 20 dominant colors
- May be more accurate than histogram comparisons

- **Color moments**
- Probability distribution characterized by its moments
- Represent color distribution by first three moments
- (Average intensity of each color component, variance and skew)
- Wij weight the contributions of different moments for each color component
- More robust in matching color images than histograms

- **Mean brightness**

Other Key-frame features

- **Texture**
- Can be conveyed with gray-level images
- How to represent texture?

1. Tamura features: quantifications of psychological attributes
   - contrast: based on statistical distribution of pixel intensities
   - directionality: computed from distributions of magnitude and direction of gradient at all pixels
   - coarseness: measure of granularity of texture
     - computed by taking averages over windows of different sizes; take optimum fit

2. SAR Model:
   - Given image of gray-level pixels: provides description of each pixel in terms of neighboring pixels
   - Depends on:
   - mean gray level of entire image
   - dependence of pixel s on neighbors r
   - use independent Gaussian random variable, zero mean, variance models noise level

- **Similarity:**
   - Texture represented as feature vector X
   - Distance between such vectors

Other Key-Frame Features

- Shape features
- Identify shapes by automatic or semi-automatic spatial segmentation algorithms
- Dominant objects obtained by
  - color-based segmentation algorithm
  - interactive outlineing algorithm
- **Compare shapes:**
- Cumulative turning angles
  - Close to human perception
  - Invariant under translation, rotation, scaling
  - Relatively easy computation

Other KeyFrame Features

- **Edge Features**
- Sobel filters
- Compare by calculating correlation between edge maps
- Limited by dependency on resolution, size and orientation

Temporal Features for Representation of Shot Content

- Camera operations
- Detect panning, tilting, zooming by analyzing either motion fields of spatiotemporal images
- Temporal variaion of brightness and colors
- represented by mean and variance of average birghtness and few dominant colors
- Salient stills:Images constructed from content of all frames in a shot
- Capture content and context of entire shot
- Requires sophisticated motion analysis, expensive
- Also: global orientation of images, object motions

Content-Based Retrieval

- Key-frame Based Retrieval
  - **Query by visual templates**
    - Retrieve key-frame images of known color patterns (sky, lawn, forest)
  - **Query by template manipulation**
    - Manipulate color distribution
  - **Query by object feature specification**
  - **Click on key-frame to view video clip**
- Shot-Based Retrieval
  - **Query based on temporal features**
    - e.g., find all shots with camera panning at rate
- Combination of temporal and key-frame features

Browsing Tools

- **Sequential access**
  - VCR-like interface
- **Random access**
  - Hierarchical video browser
  - Top level: few key frames represent entire video
  - Lower levels: smaller groups of shots, then shots, then frames
- Currently: icons at highest level based on sequential behavior over time
- Better to have icons represent shots of similar content