Temporal video segmentation and classification using spatial / spatio-temporal features

Vasileios Mezaris

Thessaloniki, December 9, 2009

Motivation

- Goal: video understanding, for
  - Indexing and retrieval
  - Summarization
  - Personalized delivery
  - …

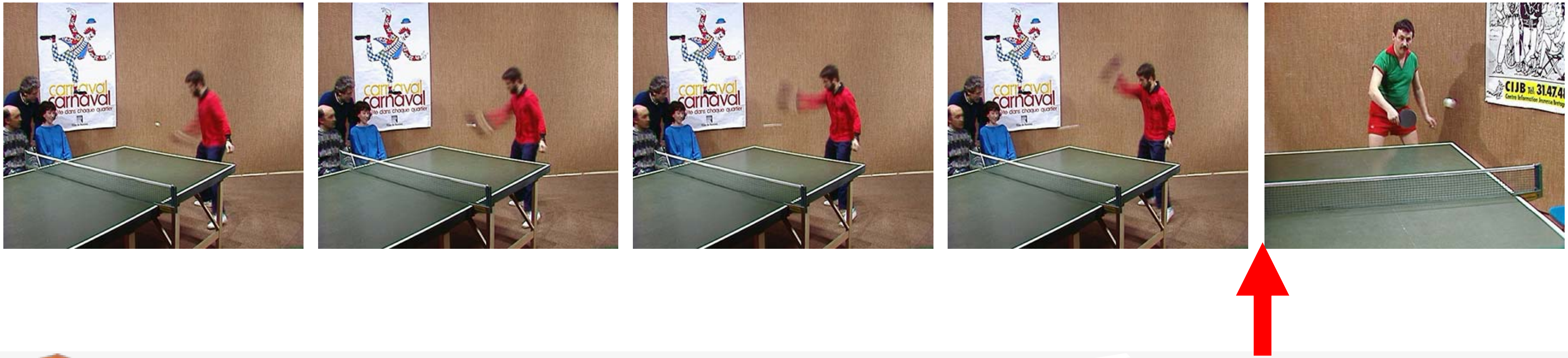
- indoor
- people
- studio
- interview
- face
- male
- war
- Iraq
- attack
- …

Video understanding

- Tasks
  - Temporal segmentation to shots & scenes
  - Spatial / spatio-temporal segmentation to regions / objects
  - Content representation (shot / scene / region / object / …) and classification (single- / multi-label)
  - Context (spatial / temporal / …) exploitation
  - Action & event detection
  - Person & face detection and recognition
  - Knowledge representation & reasoning
  - Associated information (audio / text / metadata / …) processing
  - Multi-modal fusion
  - …

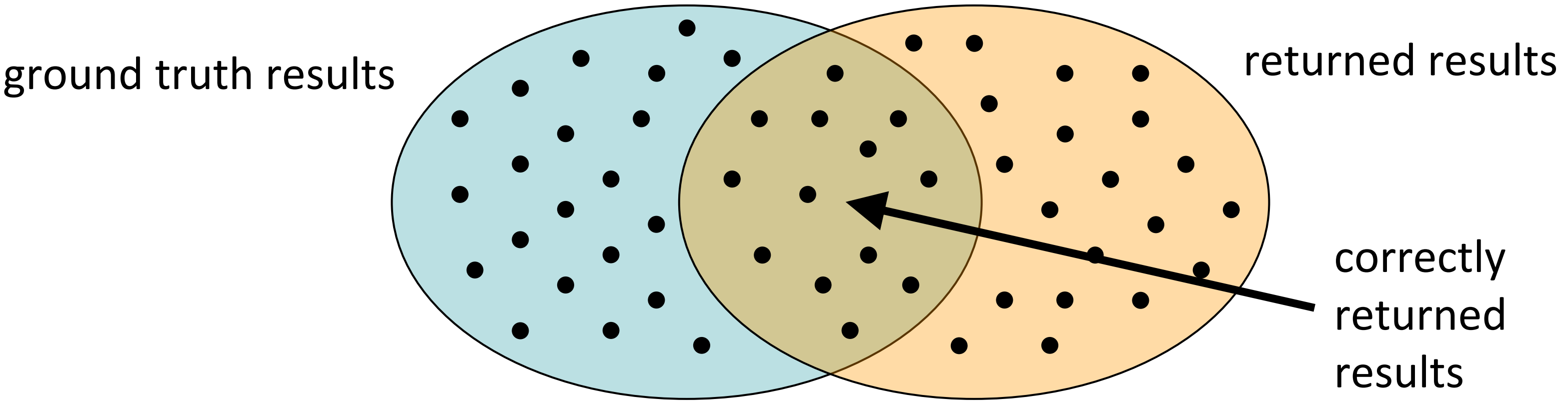
Shot segmentation

- What is a shot?
  - A shot is defined as a sequence of consecutive frames taken without interruption by a single camera
- Shot change is manifested by a change in visual content
  - Abrupt transition
  - Gradual transition (dissolve, fade in / fade out, wipe,…)



Shot segmentation

- Evaluation
  - Precision-Recall
    - Precision: # correctly returned results / # returned results
    - Recall: # correctly returned results / # ground truth results
  - Other criteria
    - e.g. temporal alignment of transition start & end frame

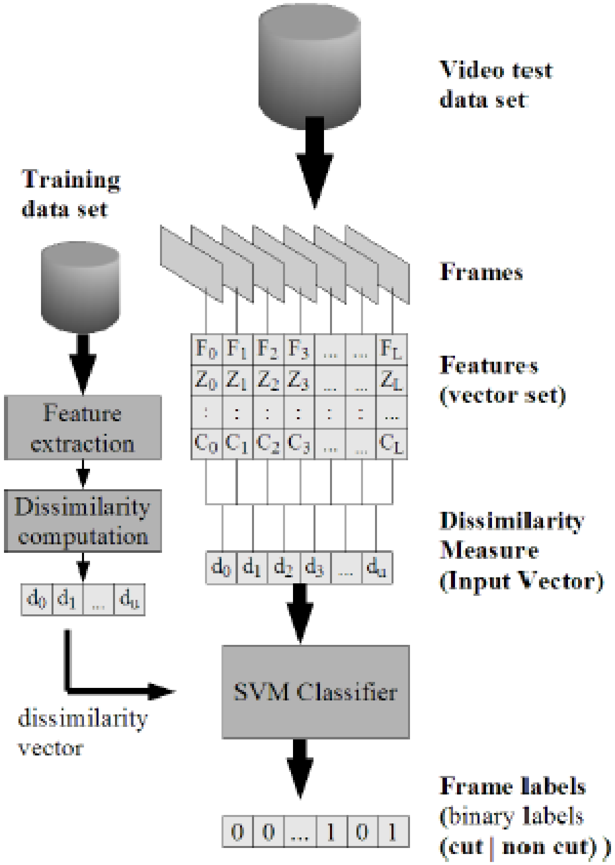


Abrupt transition detection

- Typically, based on pair-wise frame comparisons
  - Raw vs. compressed video
  - Frame representation (features)
  - Pre-processing (e.g. motion compensation)
  - Single vs. multiple criteria
  - Simple thresholding vs. learning

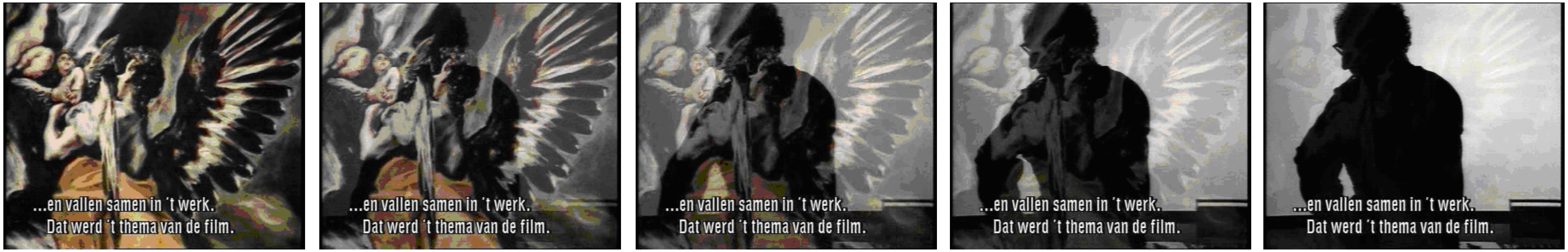
Abrupt transition detection

- Chavez et al., 2006
  - Frame representation
    - Color histograms in RGB, HSV, opponent color space
    - Shape descriptors (Zernike moments, Fourier-Mellin moments
    - Projection histograms
    - Motion
  - Learning
    - Support Vector Machines (SVM)
    - Input vector: feature distances
    - Precision, recall >90% for abrupt transitions



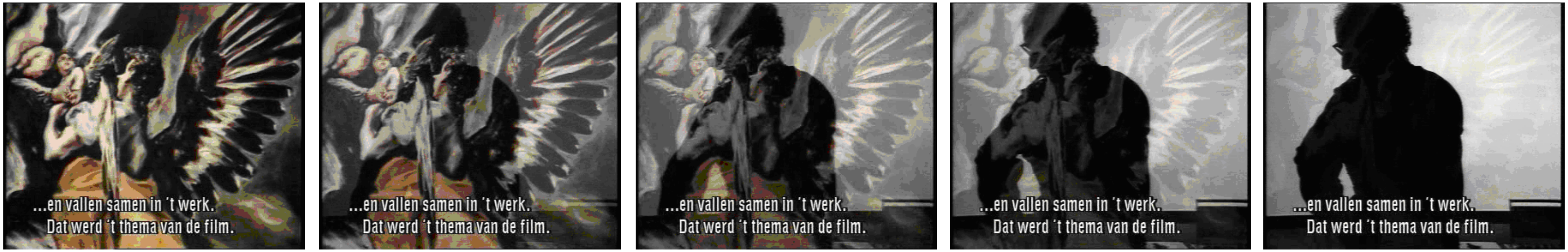
Gradual transition detection

- Pair-wise frame comparisons are not sufficient
  - Gradual transitions are effects with clear temporal dimension
  - Differences between successive frames small, easily confused with normal variations within a shot due to
    - Camera motion
    - Local motion
    - Illumination changes
    - …
  - Temporal evolution of features / feature differences is important

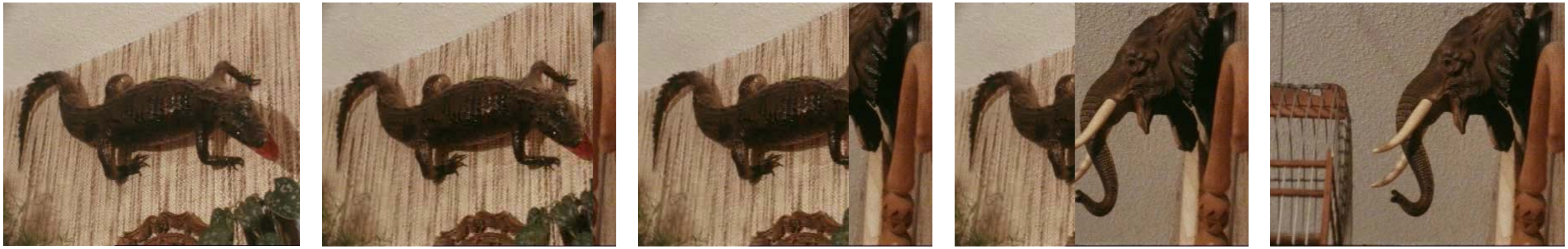

...en vallen samen in 't werk.
Dat werd 't thema van de film.
...en vallen samen in 't werk.
Dat werd 't thema van de film.
...en vallen samen in 't werk.
Dat werd 't thema van de film.
...en vallen samen in 't werk.
Dat werd 't thema van de film.
...en vallen samen in 't werk.
Dat werd 't thema van de film.

Gradual transition detection

- Common types of gradual transitions
  - Dissolve

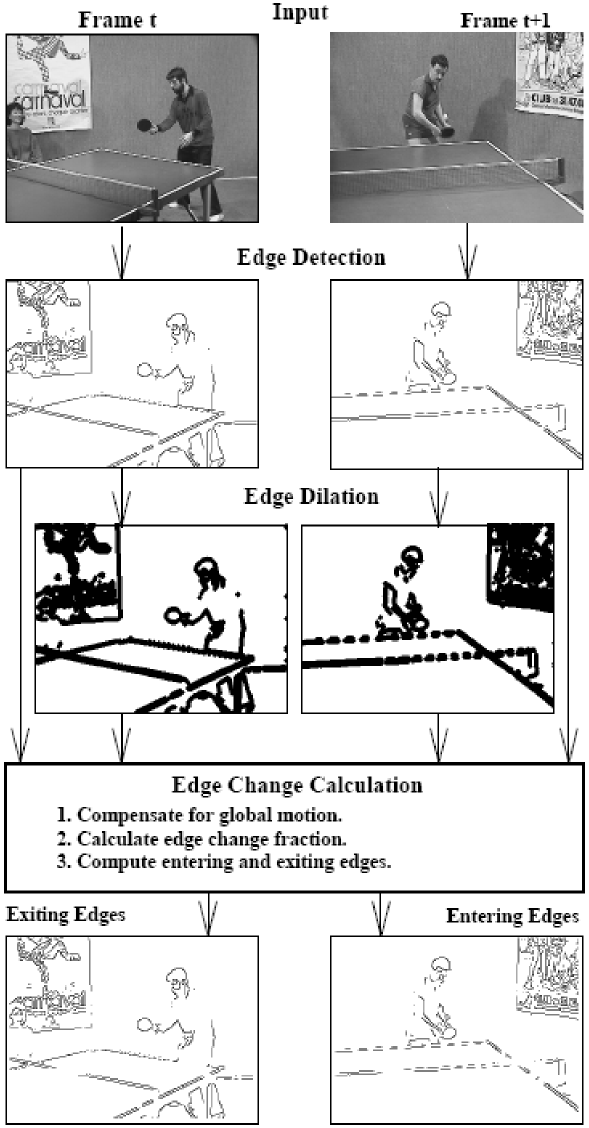

...en vallen samen in 't werk.
Dat werd 't thema van de film.

  - Fade in / fade out: special case of dissolve
  - Wipe



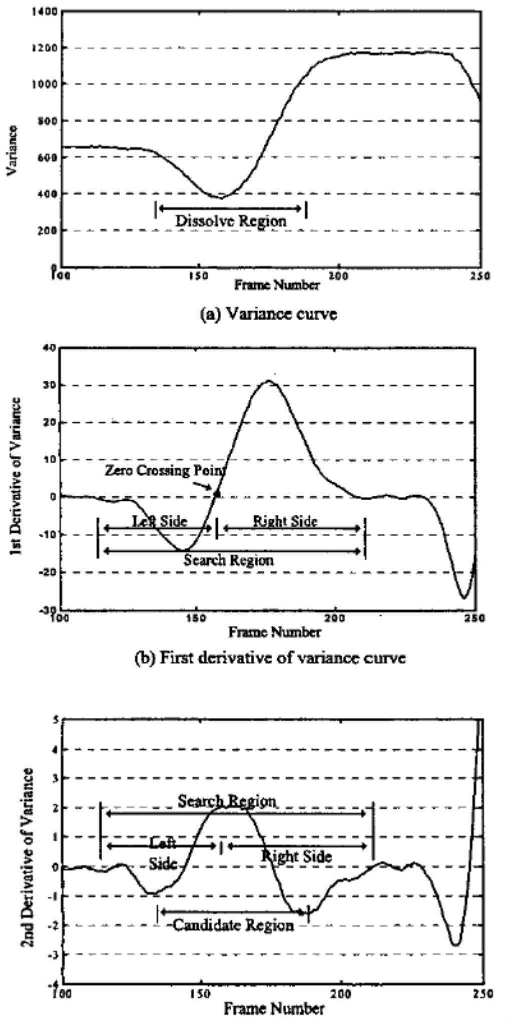
Gradual transition detection

- Zabih et al., 1999
  - Detection of abrupt and different gradual transitions
  - Edge detection
  - Motion compensation
  - Identify exiting / entering edge pixels
  - Edge change fraction (maximum of exiting / entering edge pixel number
  - Find peaks


Frame t
Input
Frame t+1
Edge Detection
Edge Dilation
Edge Change Calculation
1. Compensate for global motion.
2. Calculate edge change fraction.
3. Compute entering and exiting edges.
Exiting Edges
Entering Edges

Gradual transition detection

- Won et al., 2003
  - Dissolve detection
  - Hypothesis: a dissolve is constructed as a linear combination of its start and end frames
  - The variance of pixel intensities within a dissolve region exhibits parabolic shape
  - Possible U-shaped regions are identified in using the first and second derivatives of the luminance variance curve and are verified using an adaptive threshold.



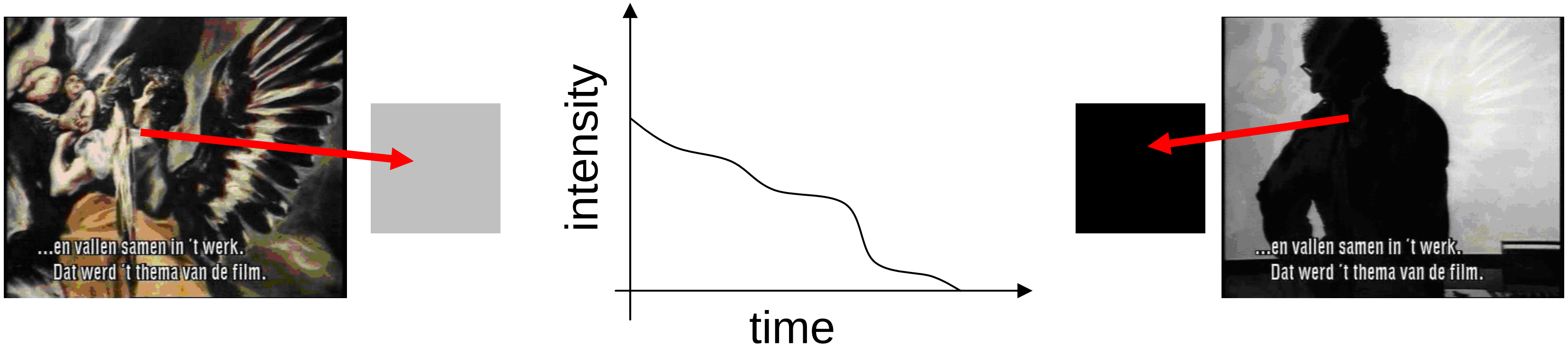
(a) Variance curve

(b) First derivative of variance curve

(c) Second derivative of variance curve

Gradual transition detection

- Su et al., 2005
  - Dissolve detection
  - Hypothesis: a dissolve is constructed as a (not necessarily linear) combination of its start and end frames,
  - …but, pixel intensity changes monotonously during the transition
  - Solution: count the pixels with monotonous intensity change within a time window


…en vallen samen in 't werk.
Dat werd 't thema van de film.
intensity
time
…en vallen samen in 't werk.
Dat werd 't thema van de film.

Gradual transition detection

- Bescos et al., 2005
  - No explicit model of the gradual transition (not limited to e.g. dissolves)
  - Distance function: a variant of the Pearson’s Test (fit of a distribution) for RGB color bands of two frames
  - Distances estimated in 8 different time scales
  - Different threshold-based criteria applied in a cascade
- Ling et al., 2008
  - Use multiple criteria
    - Intensity Pixel-wise Difference
    - Edge histogram differences
    - Color Histogram Difference in HSV Space
  - Extract them for consecutive frames (after removal of smooth intervals of video)
  - Use an SVM classifier

Gradual transition detection

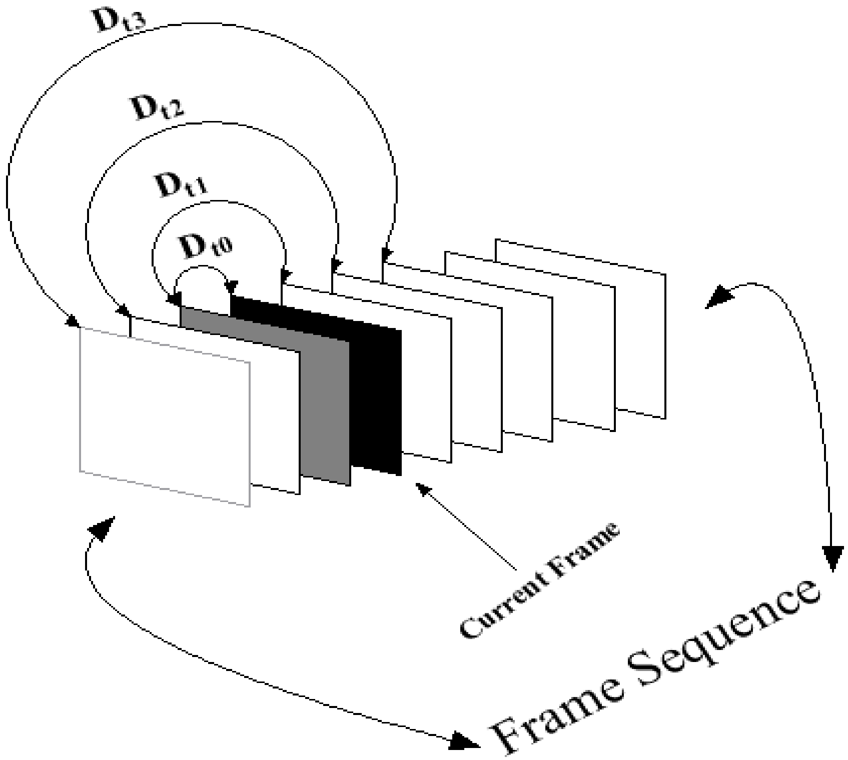
- Proposed approach
  - Combine multiple criteria
    - Individual criteria that exhibit less sensitivity to local or global motion than previously proposed ones
    - Evaluate them simultaneously rather than in a cascade
  - Evaluate criteria at different timescales
  - Use machine learning for classifying pairs of frames to shot change / non-shot-change classes
    - No need for threshold selection

Gradual transition detection

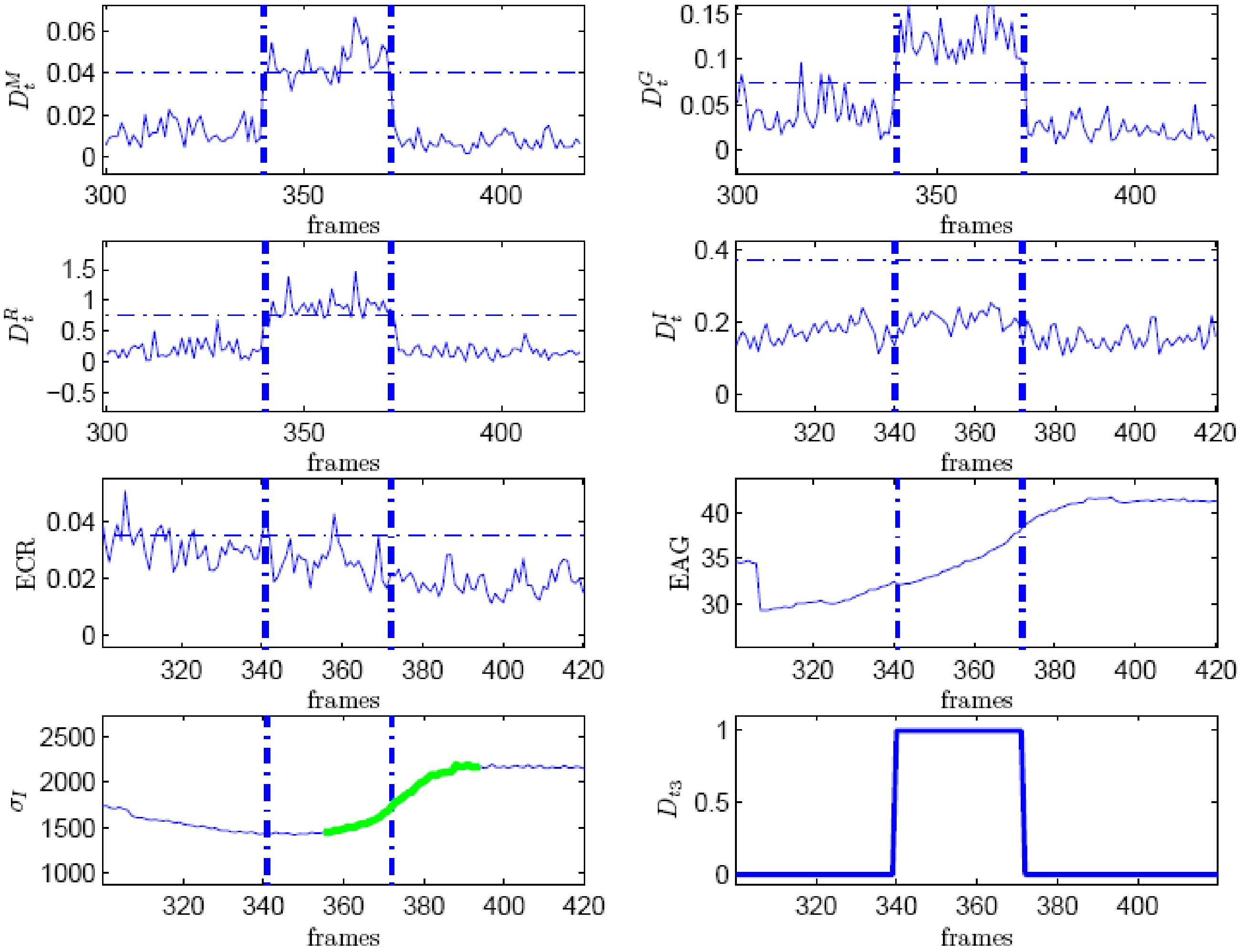
- Individual criteria
  - Macbeth Color Histogram Change (D_t^M)
    - Macbeth pallet consists of twenty four colors, selected according to human color perception
    - Limited number of clusters ensures robustness to slight color variations or noise effects
  - Color Coherence Change (D_t^G)
    - Color Coherence Vectors (CCV) have been proposed for image retrieval applications
    - Distinguish pixels to coherent ones (:belonging to contiguous regions of size greater than x*)*, and incoherent ones
  - Luminance Center of Gravity Change (D_t^R)
    - During a gradual transition the spatial distribution of pixel intensities changes

Gradual transition detection

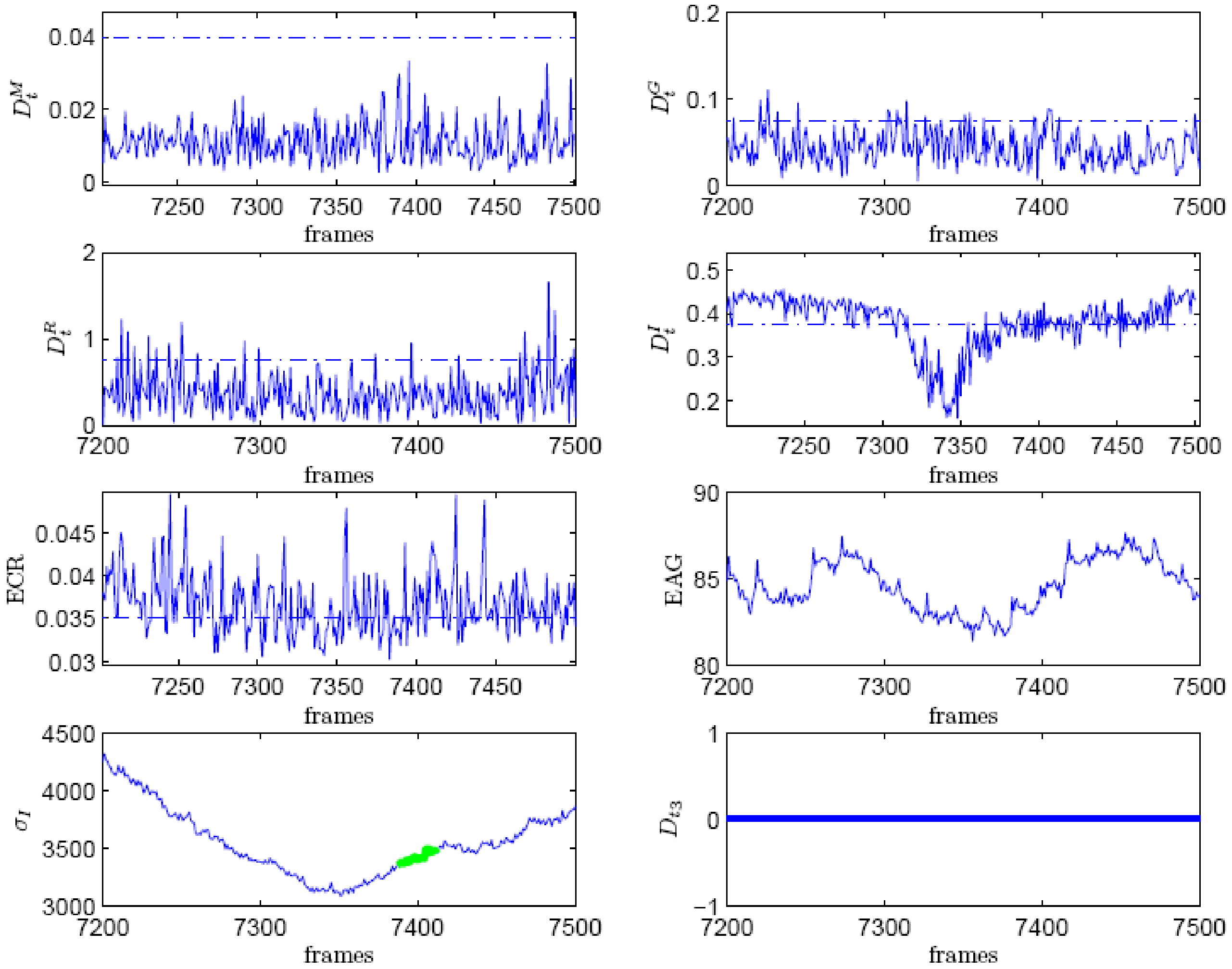
- Timescales and criteria evaluation
  - One feature vector extracted for every frame, with one element per criterion and timescale
  - Input to SVM classifier
  - Final post-processing of SVM output sequence



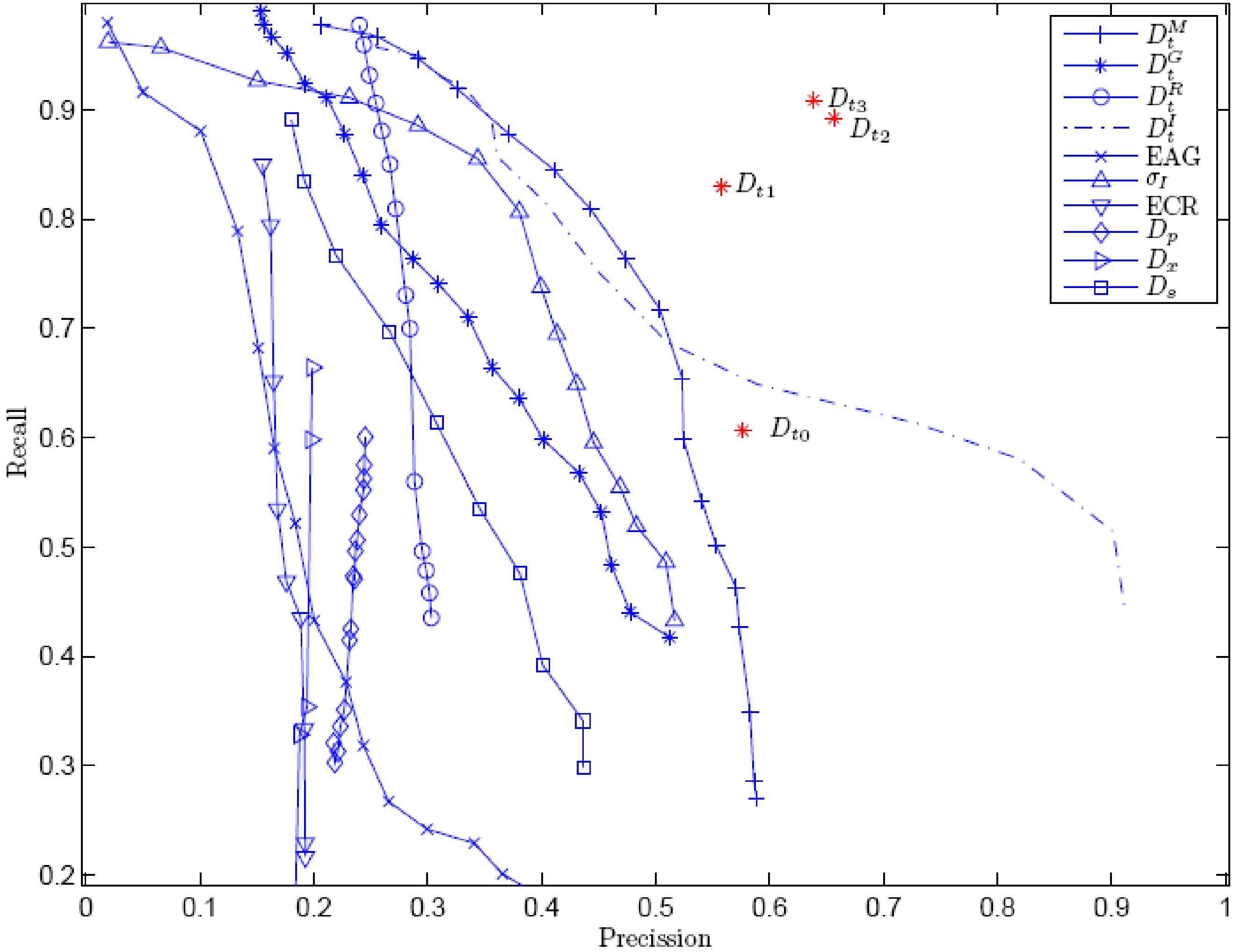
Gradual transition detection



Gradual transition detection



Gradual transition detection

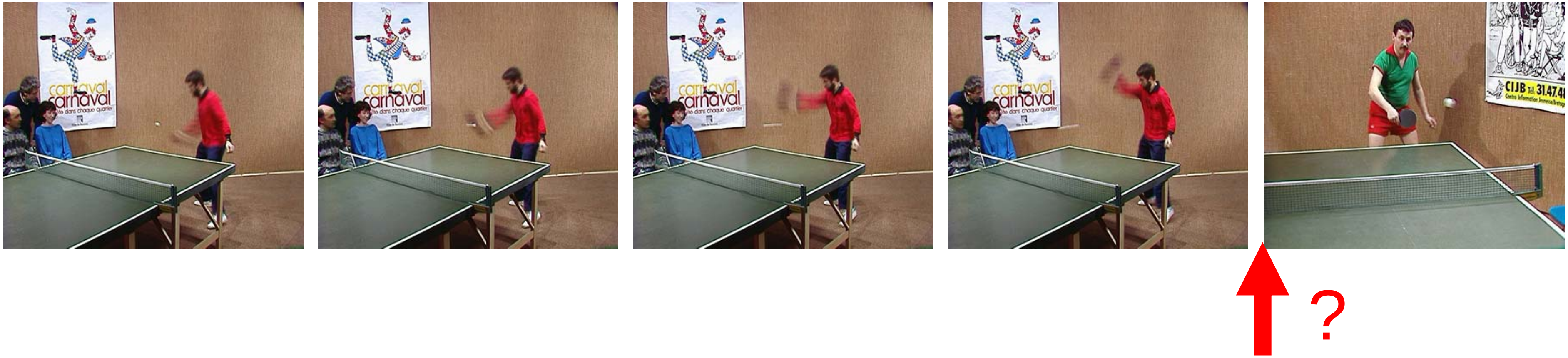


Conclusions

- Abrupt transition detection is a solved problem
  - Precision, recall >90%
  - Sufficient for any practical application
- Gradual transition detection is a bit more difficult
  - Usually, also less important in practice
  - Very high reported scores not necessarily reproducible on different datasets (heuristics / learning affected)
  - Good performance can be achieved by
    - Using a combination of criteria
    - Examining change in different timescales
    - Using machine learning rather than thresholding
- Is shot segmentation all we need?

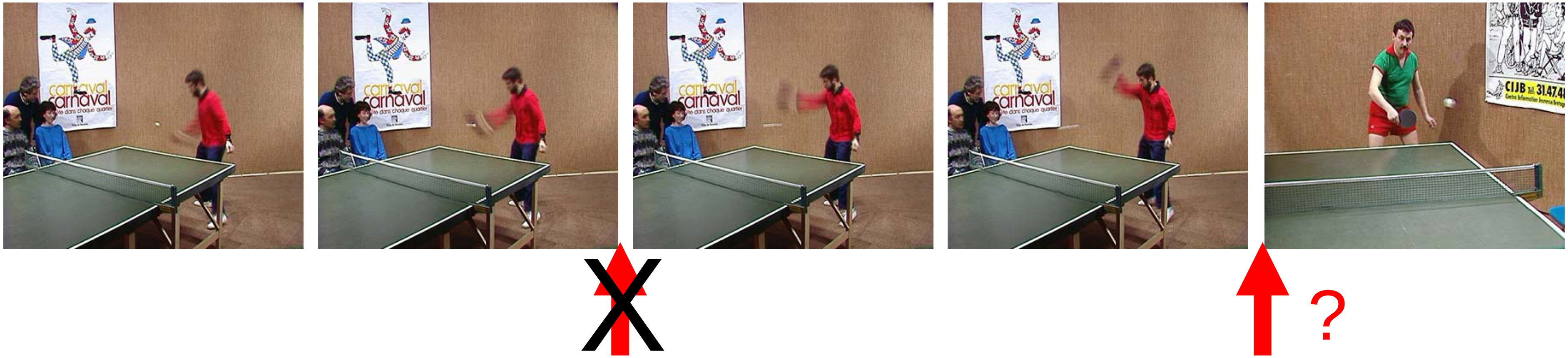
Scene segmentation

- What is a scene?
  - A scene is generally defined as a temporal video segment that is elementary in terms of semantic content
  - Scene segmentation is important for summarization, indexing, video browsing,…
- Scene change is not manifested by a change in visual content alone



Scene segmentation

- Basic assumption
  - A shot cannot belong to more than one scenes
  - Scene boundaries are a subset of the visual shot boundaries of the video
  - Scene segmentation typically performed by
    - Shot segmentation, and
    - Shot grouping



Scene segmentation

- Common semantic content can be defined in more than one ways
  - Logical Story Unit (LSU) Definition: A series of temporally contiguous shots characterized by overlapping links that connect shots with similar visual/audio content [Hanjalic et al., 1999]
  - Film Production Scene Definition: A small number of interrelated shots that are unified by location or dramatic incident [Beaver, 1994]
  - Other: A temporal video segment for which three properties, event or dramatic incident, setting and time are consistent [Petersohn, 2009]

Scene segmentation

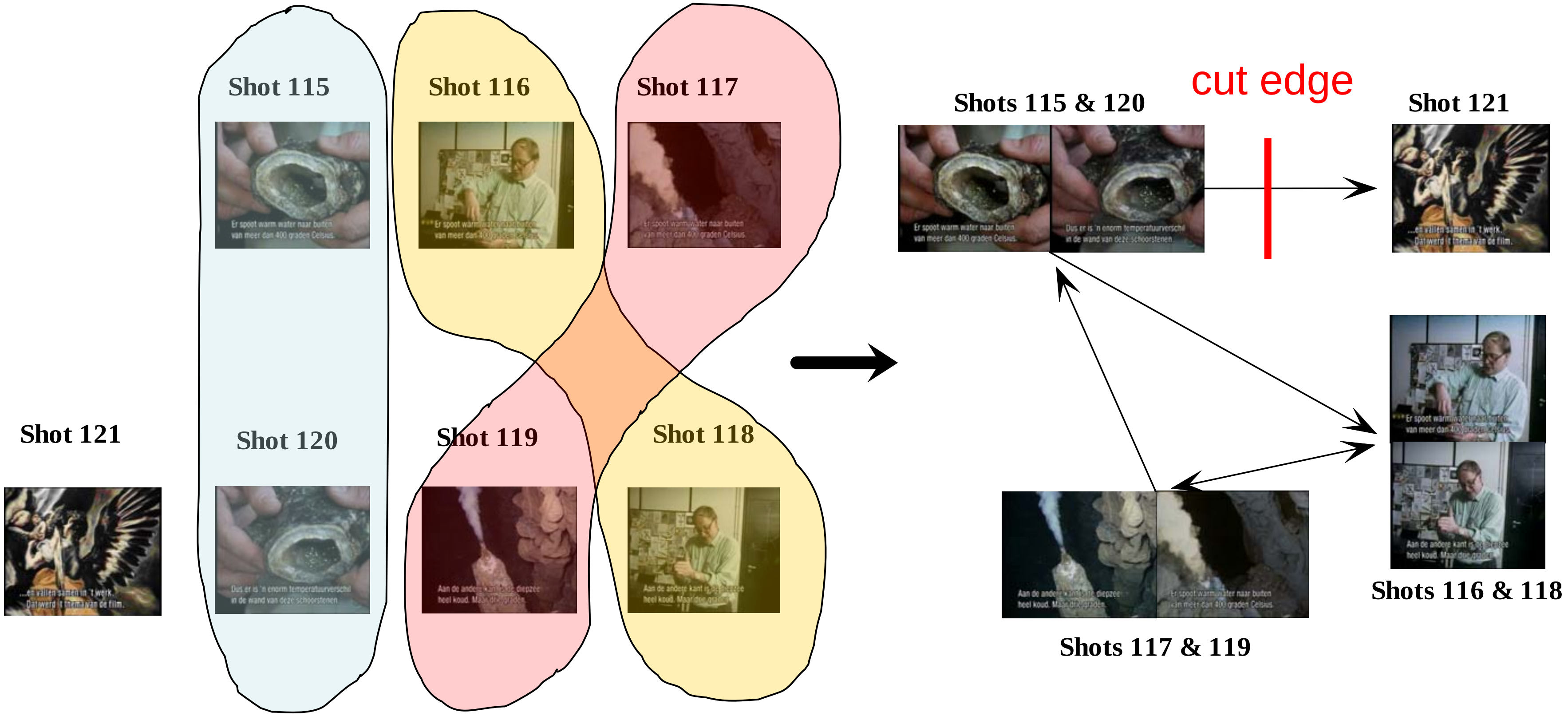
- Literature
  - Differences in applicability
    - Domain-specific techniques
      - News video (using knowledge on news structure)
      - TV broadcasting (based on advertisement detection)
      - …
    - Domain-independent techniques
  - Differences in employed information
    - Uni-modal techniques
      - Visual information only
    - Multi-modal techniques
      - Visual, audio, speech transcripts, …

Uni-modal scene segmentation

- Yeung et al., 1998
  - Key-frame description using HSV histogram
  - Inter-shot similarity: the minimum distance between pairs of feature vectors belonging to different shots
  - Pair of shots with similarity and temporal distance less than empirical thresholds are grouped into the same cluster
  - A Scene Transition Graph (STG) is constructed
    - Nodes represent shot clusters
    - A directed edge is drawn from a node to another if there is a shot represented by the first node that immediately precedes any shot represented by the second node
  - The set of cut-edges is the set of scene boundaries
    - Cut-edge is defined as an edge, which if removed, results in two disconnected graphs

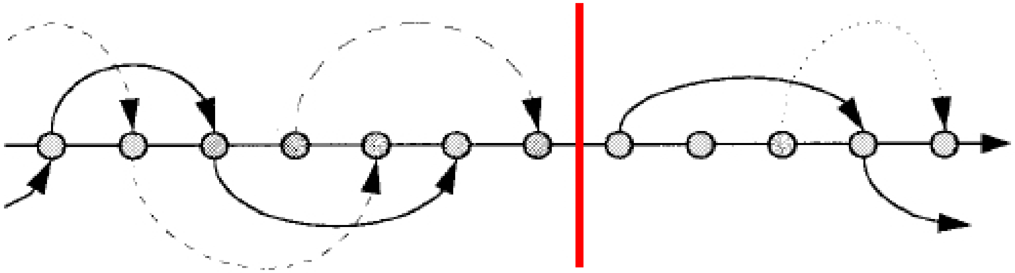
Uni-modal scene segmentation

- Scene transition graph



Uni-modal scene segmentation

- Hanjalic et al., 1999
  - Shot segmentation and key-frame extraction, “shot image” created by merging all key-frames of a shot
  - Shot images are divided into 8x8 blocks; average color in L*u*v* space is calculated for each block
  - Shot dissimilarity is computed by assigning each block of the first shot image into a block of the second shot image and summing the block distances
  - Shot links are extracted through thresholding shot dissimilarities; All shots boundaries in the interior of a shot link are not scene boundaries

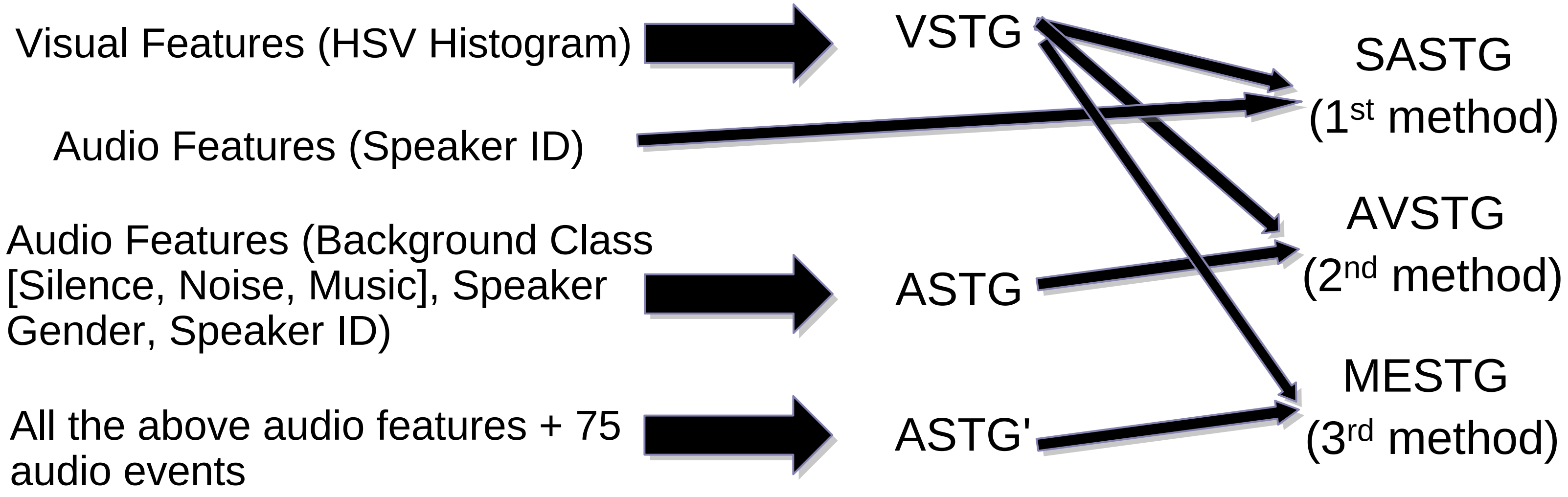


Multimodal scene segmentation

- Chen et al. 2002
  - Shot audio descriptors (volume, energy, spectral flux etc.)
  - Consecutive shots whose audio description difference exceeds an empirical threshold are assigned to different scenes
- Nitanda et al., 2005
  - Decomposition of auditory channel into audio segments, each belonging to one of 5 classes (silence, speech, music etc.).
  - Scene change: shot boundary exist within an empirical time interval before or after an audio segment boundary
- Goela et al. 2007
  - Low- and higher-level audio features (MFCC coefficients; music, speech, laughter, silence classification); visual features (average shot count within a time window)
  - Feature vector serves as input to a SVM classifier

Multimodal scene segmentation

- Proposed approaches
  - Multi-modal extensions of STG
  - Significantly improved performance over visual-only STG
  - Reduced sensitivity to STG construction thresholds

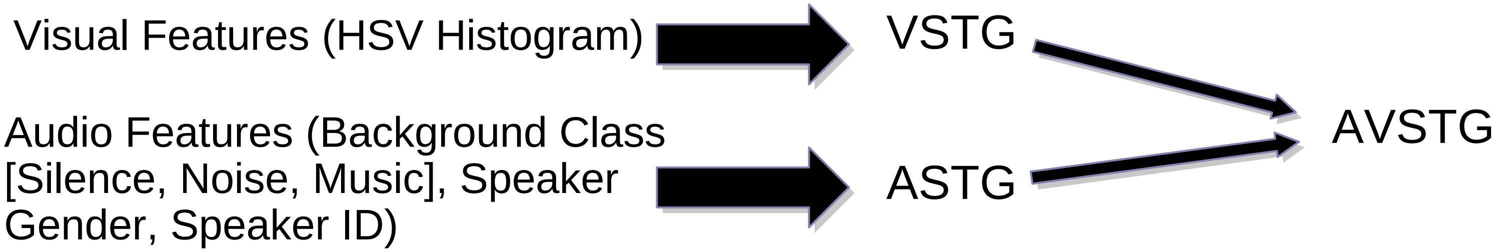

Visual Features (HSV Histogram)
Audio Features (Speaker ID)
Audio Features (Background Class [Silence, Noise, Music], Speaker Gender, Speaker ID)
All the above audio features + 75 audio events
VSTG
ASTG
ASTG'
SASTG
(1st method)
AVSTG
(2nd method)
MESTG
(3rd method)

Multimodal scene segmentation

- Speaker Assisted Scene Transition Graph (SASTG)
  - Uses speaker segmentation and clustering results (Speaker ID) to post-process a visual STG
  - Algorithm
    - A visual STG is constructed and initial scene boundaries are estimated
      - Construction parameters favor over-segmentation
    - For each initially extracted scene boundary
      - Construct the two sets of speaker segments within a time window before / after it
      - If speaker segments with the same speaker are included in both sets and in temporal distance less than an empirical threshold, the scene boundary is rejected

Multimodal scene segmentation

- Audio-Visual Scene Transition Graph (AVSTG)
  - Uses background class (Silence, Noise, Music), speaker gender, speaker ID results
  - Algorithm
    - A visual STG (VSTG) is constructed
    - An audio STG (ASTG) is constructed
    - The two STGs are “merged”



Visual Features (HSV Histogram)

Audio Features (Background Class [Silence, Noise, Music], Speaker Gender, Speaker ID)

VSTG

ASTG

AVSTG

Multimodal scene segmentation

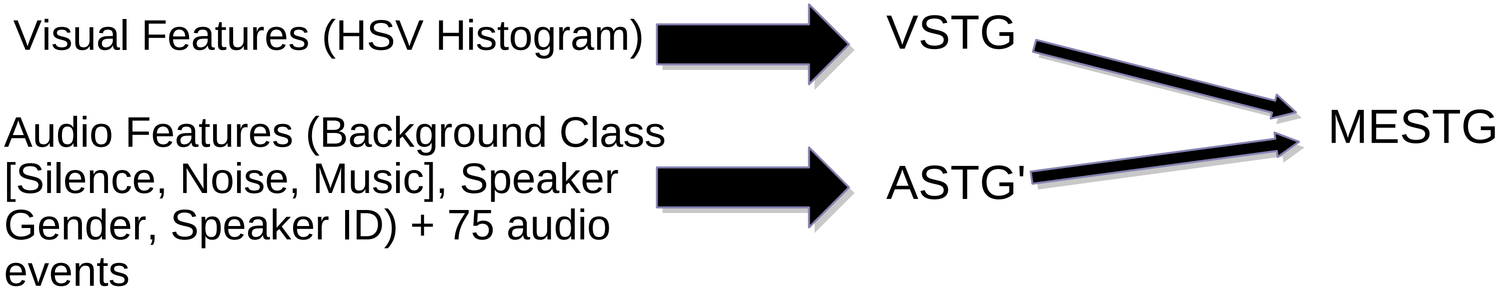
- ASTG construction assumptions
  - Each set of temporally consecutive audio segments that share the same speakers and background conditions cannot belong to more than one scenes
  - The distribution of speaker identities can serve as a measure of audio segment similarity
- Algorithm
  - Temporally adjacent audio segments with common speakers and background conditions are merged
  - Visual shots are merged to video units according to the audio segmentation (if two or more shots overlap with a single audio segment)
  - Video units are clustered according to speaker identity distribution similarity and time adjacency
  - A graph is formed, with nodes representing video unit clusters and directed edges connecting two nodes if a video unit in the first node is succeeded by one in the second node

Multimodal scene segmentation

- ASTG and VSTG merging
  - ASTG
    - Multiple ASTGs with random parameter sets are constructed
    - For each shot boundary, the probability (or “confidence”) of it also being a scene boundary is estimated
  - VSTG
    - Similar process (multiple VSTGs constructed, confidence values estimated)
  - The two confidence values coming from ASTG, VSTG are linearly combined
  - If the combination exceeds a threshold, a scene change is declared
  - Linear combination weights can be learned (e.g. estimated by LSE)
- Advantages
  - Improved performance
  - Lower dependence on STG construction parameters (similarity thresholds, temporal windows,…)

Multimodal scene segmentation

- Multiple Evidence Scene Transition Graph (MESTG)
  - Uses background class (Silence, Noise, Music), speaker gender, speaker ID results + **75 audio events**
  - Algorithm
    - A visual STG (VSTG) is constructed
    - An audio STG (ASTG) is constructed
    - The two STGs are “merged”



Visual Features (HSV Histogram)

VSTG

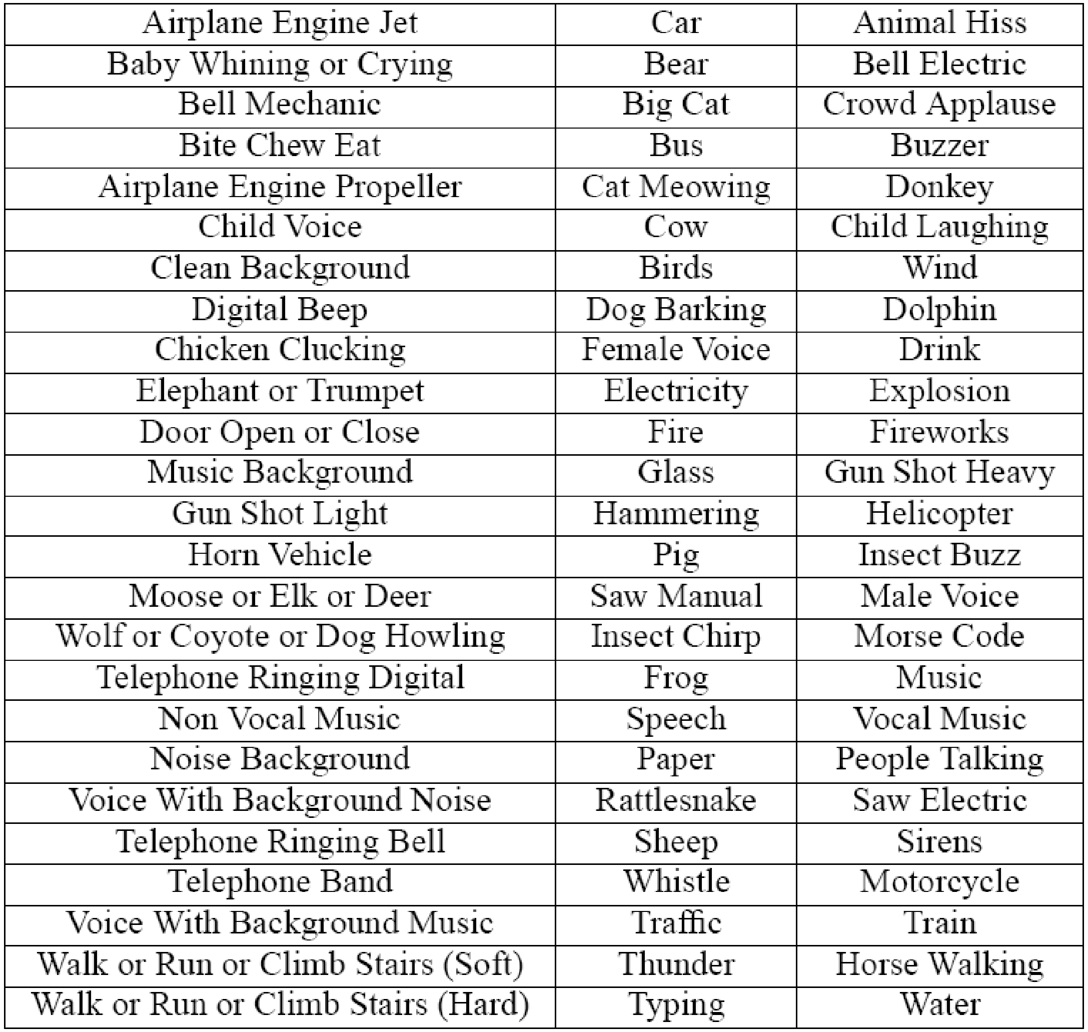
Audio Features (Background Class [Silence, Noise, Music], Speaker Gender, Speaker ID) + 75 audio events

ASTG'

MESTG

Multimodal scene segmentation

- Audio events
  - Semantically elementary audio content
  - Extracted using
    - Multi-layer Perceptrons or GMMs (14 events)
    - SVMs (61 events)
  - A confidence value in [0,1] is returned for each event and each audio segment → 75 element audio event vector



Airplane Engine Jet	Car	Animal Hiss
Baby Whining or Crying	Bear	Bell Electric
Bell Mechanic	Big Cat	Crowd Applause
Bite Chew Eat	Bus	Buzzer
Airplane Engine Propeller	Cat Meowing	Donkey
Child Voice	Cow	Child Laughing
Clean Background	Birds	Wind
Digital Beep	Dog Barking	Dolphin
Chicken Clucking	Female Voice	Drink
Elephant or Trumpet	Electricity	Explosion
Door Open or Close	Fire	Fireworks
Music Background	Glass	Gun Shot Heavy
Gun Shot Light	Hammering	Helicopter
Horn Vehicle	Pig	Insect Buzz
Moose or Elk or Deer	Saw Manual	Male Voice
Wolf or Coyote or Dog Howling	Insect Chirp	Morse Code
Telephone Ringing Digital	Frog	Music
Non Vocal Music	Speech	Vocal Music
Noise Background	Paper	People Talking
Voice With Background Noise	Rattlesnake	Saw Electric
Telephone Ringing Bell	Sheep	Sirens
Telephone Band	Whistle	Motorcycle
Voice With Background Music	Traffic	Train
Walk or Run or Climb Stairs (Soft)	Thunder	Horse Walking
Walk or Run or Climb Stairs (Hard)	Typing	Water

Multimodal scene segmentation

- Audio event usage
  - Confidence level normalization
    - Diversity of the distribution of confidence values among different event detectors; differences in the actual frequency of appearance of different events within a video
    - Simple normalization used

$$\tilde{ev}(j) = \frac{ev(j)}{maxev_j}$$

  - Similarity metric
    - Similarly high confidence levels for one event in two segments reveal significant segment similarity; similarly low confidence levels do not
    - Minkowski distance unsuitable (depends only on the difference of event vectors)
    - Variant of Chi-test distance

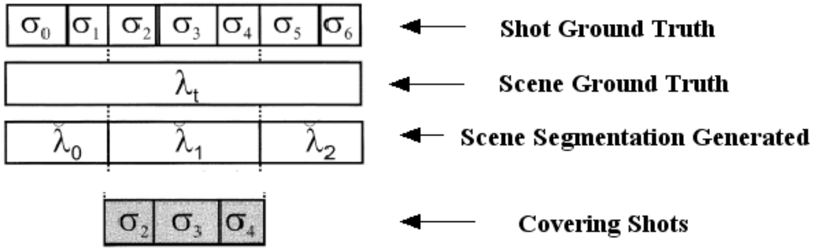
$$D(\tilde{EV}_1, \tilde{EV}_2) = \sqrt{\sum_{j=1}^{J} \frac{(\tilde{ev}_1(j) - \tilde{ev}_2(j))^2}{\tilde{ev}_1(j) + \tilde{ev}_2(j)}}$$

Multimodal scene segmentation

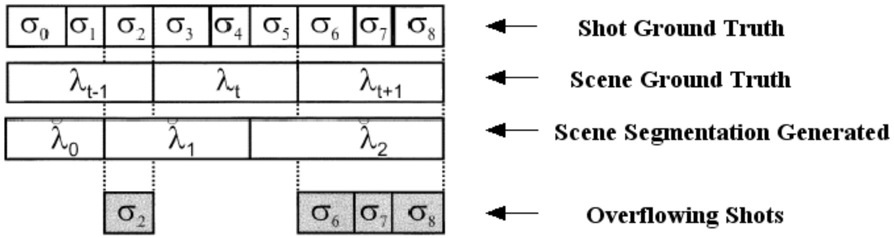
- ASTG' construction assumptions
  - Each set of temporally consecutive audio segments that share the same speakers and background conditions **and exhibit similar audio events** cannot belong to more than one scenes
  - The distribution of speaker identities **and audio events** can serve as a measure of audio segment similarity
- Algorithm
  - Temporally adjacent audio segments with common speakers and background conditions **and similar audio event vectors** are merged
  - Visual shots are merged to video units according to the audio segmentation (if two or more shots overlap with a single audio segment)
  - Video units are clustered according to speaker identity distribution similarity, **audio event vector similarity** and time adjacency
  - A graph is formed, with nodes representing video unit clusters and directed edges connecting two nodes if a video unit in the first node is succeeded by one in the second node

Scene segmentation

- Evaluation
  - Precision-Recall could be used, but they do not capture the temporal displacement of scene boundaries
  - Coverage-Overflow [Vendrig et al. 2002]
    - Coverage: The fraction of shots of automatically generated scenes that overlap the most with the ground-truth units (best: 100%)
    - Overflow: The mean overlap of automatically generated scenes that cover any ground-truth scene with the previous and subsequent ground-truth units (best: 0%)



σ_0 σ_1 σ_2 σ_3 σ_4 σ_5 σ_6
Shot Ground Truth
λ_t
Scene Ground Truth
λ_0 λ_1 λ_2
Scene Segmentation Generated
σ_2 σ_3 σ_4
Covering Shots



σ_0 σ_1 σ_2 σ_3 σ_4 σ_5 σ_6 σ_7 σ_8
Shot Ground Truth
λ_{t-1} λ_t λ_{t+1}
Scene Ground Truth
λ_0 λ_1 λ_2
Scene Segmentation Generated
σ_2 σ_6 σ_7 σ_8
Overflowing Shots

Scene segmentation

- Experiments
  - Test Database
    - 7 documentary films from the collection of the Netherlands Institute for Sound and Vision - duration 229 minutes
    - Ground truth manually generated - 237 scenes

Method	VSTG	SASTG	AVSTG	MESTG	[Nitanda et al.]
Coverage (%)	79.18	81.22	83.86	**85.75**	77.93
Overflow (%)	17.81	12.28	11.05	**10.71**	13.88

Conclusions

- The Scene Transition Graph is a suitable technique for scene segmentation
  - It is possible to limit the influence of STG construction parameters (similarity thresholds, temporal windows,…) on the results
- Good performance can be achieved by
  - Using visual and high-level audio information (speaker IDs, background classification etc.)
  - Audio events (although their detection is imperfect)
  - The richer the information employed, the better the results
- Next issue: how do we classify shots or scenes into semantic classes?

Shot representation and classification

- Objective: associate shots with classes of content
  - What kind of content? What kind of classes? What kind of association? (one-to-one / one-to-many; hard / soft; …)
    - TRECVID high-level feature extraction task
- Shot representation
  - Keyframe features (color / texture / structure / …; global / local; …)
    - Interest points & Bag-of-Words
  - Motion information (global / local motion)
    - Spatio-temporal interest points & interest point tracks
  - Temporal evolution of features; audio features; text transcripts; …
- Classification
  - Support Vector Machines
  - Hidden Markov Models
  - …

Interest points

- Definition (Wikipedia)

  An interest point is a point in the image that

  - Has a clear, preferably mathematically well founded, definition
  - Has a well defined position in image space
  - Has a neighborhood rich in local information content
  - Is stable under perspective, scale, illumination variations
  - Has a high degree of reproducibility

Interest points

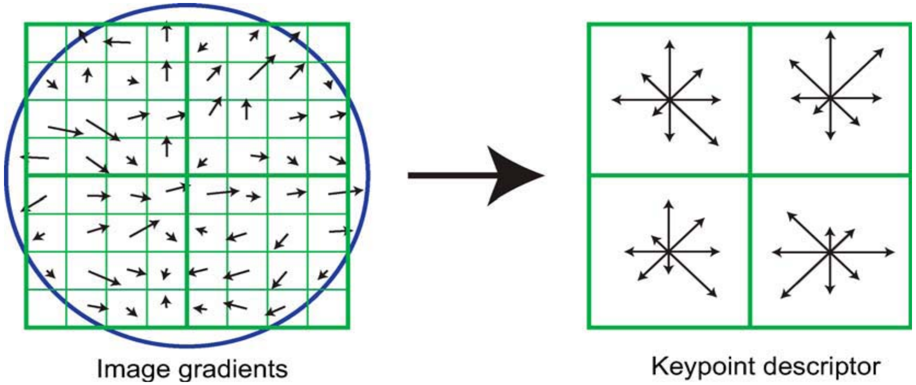
- Detectors
  - Difference of Gaussians
  - Laplacian of Gaussians
  - Harris detector (second moment matrix)
  - Harris - Affine
  - Harris - Laplace
  - Hessian
  - …

Interest points

- Descriptors
  - SIFT (Scale-invariant feature transform)
  - SURF (Speeded Up Robust Features)
  - GLOH (Gradient Location and Orientation Histogram)
  - LESH (Local Energy based Shape Histogram)
  - Color SIFT (several variations)
  - Affine SIFT
  - …

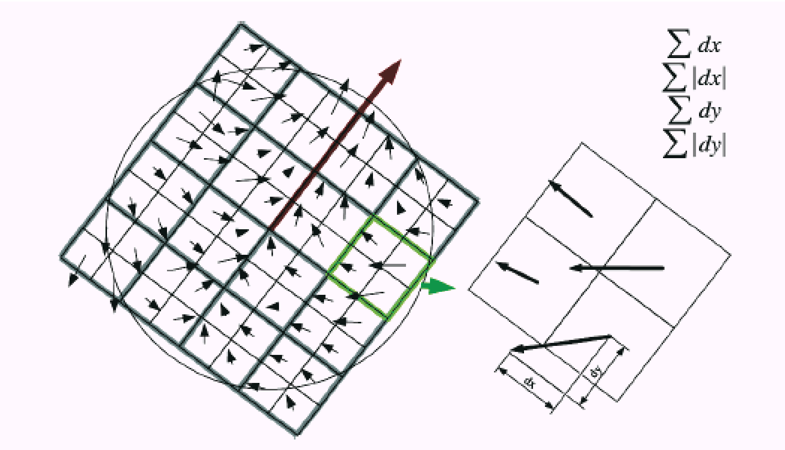
Interest points

- Lowe, 2004
  - SIFT (Scale-invariant feature transform)
  - Difference-of-Gaussian Based Interest Points
  - Interest Point Description
    - Orientation Assignment
    - Descriptor based on histograms of local gradients and orientations



Interest points

- Bay et al., 2008
  - SURF (Speeded Up Robust Features)
  - Hessian Matrix Based Interest Points
  - Interest Point Description
    - Orientation Assignment
    - Descriptor based on Sum of Haar Wavelet Responses

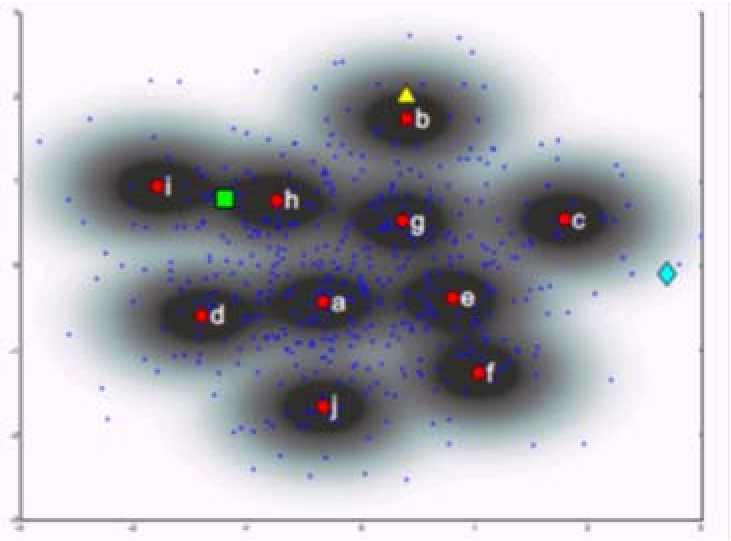


Bag-of-Words

- Codebook generation
  - Partition descriptor space in N subspaces by clustering
  - Define Codewords as cluster centroids
- Image representation
  - Assign descriptors to Codewords
  - Create histogram of Codewords
    - Hard assignment
    - Soft assignment (Visual Word Ambiguity)

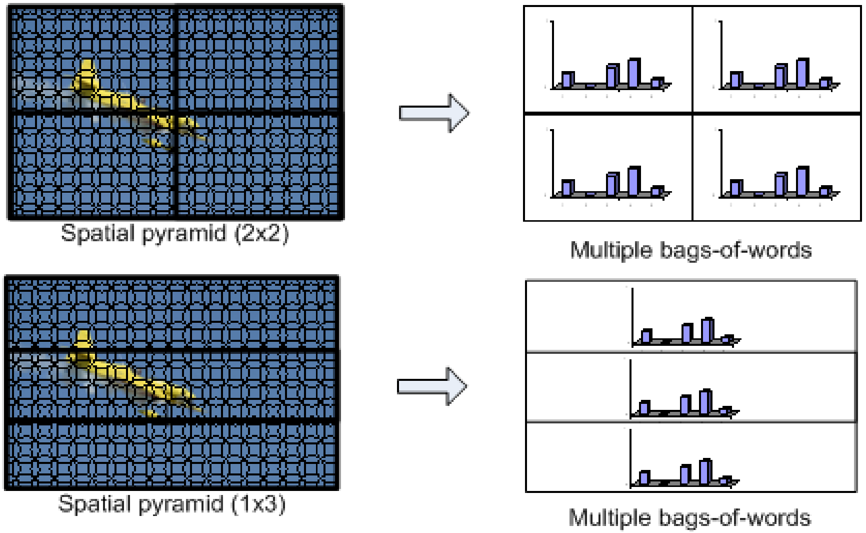
Bag-of-Words

- van Gemert et al., 2010
  - Problems in BoW model
    - Visual word Uncertainty
    - Visual word Plausibility
  - Solution: Kernel Codebooks
    - Compute distance to each codeword
    - Employ (gaussian) kernel to model codeword probability
    - Create histogram of codeword occurrence probabilities



Bag-of-Words

- Pyramidal decomposition [Lazebnik et al.,2006]
  - Goal: Introduce localization awareness in BoWs
  - Create spatial pyramids (2x2, 1x3)
  - Extract SIFT descriptors for each sub-region
  - Create multiple Bags-of-Words



Spatial pyramid (2x2)

Multiple bags-of-words

Spatial pyramid (1x3)

Multiple bags-of-words

Bag-of-Words

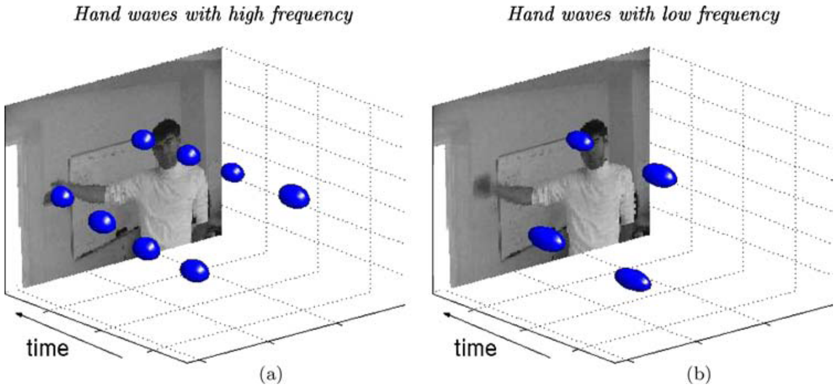
- Snoek et al.,2008
  - Goal: Exploit color channel information
  - Solution: use multiple color SIFT descriptors
  - Color SIFT in different color spaces
    - Opponent-SIFT
    - C-SIFT
    - rgSIFT
    - Transformed Color-SIFT
    - Hue-SIFT
  - Multiple Bag-of-Words for color SIFT variants
  - Concatenate BoWs in single descriptor

Spatio-temporal extensions

- Motivation: Exploit motion information
- Methodologies:
  - Indentify & describe interest points (volumes) in space-time
  - Identify & describe trajectories of 2D interest points

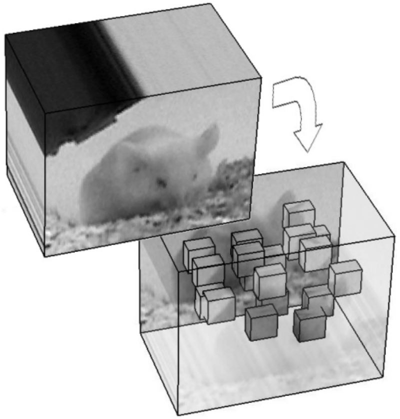
Spatio-temporal extensions

- Laptev et al., 2005
  - Goal: event detection, human motion analysis
  - Solution: Space-Time Interest Points
    - Points having large variations along both spatial and temporal directions (spatio-temporal corners, e.g. when the motion of a part is reversed)
    - Describe space-time interest points with scale-invariant descriptors
    - Cluster points
    - Model events based on cluster membership, position, time of observed points



Spatio-temporal extensions

- Dollar et al., 2005
  - Similar idea: find space-time interest points
    - Different detector (with a high response not only in spatio-temporal corners but also in volumes where periodic motions occur etc.)
  - Identify cuboids (3D regions around interest points)
  - Test different cuboid descriptors (e.g. local histograms of brightness, gradient, etc)
  - Employ a bag-of-cuboids approach
- In general, interest points in space-time
  - Have difficulties under global (camera) motion
  - Do not exploit long-term motion information



Spatio-temporal extensions

- Loccoz et al., 2006
  - Goal: event based video indexing
  - Solution: use local motion
    - Identify trajectories of similar scale-invariant interest points
    - Quantize local trajectories and describe with multi-scale histograms
    - Bag-of-words for the trajectories (motion information only)
- Anjulan et al., 2009
  - Goal: video object retrieval
  - Solution:
    - Create tracks of interest points with similar SIFT descriptors
    - Describe shots with the average SIFT descriptor for each track
    - Cluster them to “objects” based on visual similarity and spatial proximity

Spatio-temporal extensions

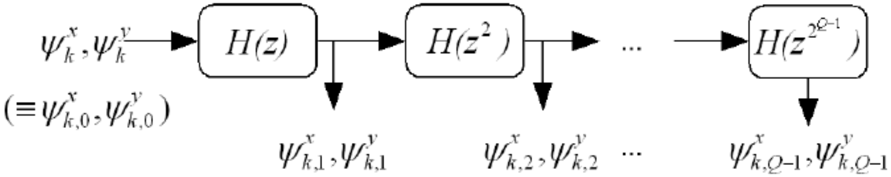
- Proposed approach:
  - Extract tracks of SIFT points from the entire shot
  - Select longer tracks to represent the shot
  - Represent tracks with appropriate descriptor
  - Create Bag-of-Spatiotemporal-Words
    - Model shots is a space of “similar in appearance, similarly moving local regions” rather than “similar in appearance local regions” or “similar local motion patterns”
- Advantages
  - Appearance (2D) and motion information are treated together
  - Long-term motion information is exploited
  - Invariance to scale, camera motion, …

Spatio-temporal extensions

- Extract tracks of SIFT points from the entire shot
  - Temporally sub-sample frames by a factor of *a*
  - Extract SIFT interest points + descriptors from every remaining frame
  - Match SIFT descriptors between successive frames and store motion vectors (append to existing tracks; start new)
  - Estimate global motion (8-parameter bilinear model) using least squares and iterative rejection, and compensate motion vectors
  - Matching process selected for simplicity; more elaborate matching possible (to account for occlusions in part of the frame sequence etc.)
- Select longer tracks to represent the shot
  - Longer tracks: more stable local regions
  - Choose *N* longer tracks for robustness

Spatio-temporal extensions

- Motion track Representation
  - Average SIFT descriptor for the track (128 elements)
  - Motion information
    - Filter-bank for capturing motion at different time scales


ψ_k^x, ψ_k^y
$(\equiv \psi_{k,0}^x, \psi_{k,0}^y)$
$H(z)$
$H(z^2)$
...
$H(z^{2^{Q-1}})$
$\psi_{k,1}^x, \psi_{k,1}^y$
$\psi_{k,2}^x, \psi_{k,2}^y$
...
$\psi_{k,Q-1}^x, \psi_{k,Q-1}^y$

    - For every time scale, histograms of motion direction are created at multiple granularity levels (4-bins, 8-bins, …)
    - Four time-scales, 3 motion direction granularity levels (112 elements)
  - Concatenated vector of 240 elements jointly describing the appearance and motion of a local region
- BoW model created as for any other descriptor

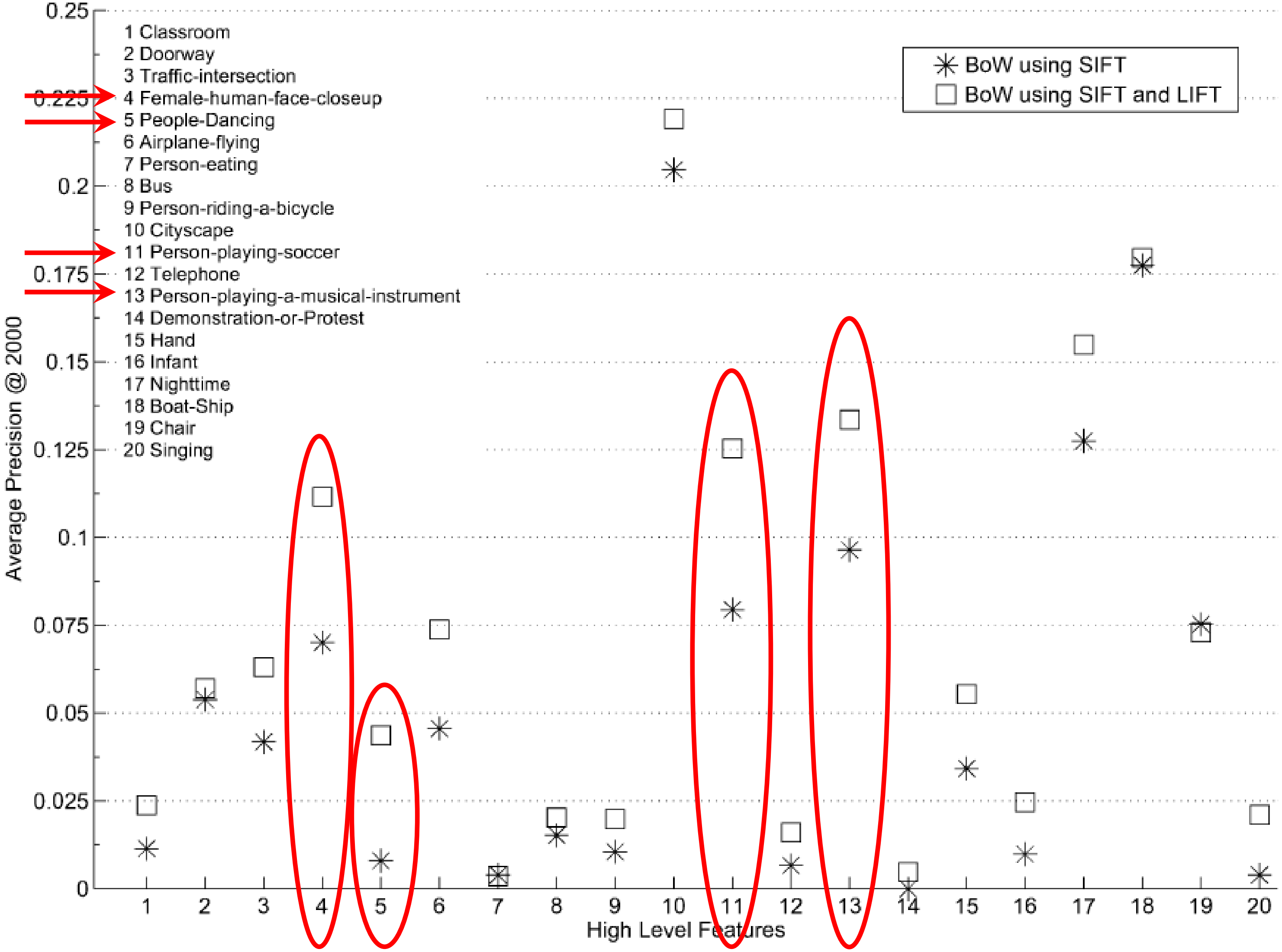
Spatio-temporal extensions

- Invariance concerns
  - Scale invariance in the 2D: SIFT
  - Camera motion: estimated using LSE and IR, and compensated
  - Motion information: only direction of each elementary motion of the track employed (rather than direction and magnitude of motion)
    - Invariance to image scale, since the same motion (e.g. a person picking up the phone) will result in different motion vector magnitudes depending on the camera focal length and distance from the plane of the motion
  - Motion direction histograms at different time scales and different granularity levels
    - Allow for partial matches when considering partial tracks and small variations in the direction of motion

Spatio-temporal extensions

- Experimentation: TRECVID HLFE
  - Goal: Encourage and benchmark visual concept detection systems
  - Benchmark video data collection (2007)
    - Netherlands Institute for Sound and Vision (~100 hours of news magazine, science news, news reports, documentaries, educational programming and archival video)
    - 50 hours annotated data for development
    - 50 hours for testing (also with ground truth annotation, at this point)
  - Evaluation using Average Precision

Spatio-temporal extensions



Conclusions

- Spatio-temporal extensions important for high-level feature extraction
  - Combined use of 2D descriptors and motion information improves results
  - Particularly for high-level features that have a strong temporal dimension
  - Tested in large set of realistic data (non-controlled subjects; camera motion;…)

Credits

- ITI colleagues
  - Anastasios Dimou
  - Panagiotis Sidiropoulos
  - Efi Tsamoura
  - Ioannis Kompatsiaris
- INESC-ID colleagues
  - Hugo Meinedo
  - Miguel Bugalho
  - Isabel Trancoso

References

- Shot segmentation

  - G. C. Chavez, M. Cord, S. Philip-Foliguet, F. Precioso, A. de A. Araujo, “Robust scene cut detection by supervised learning”, EUSIPCO 2006.
  - R. Zabih, J. Miller, K. Mai, “A feature-based algorithm for detecting and classifying production effects”, Multimedia Systems, vol. 7, no. 2, pp. 119–128, 1999.
  - J.U. Won, Y.S. Chung, I.S. Kim, J.G. Choi, K.H. Park, “Correlation based video-dissolve detection”, Proc. Int. Conf. on Information Technology Research and Education (ITRE2003), pp. 104 – 107, August 2003.
  - C.W. Su, H.Y.M. Liao, H.R. Tyan, K.C. Fan, L.H. Chen, “A motion-tolerant dissolve detection algorithm”, IEEE Transactions on Multimedia, vol. 7, no. 6, pp. 1106–1112, December 2005.
  - J. Bescos, G. Cisneros, J.M. Martinez, J.M. Menendez, J. Cabrera, “A unified model for techniques on video-shot transition detection”, IEEE Transactions on Multimedia, vol. 7, no. 2, pp. 293–307, April 2005.
  - X. Ling, L. Chao, L. Huanand, X. Zhang, “A general method for shot boundary detection”, International Conference on Multimedia and Ubiquitous Engineering (MUE08), pp. 394–397, April 2008.
  - E. Tsamoura, V. Mezaris, I. Kompatsiaris, "Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework", IEEE International Conference on Image Processing, Workshop on Multimedia Information Retrieval (ICIP-MIR 2008), San Diego, CA, USA, October 2008, pp. 45-48.

References

- Scene segmentation

  - A. Hanjalic, R. L. Lagendijk, “Automated high-level movie segmentation for advanced video-retrieval systems”, IEEE Trans. On Circuits and Systems for Video Technology, vol. 9, pp. 580–588, June 1999.
  - F. Beaver, “Dictionary of Film Terms”, Twayne Publishing, New York, 1994.
  - C. Petersohn, “Temporal video structuring for preservation and annotation of video content”, Proc. IEEE ICIP 2009.
  - M. Yeung, B.-L. Yeo, “Segmentation of video by clustering and graph analysis”, Computer Vision and Image Understanding, vol. 71, pp. 94–109, July 1998.
  - S.-C. Chen, M.-L. Shyu, W. Liao, C. Zhang, “Scene change detection by audio and video clues”, Proc. IEEE ICME, pp. 365-368, August 2002.
  - N. Goela, K. Wilson, F. Niu, A. Divakaran, “An svm framework for genre-independent scene change detection”, Proc. IEEE ICME, pp. 532-535, July 2007.
  - N. Nitanda, M. Haseyama, H. Kitajima, “Audio signal segmentation and classification for scene-cut detection”, IEEE ISCAS, pp. 4030-4033, May 2005.
  - J. Vendrig, M. Worring, “Systematic evaluation of logical story unit segmentation”, IEEE Trans. On Multimedia, vol. 4, no. 4, pp. 492-499, December 2002.
  - P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, I. Trancoso, "Multi-Modal Scene Segmentation Using Scene Transition Graphs", Proc. ACM International Conference on Multimedia (MM09), Beijing, China, October 2009, pp. 665-668.

References

- Shot representation and classification
  - S. Lazebnik, C. Schmid, J. Ponce,”Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”, Computer Vision and Pattern Recognition, 2006.
  - H. Bay, A. Ess, T. Tuytelaars, L. Van Gool,“Speeded-Up Robust Features (SURF)”, Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346--359, 2008
  - J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, J.M. Geusebroek,” Visual Word Ambiguity”, IEEE Transaction On Pattern Analysis and Machine Intelligence, 2010.
  - D. G. Lowe,”Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, 2004.
  - C.G.M. Snoek et. al., ”The MediaMill TRECVID 2008 Semantic Video Search Engine”, In Proc. TRECVID Workshop, 2008.
  - A. Anjulan, N. Canagarajah, ”A Unified Framework for Object Retrieval and Mining”, IEEE Transactions On Circuits and Systems for Video Technology, vol. 19, no. 1, 2009.
  - P. Dollar, V. Rabaud, G. Cottrell, S. Belongie,” Behavior Recognition via Sparse Spatio-Temporal Features”, Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.
  - N. Moenne-Loccoz, E. Bruno, and S. Marchand-Maillet, “Local Feature Trajectories for Efficient Event-Based Indexing of Video Sequences”, Proc. CIVR 2006.
  - I. Laptev, "On Space-Time Interest Points“, International Journal of Computer Vision, vol. 64, number 2/3, pp 107–123, 2005.

Thank you for your attention!

Questions?