



National Technical University of Athens  
Department of Electrical and Computer Engineering

# **Video Content Representation using Optimal Extraction of Frames and Scenes**

*Nikolaos D. Doulamis, Anastasios D. Doulamis,  
Yannis S. Avrithis and Stefanos D. Kollias*

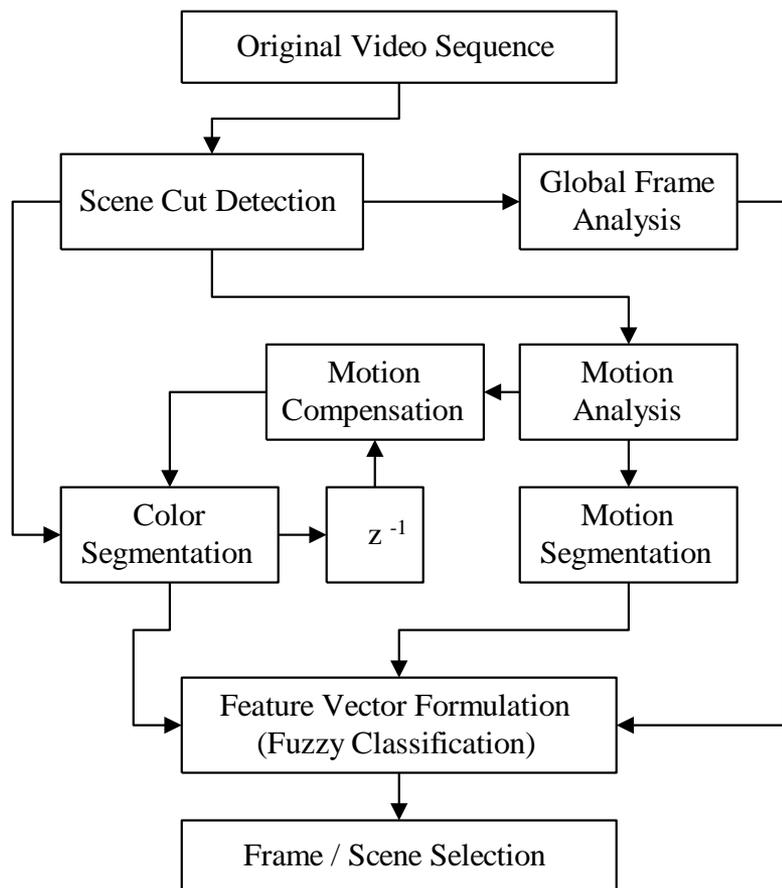
# Objective

- Automatic selection of a limited number of *key frames and scenes* from MPEG video streams
- Key frames and scenes provide sufficient information about the content of video sequences
- Representation of video sequences by multi-dimensional *feature vectors* of key frames and scenes, containing color & motion information
- *Video queries* applied directly on feature vectors of key frames and scenes

# Applications

- *Multimedia database management*: reduction of storage requirements for search capabilities, direct content-based retrieval, faster and more efficient video queries, improvement of user interface
- *Multimedia interactive services*: production of low resolution video clip previews (trailers) or still image mosaics, browsing of video databases on web pages

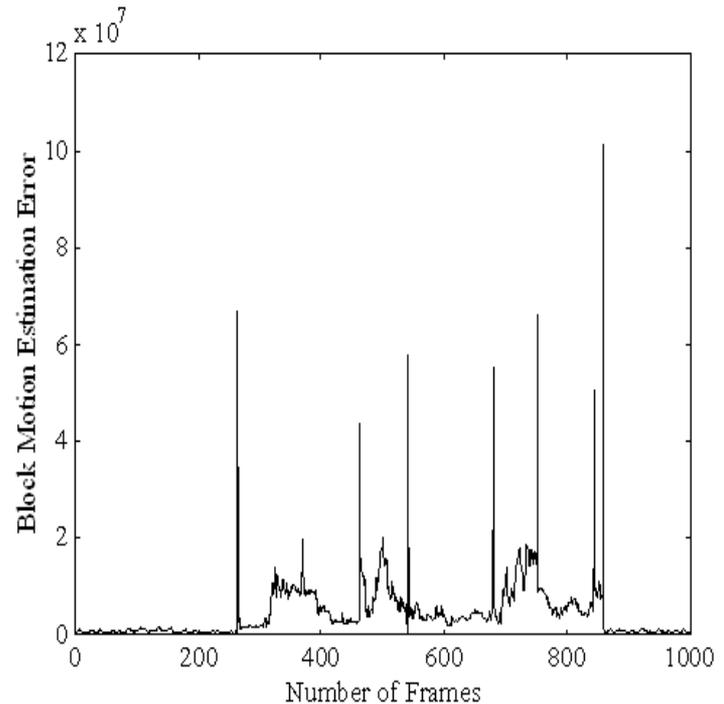
# Proposed System Architecture



- Scene cut detection
- Feature extraction for each frame
- Formulation of scene feature vectors
- Selection of the most representative scenes
- Extraction of key frames for each scene

# Scene Cut Detection

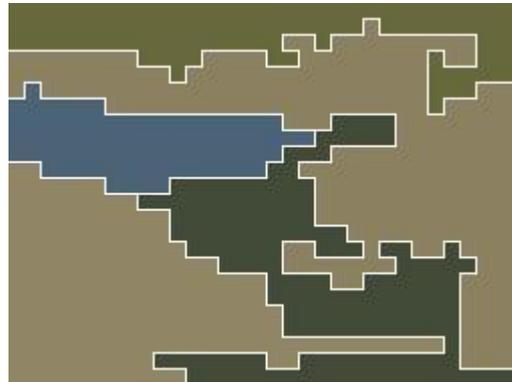
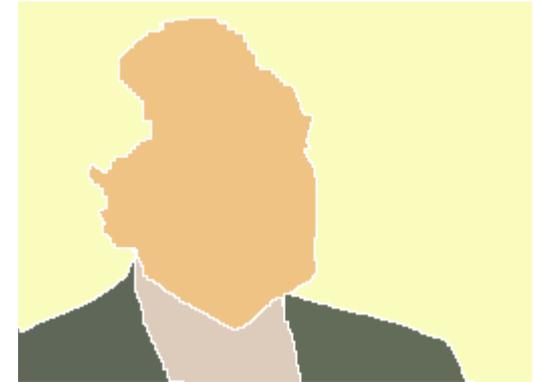
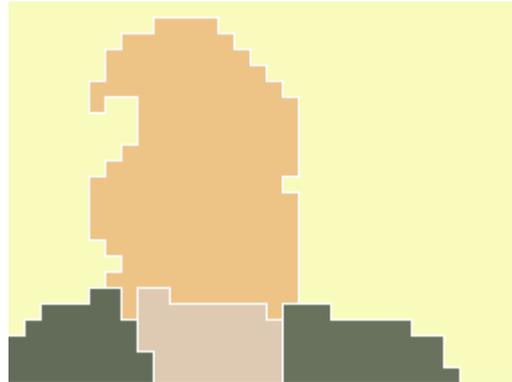
- Computation of the sum of the block motion estimation error
- Selection of frames for which sum exceeds a certain threshold
- Computations applied directly to MPEG-coded sequences



# Color Segmentation

- Segmentation according to *spatial homogeneity*
- *Block resolution* (reduction of computational time, exploitation of MPEG information)
- *Hierarchical merging* of similar segments (depending on color homogeneity & segment size)
- *Color features*: number of segments, location, size & mean color of each segment
- *Object tracking*: comparison with motion compensated segmentation results of previous frames (connected regions are encouraged to remain connected in successive frames)

# Color Segmentation Results

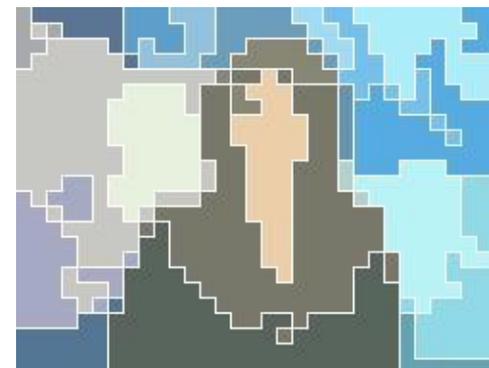
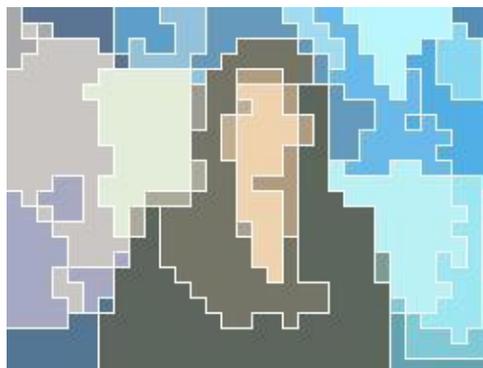
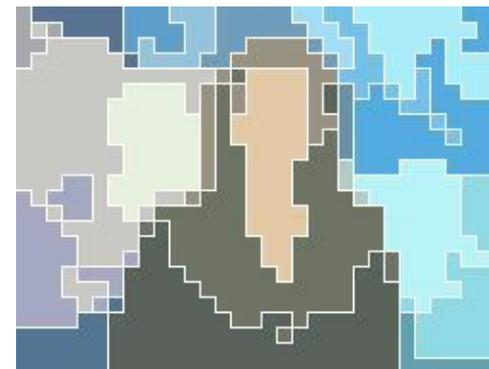
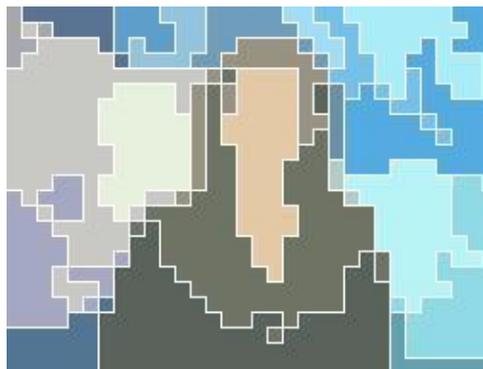


Two original frames

First stage of segmentation

Final result (full resolution)

# Object Tracking Capabilities



Two original  
successive frames

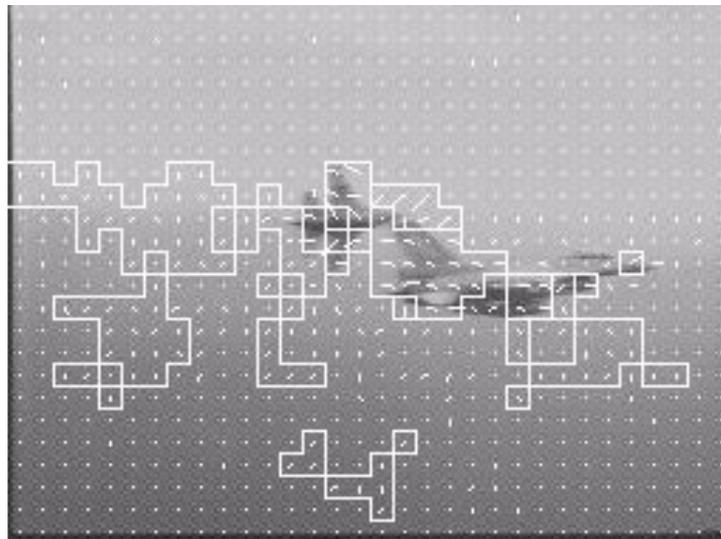
Segmentation  
without tracking

Segmentation  
with tracking

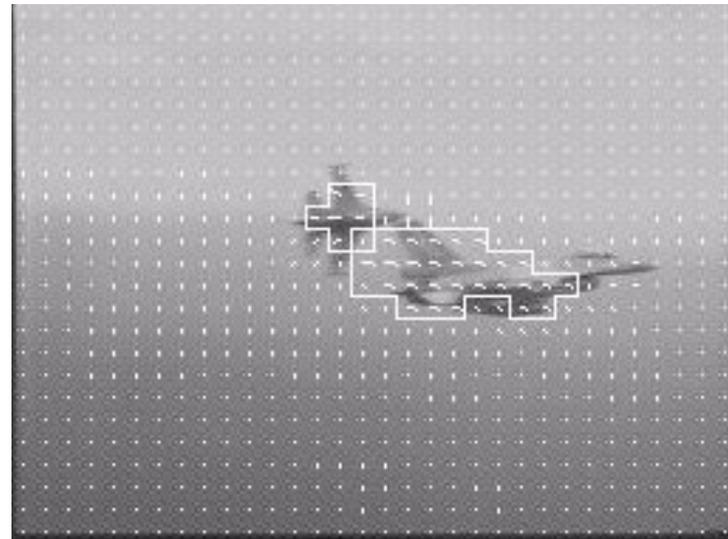
# Motion Segmentation

- Segmentation according to *spatial homogeneity*
- *Block resolution* (reduction of computational time)
- Motion vectors derived from motion analysis, or directly from MPEG stream
- *Median filtering* of derived motion vectors: elimination of “noise”, preservation of “edges”
- *Motion features*: number of segments, location, size & mean motion vector of each segment

# Motion Segmentation Results



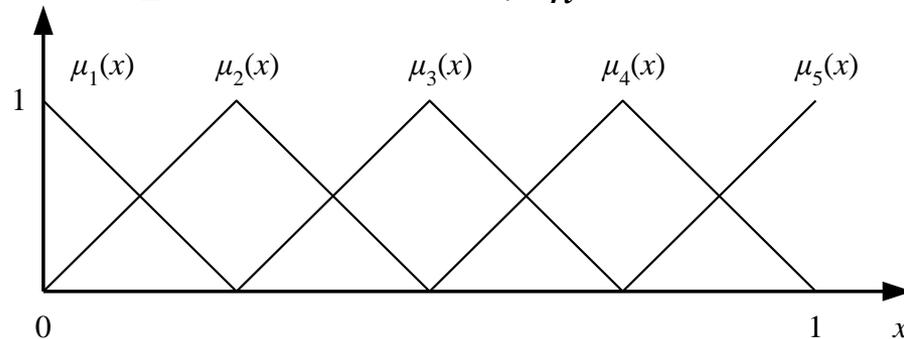
Motion segmentation  
without smoothing



Motion segmentation  
with smoothing

# Feature Vector Formulation

- *Multidimensional “histogram”*: classification of color and motion segments into pre-determined classes
- *Fuzzy classification*: normalization of each feature  $x$  to  $[0,1]$ , and partitioning into  $Q$  classes defined by membership functions  $\mu_n(x) \in [0,1]$ ,  $n = 1, \dots, Q$



- Histogram construction possible even with small number of samples  $x$ .

# Multidimensional Fuzzy Classification

- *Degree of membership* allocated to each class

$$F(n_1, \dots, n_L) = \sum_{i=1}^K \left\{ \prod_{j=1}^L \mu_{n_j}(f_j^{(i)}) \right\}$$

where  $n_j \in \{1, 2, \dots, Q\}$ : classification index for  $j$ th feature,  $Q$ : no. of partitions,  $L$ : no. of features,  $K$ : no. of segments,  $f_j^{(i)}$ :  $j$ th feature of  $i$ th segment,  $\mu_n(f)$ : degree of membership of feature  $f$  in partition  $n$

- Feature vector formed by degrees of membership for all  $M=Q^L$  combinations of  $n_1, \dots, n_L \in \{1, 2, \dots, Q\}$
- *Global frame characteristics* included in feature vector (color histogram, etc.)

# Scene Selection Mechanism

- *Scene feature vector* constructed based on frame feature vectors over duration of scene
- *Clustering* of similar scene feature vectors  $\mathbf{s}_i \in \mathfrak{R}^M$ ,  $i=1, \dots, N_S$  and selection of cluster *representatives*  $\mathbf{c}_i$ ,  $i=1, \dots, K_S$
- *Average distortion*  $D(\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{K_S}) = \sum_{i=1}^{K_S} \sum_{\mathbf{s} \in Z_i} d(\mathbf{s}, \mathbf{c}_i)$  should be minimized, where

$$Z_i = \{\mathbf{s} \in S : d(\mathbf{s}, \mathbf{c}_i) < d(\mathbf{s}, \mathbf{c}_j) \forall j \neq i\}$$

is the *influence zone* of  $\mathbf{c}_i$ :

$$(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \dots, \hat{\mathbf{c}}_{K_S}) = \arg \min_{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{K_S} \in \mathfrak{R}^M} D(\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{K_S})$$

# Optimization Method for Scene Selection

- *Generalized Lloyd* or *K-means* algorithm is used
- Arbitrary initial representatives  $\mathbf{c}_i(0)$ ,  $i=1,\dots,K_S$
- Minimization performed in iterative way:

$$Z_i(n) = \{\mathbf{s} \in S : d(\mathbf{s}, \mathbf{c}_i(n)) < d(\mathbf{s}, \mathbf{c}_j(n)) \forall j \neq i\}$$

$$\mathbf{c}_i(n+1) = \text{cent}(Z_i(n)) = \frac{1}{|Z_i(n)|} \sum_{\mathbf{s}_i \in Z_i(n)} \mathbf{s}_i$$

- $K_S$  most representative scenes selected as closest to optimal scene representatives:

$$\hat{\mathbf{s}}_i = \arg \min_{\mathbf{s} \in S} d(\mathbf{s}, \hat{\mathbf{c}}_i), \quad i = 1, 2, \dots, K_S$$

# Key Frame Selection Mechanism

- Minimization of a *correlation criterion*: key frames should not be similar to each other
- *Correlation measure* of feature vectors  $\mathbf{f}_i$ ,  $i = x_1, \dots, x_{K_F}$

$$R(\mathbf{x}) = R(x_1, \dots, x_{K_F}) = \left( \sum_{i=1}^{K_F-1} \sum_{j=i+1}^{K_F} (\rho_{x_i, x_j})^2 \right)^{1/2}$$

where  $\rho_{ij}$ : correlation coefficient of vectors  $\mathbf{f}_i$ ,  $\mathbf{f}_j$   
and  $\mathbf{x} = (x_1, \dots, x_{K_F}) \in W$ : index vector  
corresponding to a set of selected frame numbers

- Correlation measure  $R(\mathbf{x})$  should be minimized:

$$\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_{N_F}) = \arg \min_{\mathbf{x} \in W} R(\mathbf{x})$$

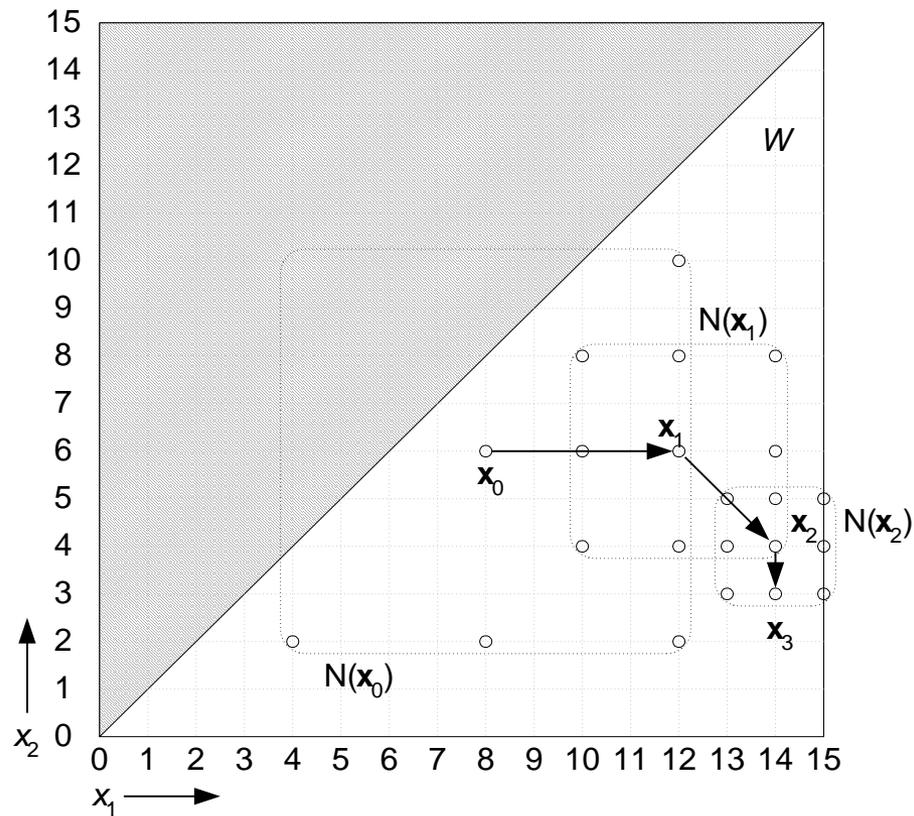
# Optimization Method for Scene Selection

- Exhaustive search is unfeasible: minimization is implemented by *logarithmic search algorithm*
- *Initial index*  $\mathbf{x}_0 = (\mu - [K_F / 2], \dots, \mu + [K_F / 2])$   
where  $\mu = 2^{L-1} - 1$  and  $N_F = 2^L$
- *Neighborhood* of index vector  $\mathbf{x}$ :

$$N(\mathbf{x}, S) = \{\mathbf{y} \in W : \mathbf{y} = \mathbf{x} + S\mathbf{p}, \mathbf{p} \in G^{K_F}\}, \quad G = \{-1, 0, 1\}$$

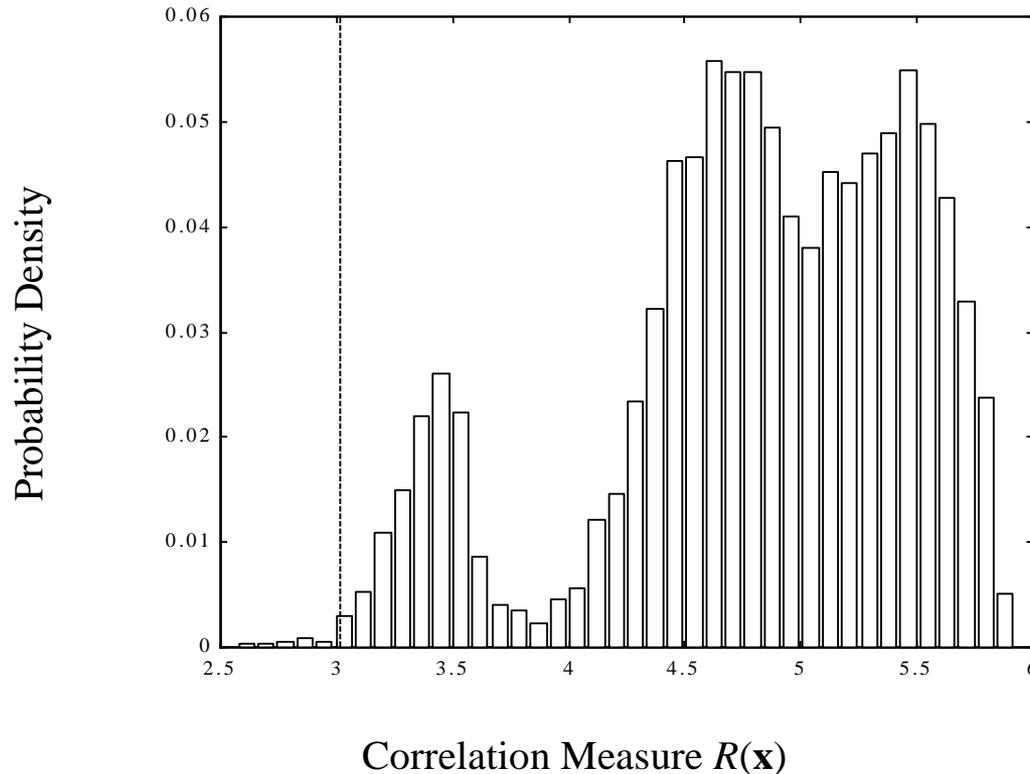
- Iteration steps:  $\mathbf{x}_n = \arg \min_{\mathbf{x} \in N(\mathbf{x}_{n-1}, S_{n-1})} R(\mathbf{x}), \quad S_n = S_{n-1} / 2$   
for  $n=1, \dots, L-2$
- Final Result:  $\hat{\mathbf{x}} = \mathbf{x}_{L-2}$
- Sub-optimal solution, but extremely fast

# Illustration of Logarithmic Search Algorithm ( $K_F=2$ )



# Logarithmic Search Performance

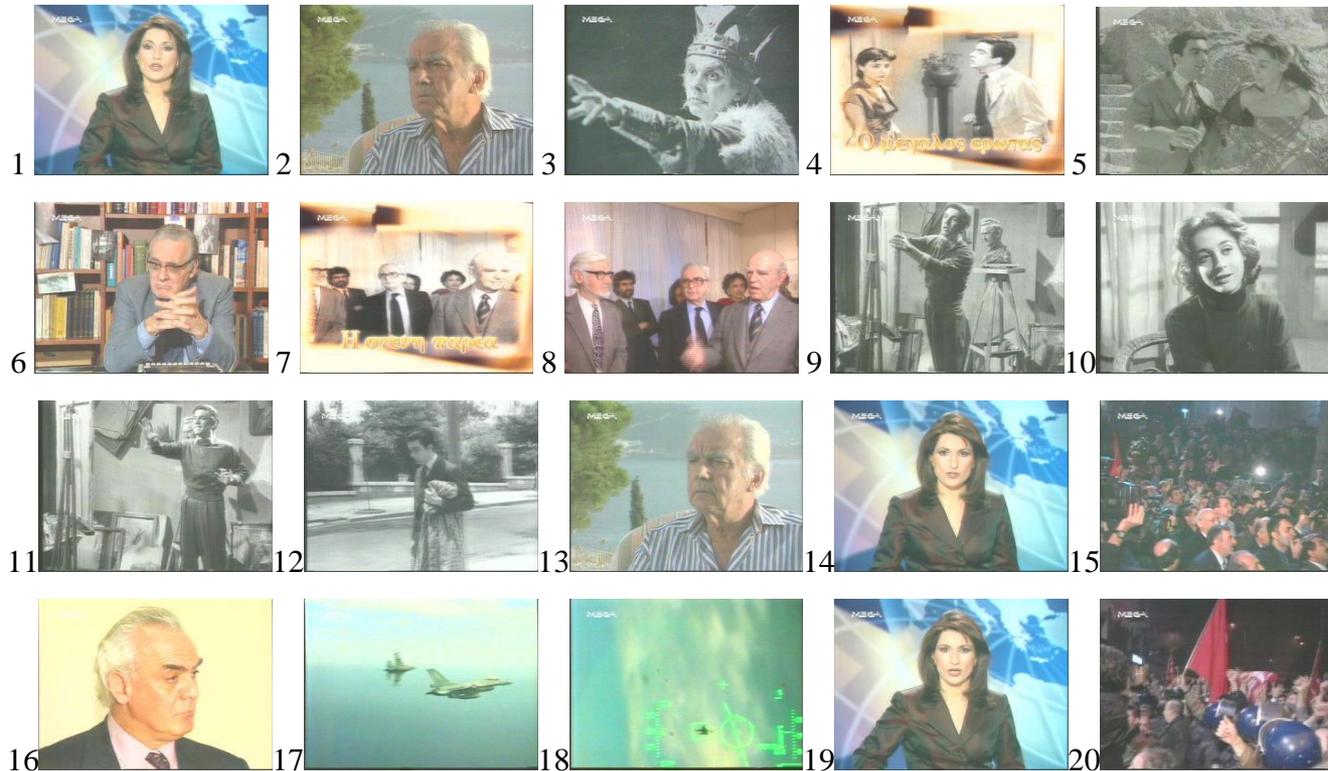
- Algorithm tested with large number of random index vectors, generating probability density function (histogram) of  $R(\mathbf{x})$ :



# Video Queries

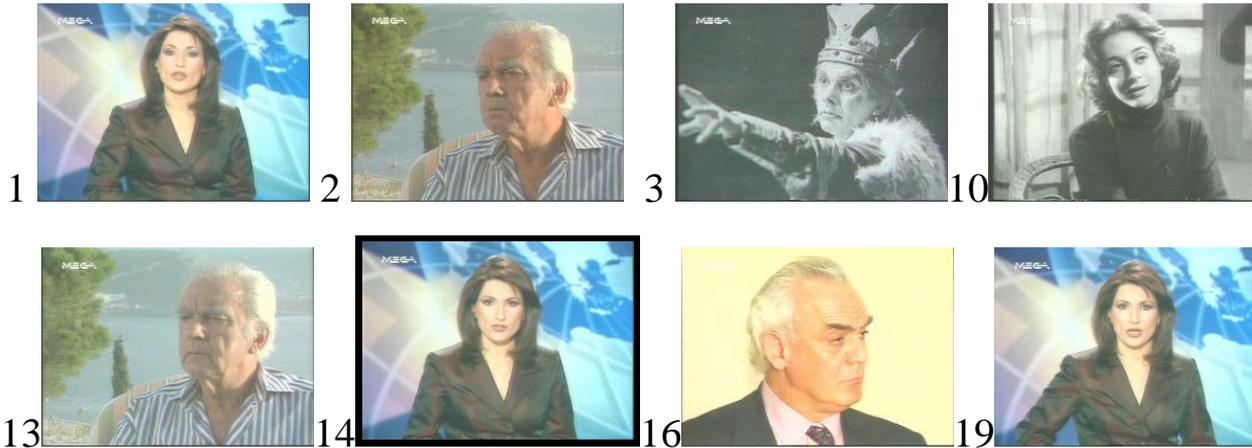
- Searching and retrieval of frames based on comparisons in the feature space
- Feature space contains all essential information, while preserving a very low dimension
- Comparisons performed on key frames only
- Dramatic reduction achieved in number of frames required for indexing, browsing or retrieval
- *Adaptive video queries* possible with parametric (weighted) distance function between feature vectors and parameter adaptation according to user requirements

# Test Video Sequence

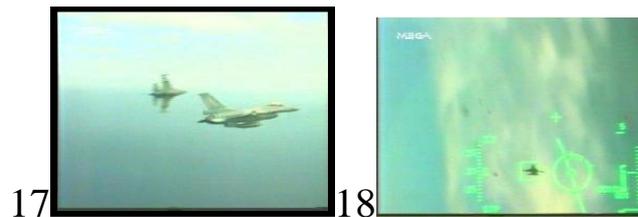


20 scenes of a 2.5 min (3750 frames) sequence  
from TV news reporting

# Scene Clustering / Selection Results



Cluster 1



Cluster 3

*Selected representative scenes shown in black border*

# Scene Clustering / Selection Results



Cluster 2



Cluster 4

*Selected representative scenes shown in black border*

# Key frame selection results



Four key frames from scene 12

# Conclusions

- *Automatic extraction* of key frames and scenes of video sequences taken from large video databases
- *Optimal selection*, based on a distortion criterion for key scenes and on a cross-correlation criterion for key frames
- *Object tracking* provides smoother feature vector trajectories and more robust frame selection
- *Direct implementation* on MPEG video streams
- *Feature vector space* enables robust and efficient frame comparisons, suitable for *video queries*

# Further Work

- Integration of color and motion segmentation results
- More robust object tracking
- More intelligent, semantic object extraction (e.g., human faces)
- Interweaving of audio and video information
- Adaptive video queries