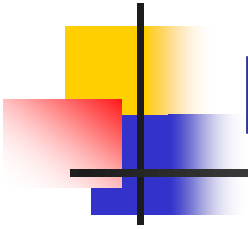
Broadcast News Parsing Using Visual Cues: A Robust Face Detection Approach

Yannis Avrithis, Nicolas Tsapatsoulis and Stefanos Kollias

Image, Video & Multimedia Systems Laboratory
Department of Electrical and Computer Engineering
National Technical University of Athens
Heroon Polytechniou 9, 157 73 Zographou, Greece

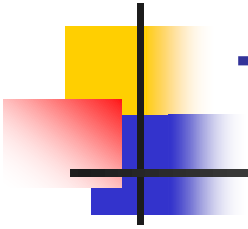


Problem Statement

- Temporal *segmentation* and *parsing* of news recordings based on *visual cues*
- Automatic content-based analysis and indexing tool for *summarization*, *browsing*, and *retrieval*
- Identification of typical news instances like *anchorpersons*, *reports* and *outdoor* shots
- Easy integration with existing systems employing *audio*, *textual* and *linguistic* cues (closed-caption tokens / teletext transcripts)

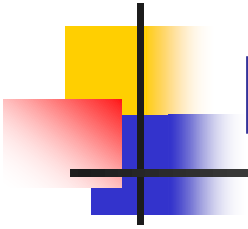
Why Digital News Archives?

- Organization of news recordings into *elementary story units*
- *Recurring appearance* of anchorpersons, reports / interviews and outdoor shots
- *Automatic parsing* valuable to data analysts in governmental and broadcast agencies, information / content providers, film studios and television / radio consumers
- *Applications* : interactive news navigation, retrieval and news-on-demand (NoD)



The Proposed Technique

- Robust *face detection* by means of color segmentation, skin color matching and shape processing
- Identification of *face close-ups* and extraction of *dominant faces*
- *Shot-change detection* based on motion-compensated frame differences
- *Shot classification* into anchors, reports / interviews, static images and outdoor shots



Face Detection

- *Color segmentation* employing the Multiresolution Recursive Shortest Spanning Tree (*M-RSST*) algorithm
- *Skin-tone color modeling* and *matching* using chrominance components of the *YCrCb* color model
- *Shape processing / filtering* using global shape features of face contours
- Fast implementation achieved with sufficient accuracy

Color Segmentation: M-RSST

- *Multiresolution decomposition* and construction of a truncated image pyramid
- All 4-connected region pairs assigned a *link weight* equal to the distance measure

$$d(X,Y) = \| \mathbf{c}_X - \mathbf{c}_Y \| \frac{a_X a_Y}{a_X + a_Y}$$

- *Recursive merging* of adjacent regions in each resolution level
- Fast algorithm, employed directly on MPEG streams with minimal decoding

Skin-Tone Color Matching

- Approximation of skin-tone color distribution with a *2-D Gaussian density function* on the *Cr-Cb* chrominance plane:

$$P(\mathbf{x} \mid \boldsymbol{\mu}_0, \mathbf{C}) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\}}{2\pi \cdot |\mathbf{C}|^{\frac{1}{2}}}$$

- *Skin-color region merging* based on estimated skin-color probability:

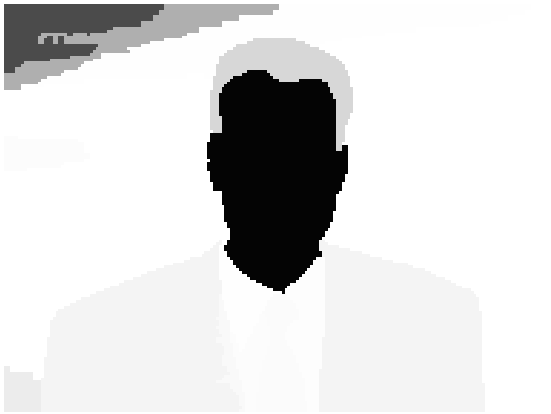
$$d_C(X,Y) = [\max(1-p_X, 1-p_Y)]^2$$

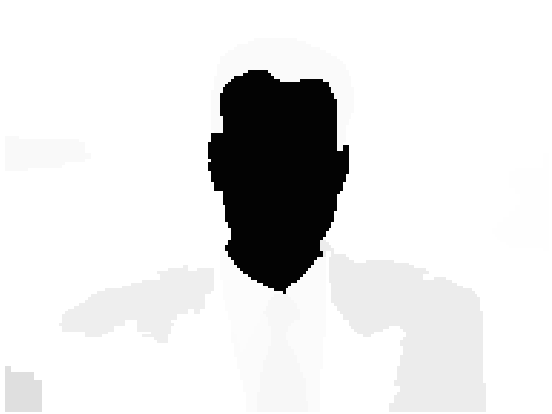
- Adjacent face segments merged – remaining partition map not affected

Shape Processing

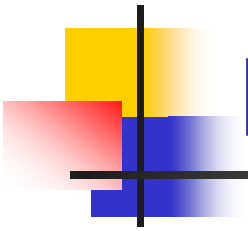
- *Global shape features* of segment contours
  - Shape *compactness* : $g_X = 4\pi\, a_X / r_X^2$
  - Shape *elongation* : $\ell_X = \sqrt{\lambda_2 / \lambda_1}$
- Both normalized in [0,1] and *invariant* to translation, scaling and rotation
- Combination with skin-color probability using non-linear functions – construction of an overall *face probability map*
- Segments with extremely irregular shape discarded

Face Detection Results



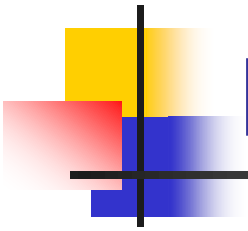


News Shot Classification (1)

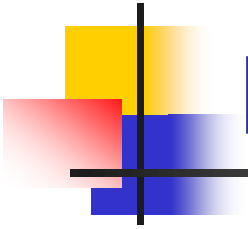


- *Shot change detection* using motion-compensated frame differences
- *Dominant face* detection based on face probability maps and facial segment size
- Temporal fluctuation of *facial segment features* & *background motion* employed for news shot classification
- Simple visual attributes (*color histograms*, *frame differences* and *motion*) used in conjunction with the derived face maps

News Shot Classification (2)



- *Shot clustering* according to background color histogram for anchor shot identification
- Shots classified into (i) *single / double anchor*, (ii) *reports / interviews*, (iii) *static images* and (iv) *outdoor shots*
- *Elementary story units* extracted between anchor shots
- Further grouping for semantic segmentation into true *news topics* requires audio and textual cues



Experiments

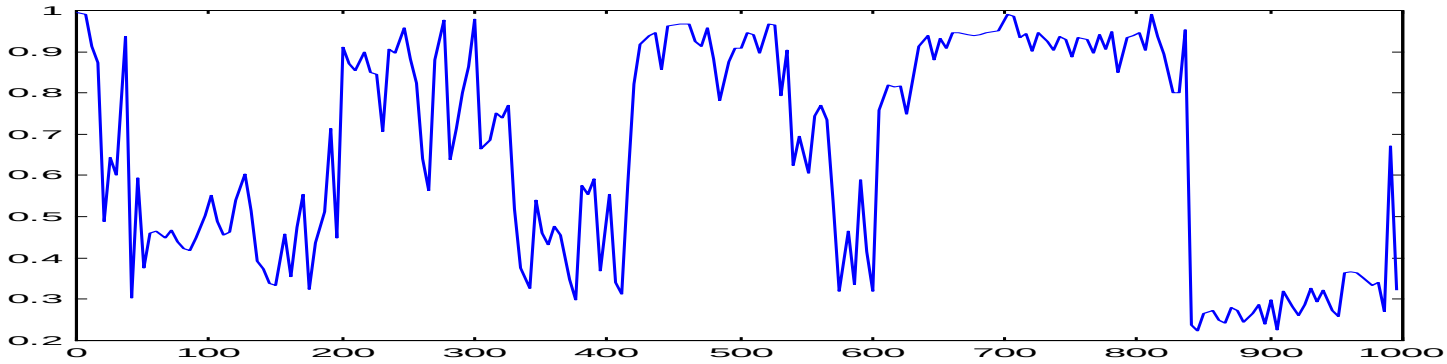
- Video database with news recordings of four Greek TV channels
- Six news broadcasts of 10 minutes at 10 fps and resolution 384×288×24bpp
- Temporal segmentation into shots, *manual shot classification & annotation* for evaluation purposes
- Performance evaluation by means of *precision / recall measurements*

Sample News Sequence

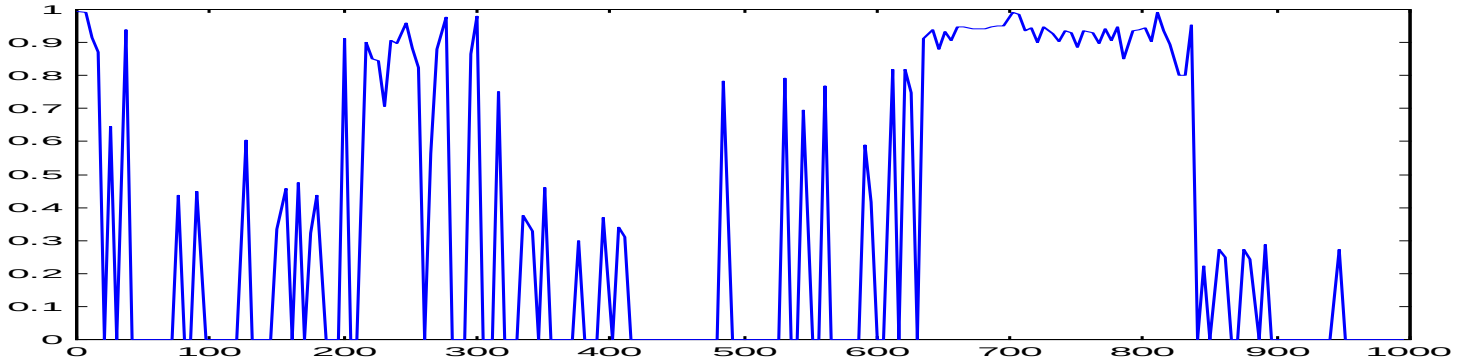
- 100 sec fragment with 1000 frames and 15 shots – 1 anchorperson shot

Results (1)

- Face probability (skin-tone color & shape)

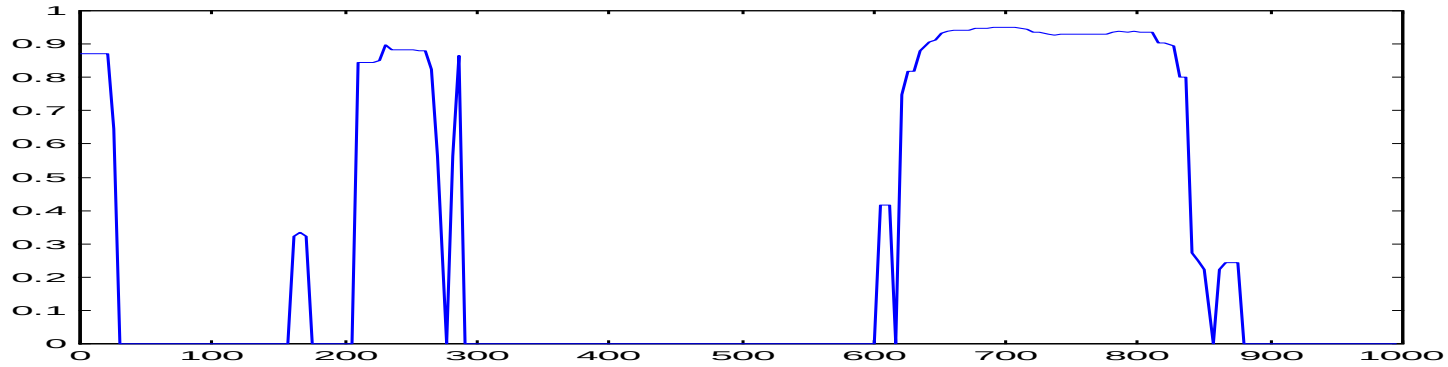


- Dominant face probability (segment size)

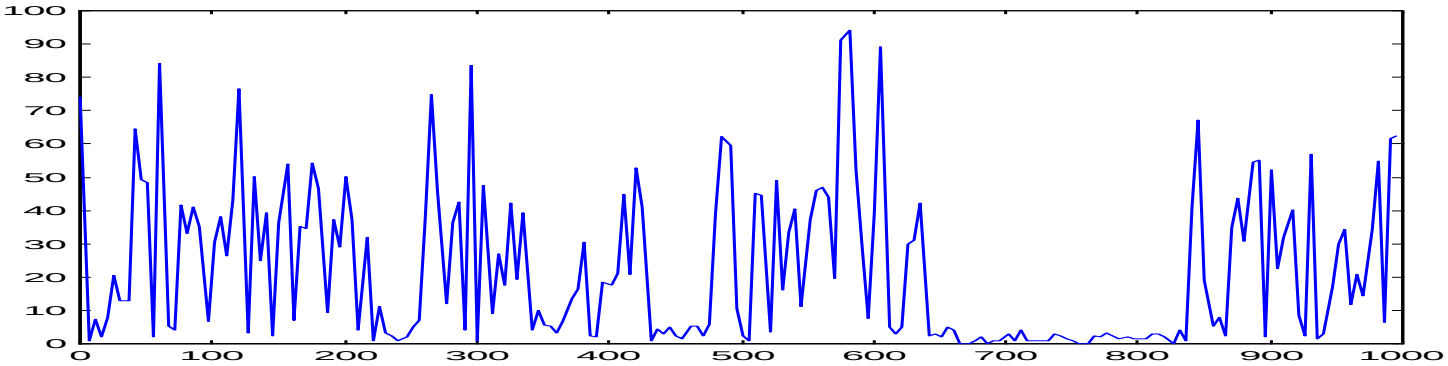


Results (2)

- Filtered probability curve



- Dominant face segment movement

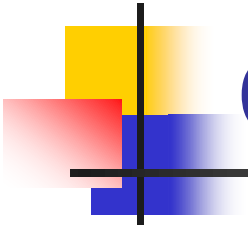


Precision / Recall Measurements

- All shots manually classified & annotated
- *Precision* : ratio of correctly aligned events to the total number of detected events (opposite of *false alarm* rate)
- *Recall* : ratio of correctly aligned events to the total number of true events (opposite of *dismissal* rate)
- *Event* : shot transition between two different shot classes, *correctly aligned* : within ± 2 frames of the corresponding true event

Precision / Recall Results

Experiment	Anchorperson		Report/ Interview		Static		Outdoor	
	P	R	P	R	P	R	P	R
A5 (a)	0.92	0.94	0.65	0.82	N/A	N/A	0.75	0.85
A5 (b)	0.95	0.65	0.83	0.94	050	1.00	0.73	0.87
ET-1	1.00	1.00	0.71	0.88	0.66	1.00	0.81	0.93
MEGA (a)	0.93	0.83	0.76	0.86	0.75	0.75	0.67	0.86
MEGA (b)	0.96	1.00	0.84	0.91	N/A	N/A	0.74	0.81
ANT1	0.93	0.71	0.77	0.88	0.75	0.66	0.85	0.86
Overall	0.95	0.93	0.76	0.88	0.67	0.85	0.76	0.86



Conclusions

- Efficient means for *temporal segmentation* and *indexing* of broadcast news programs using simple visual features
- *Reliable parsing* in the absence of other cues
- Semantic segmentation into true story segments requires *closed caption* transcripts or *audio information* (speaker identification)
- Either employed as *stand-alone application* or *integrated* with audio and textual cues of existing systems