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

Yannis Avrithis, Nicolas Tsapatsoulis and Stefanos Kollias

Department of Electrical and Computer Engineering
National Technical University of Athens
Heroon Polytechniou 9, 157 73 Zographou, Greece

e-mail: {ntsap,iavr}@image.ntua.gr

Abstract: *Automatic content-based analysis and indexing of broadcast news recordings or digitized news archives is becoming an important tool in the framework of many multimedia interactive services such as news summarization, browsing, retrieval and news-on-demand (NoD) applications. Existing approaches have achieved high performance in such applications but heavily rely on textual cues such as closed caption tokens and teletext transcripts. In this work we present an efficient technique for temporal segmentation and parsing of news recordings based on visual cues that can either be employed as stand-alone application for non-closed captioned broadcasts or integrated with audio and textual cues of existing systems. The technique involves robust face detection by means of color segmentation, skin color matching and shape processing, and is able to identify typical news instances like anchorpersons, reports and outdoor shots.*

1. Introduction

Due to the massive amount of multimedia data recently available from a multitude of sources like television broadcasts, films, surveillance videos, interactive web pages and digital video archives, there is an increasing need for new tools for automatic content-based analysis, parsing and indexing [1]. Such tools will be necessary for multimedia interactive services such as summarization, browsing and content-based retrieval, as reflected in the recent multimedia coding standards MPEG-4 and MPEG-7 [5].

Broadcast news and digital news archives are very important sources of multimedia data for two main reasons. First, both audio and video news resources are *unstructured* from the point of view of content like all audiovisual material [14]; however, the typical organization of news recordings into story segments involving recurring appearance of anchorpersons, reports and outdoor shots, allows easier *parsing*, i.e. temporal segmentation into elementary units and extraction of content information from low-level audiovisual features and semantic primitives. Second, fully automatic indexing and annotation of news recordings will be valuable to data analysts in governmental and broadcast agencies, information / content providers, film studios and television / radio consumers [7]. Costly manual approaches are currently employed for such purposes.

Meanwhile, several prototype systems have emerged allowing automatic or semi-automatic parsing and annotation of news recordings allowing interactive news navigation, content-based retrieval and news-on-demand (NoD) applications [3][6][8][14]. Most of them, however, heavily rely on *textual and linguistic cues* such as closed-caption tokens and teletext transcripts, while it is commonly agreed that *audio and visual cues* should play a more important role in the future [12] and artificial intelligence techniques will be required for integration of all content descriptors into semantic news segmentation. Currently audiovisual cues are limited to trivial processing like silence period or black screen detection [7] while cue integration has been tackled with dynamic programming [8], finite state machines [7] and hidden Markov models [4]. Moreover, the lack of textual information from a considerable number of television news broadcasts and mainly from historical news archives cannot be underestimated.

For the above reasons, we present an efficient technique for temporal segmentation and parsing of news recordings based on visual cues. Since a dominant portion of news video is occupied by human activities, human images and especially faces play an important role; thus, the proposed technique involves *face detection* by means of color segmentation, skin color matching and shape processing [10]. Shot detection in conjunction with some simple rules for combining dominant face properties (position, size and movement) with background motion information allows identification of typical news instances like *anchorpersons*, *reports* and *outdoor shots*. Hence, the proposed technique can either be employed as stand-alone application for non-closed captioned broadcasts or integrated with audio and textual cues of existing systems.

2. Face Detection

In the past the term *face detection* was strongly related with the face recognition task [9]; in order to achieve the required accuracy of detection, exhausting search procedures involving template matching, image invariants or low level features for the detection of local facial features like eyes, nose and mouth have been employed [11]. Fast implementations with sufficient accuracy have recently emerged for multimedia applications, mainly based on skin color modeling and matching through the use of the chrominance components of the *YCrCb* color model [13]. Enhanced performance with comparable computational complexity has been achieved in [10], based on color segmentation and shape processing apart from skin color matching. A similar approach is adopted in this work, as described in the sequel.

2.1. Color Segmentation

The Multiresolution Recursive Shortest Spanning Tree (M-RSST) algorithm, first introduced in [1], is our basis for color segmentation. It is a considerably fast algorithm and

can be employed for direct segmentation of MPEG video streams with minimal decoding. Initially a multiresolution decomposition of an input image is performed so that a hierarchy of frames is constructed, forming a truncated image pyramid. A partition of regions (segments) of size 1 pixel each is then created at the lowest resolution and links are generated for all 4-connected region pairs. Each link is assigned a weight equal to the distance

$$d(X,Y) = \| \mathbf{c}_X - \mathbf{c}_Y \| \frac{a_X a_Y}{a_X + a_Y} \quad (1)$$

for two adjacent regions X and Y, where, using the *YCrCb* color space, $\mathbf{c}_X = [\, Y_X,\ Cr_X,\ Cb_X \,]^T$ contains the average color components of region X and a_X is its area, i.e. the number of pixels within the region.

Region pairs are recursively merged in ascending order of the corresponding link weights; each boundary pixel of the resulting regions is then split into four new regions, and the "split-merge" procedure is repeated until the highest resolution image is reached. It has been observed in [10] that there still exist several cases – especially for large face segments – where even an optimal selection of the distance threshold cannot yield a single segment for the facial area without merging this segment with neighboring image areas. For this reason, a second step of segment merging is applied, based on skin-tone color distribution.

2.2. Skin –Tone Color Matching

In certain recent studies [13], efficient face detection has been achieved by exploiting the fact that skin-tone colors are spread over a small area of the *Cr-Cb* chrominance plane of the *YCrCb* color model. In our work, we approximated skin-tone color distribution using a two-dimensional Gaussian density function. Assuming that the mean vector $\boldsymbol{\mu}_0$ and the covariance matrix $\mathbf{C}$ are robustly estimated (e.g., through training from facial pixels of different races, obtained from TV news recordings), the likelihood of an input pattern $\mathbf{x}$ is given by:

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

Color segments represented by their average chrominance components can be assessed using the above model to give a probability of being skin segments. This *skin-color probability* is further exploited for segmentation. In particular, a small distance threshold is chosen for termination of the above segmentation procedure, and a second *skin-color merging* iteration is then applied, using the distance measure

$$d_C(X,Y) = [\max(1-p_X, 1-p_Y)]^2 \quad (3)$$

for link weights instead of $d(X,Y)$ given in (1), where p_X , p_Y are the *skin-color probabilities* associated to segments X and Y, respectively. Consequently all adjacent face segments are merged into a single segment while the remaining image partition map is not affected.

2.3. Shape Processing

Since objects unrelated to human faces but whose color chrominance components are still similar to those of the skin might be present in an image, object contour shape is also taken into account. Although the similarity of object shape to an ellipsis can be exploited through shape matching, for instance, object contours obtained through color segmentation are far from the ideal in most realistic cases; hence, only global shape features are used. The *compactness* of each shape is first obtained as

$$g_X = 4\pi \frac{a_X}{r_X^2} \quad (4)$$

where r_X denotes the perimeter (number of contour points) and a_X the area of segment X. The shape *elongation* (or aspect ratio) is then obtained through its Hotelling, or discrete Karhunen-Loeve transformation. Specifically, the 2×2 covariance matrix of the contour points and its eigenvalues are calculated, so that the corresponding shape elongation of segment X is given by

$$\ell_X = \sqrt{\lambda_2 / \lambda_1} \quad (5)$$

where λ_1, λ_2 are maximum and minimum eigenvalues, respectively. The above global shape features are fairly robust to segmentation noise and normalized in the interval [0,1]. They are also invariant to translation, scaling and rotation.

Experimental results have shown that typical values corresponding to face segments range from 0.44 to 0.79 for shape compactness and from 0.59 to 0.91 for elongation. Consequently, shape matching is achieved by transforming compactness and elongation with appropriate non-linear functions taking values in the range [0,1], similarly to fuzzy membership functions. Finally, the transformed compactness g'_X and elongation ℓ'_X are combined with the skin-color probability p_X using a weighted geometric mean, and an overall *face probability* is obtained, denoted as f_X. Appropriate weights are assigned so that shape features are in effect only used to discard face segments that possess extremely irregular shape although they match the skin-color probabilistic model.

3. News Shot Classification

Once a reliable face probability map is available for each video frame of a news sequence, it can be used for the detection of face close-ups, which are a good hint for the existence of anchorpersons and reporters or interviewed persons. In particular, the size of the detected face segments is first used to isolate large faces and discard smaller segments – that could correspond persons in the background, hands or unrelated objects. Moreover, research in face recognition has revealed that faces with resolution less than 32x32 pixels are not recognizable. Based on this inference we pose a threshold for discarding face segments that capture less than 3% of the image area. Usually one, two or rarely three *dominant faces* are retained this way. Fuzzy membership functions are again employed on dominant face location and size to classify frame image as either containing none, one or two face close-ups.

Temporal fluctuation of these properties as well as background motion are then taken into account to give a finer classification: (*i*) one or two face close-ups with a static background are classified as *single* or *double anchor*,

(*ii*) one or two face close-ups with a moving background are classified as *reports / interviews*, (iii) fixed backgrounds with no dominant faces are classified as *static images* (e.g. financial reports or weather forecasts) and (iv) finally, other cases with significant motion and small or no faces at all are classified as *outdoor shots*. Background motion is estimated by absolute frame differences; no motion compensation is required (as in shot cut detection) since fixed backgrounds are only of interest, mainly to distinguish between studio and outdoor shots.

Temporal segmentation of news recordings into elementary units is achieved by *shot change detection* and *shot classification* using the above criteria. Thresholding of motion-compensated frame differences is employed for shot change detection; this is a simple but fast approach for detecting shot cuts. More complex techniques for detecting shot transitions like zoom, wipe and dissolve effects can also be applied [2]. Shot classification is performed by estimating the class (anchor, report, static or outdoors) that best describes a whole shot, i.e. the majority of frames within the shot.

Since face close-ups with fixed background can occur in circumstances other than that of the main anchorperson (e.g., a reporter in an indoors location), anchor shots are further filtered. Specifically, they are clustered according to the color histogram of their background, and the single cluster with the most members – corresponding to the shot with the most occurrences – is selected. Anchor shots are thus limited and the remaining shots between successive anchor shots are grouped into *elementary story units*. Of course, further unit grouping is necessary for higher-level, semantic segmentation into true *news topics*, but such grouping would also require audio and textual cues.

4. Experimental Results

The visual content used in our experiments included a database video created from news recordings of Greek TV channels, namely A5, ET1, MEGA and ANT1. Six news broadcasts of duration 10 minutes each were recorded at 10 frames per second with a resolution of 384×288×24bpp. First, face detection is demonstrated in Figure 1 for a typical anchorperson image. It can be observed that with the proposed integration of color segmentation, skin color matching, shape processing and size features, an accurate face probability map is obtained, giving exactly the dominant face segment of the anchorperson.

Temporal segmentation is illustrated for a sample fragment of the MEGA sequence, shown in Figure 2. The fragment, of total duration 100 seconds, contains 15 shots, one of which corresponds to the anchorperson. Figure 3(a) illustrates the maximum probability of the face probability map obtained from each frame, before taking segment size into account. Similarly, Figure 3(b) presents the *dominant face* probability curve after eliminating small face segments. It can be seen that certain shots containing small segments with high face probability are now discarded; otherwise they would introduce false alarms. A median filter is then applied to the dominant face probability curve, as depicted in Figure 3(c). Filtering is necessary to alleviate face shots with very short duration. For example an anchorperson shot rarely has duration less than few seconds. Appropriate choice of the filter's window can be applied to account for the required shot duration. It is deducted that there are three possible face shots. The first does not correspond to a real face shot, giving a *false alarm*. This happens when a segment with shape similar to that of a face, fits the skin-color model. The second is a true face shot, while the third actually corresponds to two successive shots containing the anchorperson and an interviewed person. The above false alarm case is alleviated by taking into account the dominant face motion, as shown in Figure 3(d). Real face segments present small movement while oscillations appear in many cases of outdoor shots that are misclassified as face shots. On the other hand, the two successive face shots have already been split by shot change detection; they could also be separated employing simple criteria such as color histogram variation, depicted in Figure 3(e).

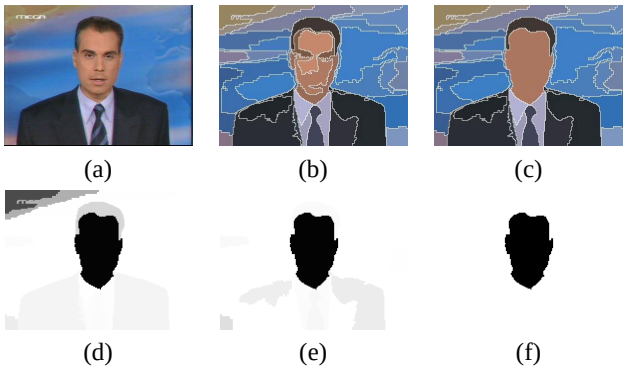

Figure 1: Face detection. (a) Original frame, (b) and (c) color segmentation before and after skin color merging respectively, (d) skin-color probability map, (e) face probability map and (f) dominant face segment.

Figure 2: A sample fragment of the MEGA sequence, containing 1000 frames (100 sec) and 15 shots, one of which corresponding to the anchorperson.

The performance of the proposed temporal segmentation technique has been evaluated in terms of precision and recall measurements. All news recordings have been segmented using shot cut detection and shots have been manually classified and annotated. Automatically extracted events can thus be compared to manually annotated ('true') ones. Similarly to [7], *precision* is defined as the ratio of correctly aligned events to the total number of detected events (i.e., the opposite of *false alarm* rate), while *recall* as the ratio of correctly aligned events to the total number of true events (i.e., the opposite of *dismissal* rate). An event is defined as a shot transition between two different shot classes; an event is correctly aligned if it occurs within ±2 frames of the corresponding true event.

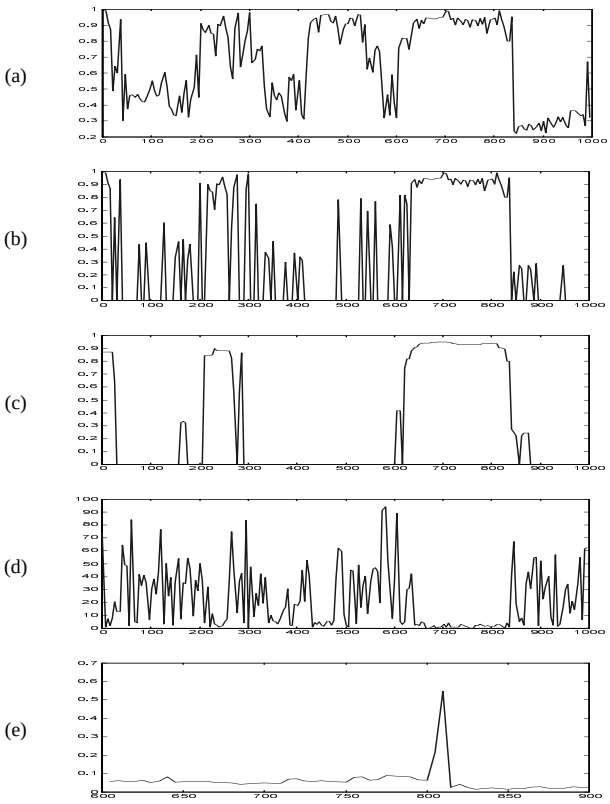

Figure 3: (a) Face probability curve vs. frame number (0-1000), (b) Dominant face probability curve (after removing smaller segments), (c) filtered curve, (d) dominant face segment movement and (e) histogram variation for frames 600-900.

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	1.00	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	0.93	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	0.94	0.77	0.88	0.75	0.66	0.85	0.86
Overall	0.95	0.97	0.76	0.88	0.67	0.85	0.76	0.86

Table I: Precision and recall measurements for shot classification of the six test news sequences.

Anchorperson shots have the best classification rates. This is expected, as this shot class is based on clustering apart from face detection. The classification rates for report/interview shots are smaller mainly due to the uncontrolled illumination conditions, which can interfere with the skin-color model. Report/interview shots are usually misclassified as outdoor shots and vice versa. Due to the limited number of static shots no reliable conclusions can be made.

5. Conclusion

The proposed technique provides with an efficient means for temporal segmentation and indexing of broadcast news using visual cues. Although semantic segmentation into true story segments would also require closed caption transcripts or audio information (e.g. speaker identification) properly combined using artificial intelligence techniques, the performance of the employed face detection procedure allows reliable parsing even in the absence of other cues. Simple visual attributes such as color histograms, frame differences and motion have been used in conjunction with the derived face maps, yielding promising results. The proposed technique can either be employed as stand-alone application for non-closed captioned broadcasts or integrated with audio and textual cues of existing systems.

6. References

[1] Y. Avrithis, A. Doulamis, N. Doulamis and S. Kollias, "A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases," *Computer Vision and Image Understanding* **75** (1/2), pp. 3-24, July 1999.

[2] P. Bouthemy, M. Gelgon and F. Ganancia, "A Unified Approach to Shot Change Detection and Camera Motion Characterization," *IEEE Trans. CSVT* **9** (7), pp.1030-1044, Oct. 1999.

[3] M. Brown, J. Foote, G. Jones, K. Sparck-Jones and S. Young, " Automatic Content-Based Retrieval of Broadcast News," *Proc. of ACM Multimedia Conference*, San Francisco, CA, Nov. 1995.

[4] S. Eickeler, A. Kosmala and G. Rigoll, "A New Approach to Content-Based Video Indexing Using Hidden Markov Models," *Proc. of WIAMIS*, Belgium, June 1997.

[5] ISO/IEC JTC1/SC29/WG11, "MPEG-7: Context and Objectives (v.5)," Doc. N1920, 1997.

[6] B. Merialdo, "Automatic Indexing of TV News," *Proc. of WIAMIS*, Belgium, June 1997.

[7] A. Merlino, D. Morey and M. Maybury, "Broadcast News Navigation Using Story Segments," *Proc. of ACM Multimedia Conference*, Seattle, WA, Nov. 1997.

[8] Y. Nakamura and T. Kanade, "Semantic Analysis for Video Contents Extraction – Spotting by Association in News Video," *Proc. of ACM Multimedia Conference*, Seattle, WA, Nov. 1997.

[9] A. Samal and P.A. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey," *Pattern Recognition*, **25** (1), pp. 65-77, 1992.

[10] N. Tsapatsoulis, Y. Avrithis and S. Kollias, "Efficient Face Detection for Multimedia Applications," *Proc. of ICIP 2000* (submitted for publication).

[11] N. Tsapatsoulis, N. Doulamis, A. Doulamis, and S. Kollias "Face Extraction from Non-uniform Background and Recognition in Compressed Domain," *Proc. of ICASSP'98*, Seattle WA, May 1998.

[12] S. Tsekeridou and I. Pitas, "Audio-Visual Content Analysis for Cont-Based Video Indexing," *Proc. of Int. Conf. on Multimedia Computing and Systems*, Florence, Italy, June 1999.

[13] H. Wang and S.-F Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video," *IEEE Trans. CSVT* **7** (4), August 1997.

[14] H.J. Zhang, S.Y. Tan, S. Smoliar and G. Yihong, "Automatic Parsing and Indexing of News Video," *Multimedia Systems* **2**, pp. 256-266, 1995.