AUDIOVISUAL ANCHORPERSON DETECTION FOR TOPIC-ORIENTED NAVIGATION IN BROADCAST NEWS

Martin Haller[1], *Hyoung-Gook Kim*[2], *Thomas Sikora*[1]

[1]Technical University of Berlin
Department of Communication Systems
EN 1, Einsteinufer 17, 10587 Berlin, Germany
{haller,sikora}@nue.tu-berlin.de

[2]Samsung Advanced Institute of Technology
Mt. 14-1, Nongseo-Ri, Giheung-Eup,
Yongin-Si, Gyeonggi-Do, Korea 449-712
hyounggook.kim@samsung.com

ABSTRACT

This paper presents a content-based audiovisual video analysis technique for anchorperson detection in broadcast news. For topic-oriented navigation in newscasts, a segmentation of the topic boundaries is needed. As the anchorperson gives a strong indication for such boundaries, the presented technique automatically determines that high-level information for video indexing from MPEG-2 videos and stores the results in an MPEG-7 conform format. The multimodal analysis process is carried out separately in the auditory and visual modality, and the decision fusion forms the final anchorperson segments.

1. INTRODUCTION

Video streams need to be indexed before they can be retrieved with content-based queries. To this end, methods for automatic content-based analysis [1, 2, 3] are needed. The automatic extraction of high-level semantic information can largely avoid manual annotation. Availability of multimedia metadata will become more and more important. Not only retrieval and browsing depend on those metadata. They are also a prerequisite for the emerging universal multimedia access and for the Semantic Web.

Anchorperson (AP) detection is an important indexing task for broadcast news. For instance, AP segments are useful for fast topic-oriented navigation within news by a viewer or for further content-based analysis. Several approaches have been suggested in the literature [3]. Many techniques are using visual analysis, but more and more techniques also incorporate other modalities to enhance the detection performance.

Ide et al. [4] proposed to detect the AP based on color histogram distances in the visual domain. After shot detection, key-frames are extracted. Key-frames with frontal faces accompanied with lip movement are considered in the clustering of color histogram distances. Then the largest and most dense cluster is chosen as AP cluster. Qi et al. [5] suggested an audiovisual technique for AP detection. They used speaker change point detection and speaker clustering in the audio domain and key-frame clustering in the visual domain. The AP clusters are chosen with two assumptions about the temporal structure independently in each domain. The segments should have the highest proportion and the distribution should be more disperse than the others. The independent results are combined with an AND operation. An approach for multilevel AP detection is proposed by Lan et al. [6]. First, candidates are identified based on visual features with SVM classification. After feature extraction in the auditory and visual modality, a multimodal associated clustering is performed. Visual features are extracted from key-frames and detected face regions. Then false alarms are removed. The cluster of the main AP is chosen from the large clusters where the elements are distributed over the whole news broadcast.

In this paper, we propose a technique to detect the main AP with the separate analysis of the auditory and visual modality. The overall results are formed with decision fusion (Fig. 1). Our technique combines existing approaches to achieve results that are computationally efficient and highly accurate. For this purpose, we apply speaker segmentation to the audio signal and detect frontal faces in the video signal. Clean speech segments and frontal faces are used for cluster analysis. The AP cluster is identified among the other clusters for each modality. To this end, our technique uses modified cluster selection criteria. The demonstration program reads MPEG-2 video streams and stores the results of the content analysis in an MPEG-7 XML document.

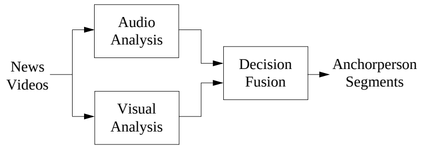

Fig. 1. Separated audiovisual analysis with decision fusion

The paper is organized as follows. The proposed technique is described in section 2. Experimental results are presented in section 3, which is followed by the conclusions and further work.

2. ANCHORPERSON DETECTION SYSTEM

2.1. Audio analysis

The steps of the audio analysis are shown in Fig. 2.

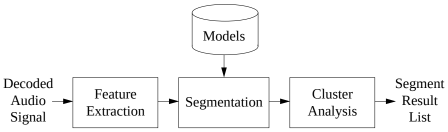

Fig. 2. Block diagram of audio analysis

The input of the *feature extraction* module is the decoded audio signal from an MPEG-2 stream. This module calculates one 13 dimensional feature vector every 10 ms. The vector consists of the logarithmic frame energy and the first 12 mel-frequency cepstral coefficients (MFCCs). These features have been widely used in audio analysis [7].

The *segmentation* module uses these features and performs a maximum likelihood (ML) classification of half a second sub-segments. Here, the segmentation distinguishes between two categories. The first category is the speech without noise or other environmental sounds in the background. The second category is the noisy speech or other possible audio signals included in broadcast news. The temporal segments with clean speech have a direct relation with the well-defined recording settings in a studio environment. Therefore, these segments are referred to as studio segments. As we expect that the AP is only included in these studio segments, the further analysis process regards only these segments. Hence, only sub-segments classified as the category studio are used to build smoothed studio segments and all other sub-segments are ignored. The ML-classification uses Gaussian Mixture Models (GMMs) [8]. As it is also possible that in one studio segment different speakers are included, pause detection determines additional potential segment boundaries. These are possible speaker change points.

The *cluster analysis* is performed in three steps. At first, the distances between the segments are computed. Afterwards, a Hierarchical Cluster Analysis (HCA) is carried out. Finally, the AP cluster is selected.

The low-level features from the feature extraction are used to determine the distances between studio segments. Inspired from the speaker segmentation [9], Kullback-Leibler divergence (KL2), Generalized Likelihood Ratio (GLR), and the Bayesian Information Criterion (BIC) are compared as distance measures. The single, complete and average linkage methods were compared for the HCA. As it was examined by experiments with the test set of videos, all three distances reached the same results in terms of precision and recall in combination with the average linkage method. The other cluster methods achieved less accurate detection results. As the BIC and KL2 measures are slightly more sensitive to the cutting threshold, our technique uses only GLR and average linkage for the cluster analysis of the studio segments with a global common threshold for dendrogram cutting. The HCA results in M_a clusters $\mathcal{A}_k$, where index a symbolizes the audio domain and the index k identifies the k-th cluster.

These clusters of studio segments are used as input for the cluster selection, where the AP cluster will be chosen. Assuming that each cluster includes segments with similar speech utterances from one speaker, the AP is identified by the verification of three cluster selection criteria for the temporal structure. First, the person has the appearance with the longest duration as he/she dominates the broadcast. Second, the appearance of the person is close to the center of the broadcast as the person introduces topics repeatedly throughout the broadcast. Third, the person has the highest variance of time points as he/she is usually present at the start and the end of the broadcast. Thus, the mean $\mu_{\mathrm{a},k}$ and the variance $\sigma^2_{\mathrm{a},k}$ of the time points of frames are determined for each cluster. The mean value is used to compute the closeness $c_{\mathrm{a},k}$ to the center t_μ of the video with

$$c_{\mathrm{a},k} \quad = \quad 1 - |t_\mu - \mu_{\mathrm{a},k}| \,/\, t_\mu \quad . \tag{1}$$

$c_{\mathrm{a},k}$ lying in the range $[0, 1]$. With the center closeness, the variance and the total frame number $N_{\mathrm{a},k}$ for the studio segments contained in the k-th cluster, the vector

$$\mathbf{a}_k^\mathrm{T} = \left(\; c_{\mathrm{a},k} \; , \; \sigma^2_{\mathrm{a},k} / \max_{\forall k} \left(\sigma^2_{\mathrm{a},k} \right) \; , \; N_{\mathrm{a},k} / \max_{\forall k} \left(N_{\mathrm{a},k} \right) \; \right) \tag{2}$$

can be defined for each cluster. The l-th cluster is chosen, if it maximizes the Euclidean norm (L_2) of all obtained vectors, so that

$$l = \operatorname*{arg\,max}_{1 \le k \le M_\mathrm{a}} \|\mathbf{a}_k\|_2 \tag{3}$$

identifies the index of the AP cluster.

Once the AP cluster is chosen, a post-processing removes the segment boundaries introduced by pause detection between AP segments by merging such segments with a distance below 6 seconds.

At the end of the audio analysis, a result list of start and end time points of AP segments found in the auditory modality is available for the decision fusion.

2.2. Visual analysis

All main steps of the analysis of the visual modality for AP detection are depicted in Fig. 3.

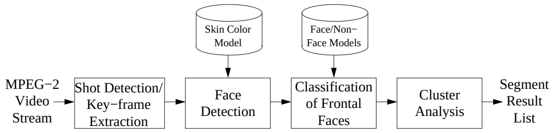

Fig. 3. Block diagram of visual analysis

With an MPEG-2 video stream as input signal, the *shot boundary detection* along with *key-frame extraction* finds the shots and reduces them to key-frame images. The distance between color histograms in the RGB color space are used for the detection of shot boundaries [10]. For a good computational performance, only I-frames with their DC coefficients from the compressed domain of the MPEG stream are considered for color histogram computation. As the I-frames appear at least twice per second for terrestrial DTV broadcasting, the temporal resolution is accurate enough for AP detection. For finding the AP segments, only hard-cut shots are considered as the predominant shot transition type in broadcast news. Furthermore, a spatial portioning scheme with four horizontal slices is used to detect shot boundaries independently in certain regions. An adaptive threshold decides about the shot boundaries. The threshold as weighted average value depends on L consecutive color histogram distances, from which $L/2-1$ values come from future distances. For every interval between two shot boundaries, a key-frame in the middle of the interval is extracted.

The *face detection* is carried out with a resolution of 180 × 144 pixels by performing a pixel-based segmentation of skin-color regions and a rule-based validation of segments as face candidates. The color segmentation takes place in the HSV color space with saturation and shifted hue only using a trained GMM instead of a predefined rectangular area of HS-values. A hard threshold is used to identify the skin-color regions in the key-frame images. A validation assures that the width, the height, the area and the aspect ratio of the face candidate images are lying in a certain range. These normalized face candidate images are passed to the frontal face classifier.

For the *classification of frontal faces*, different features, transformations and classification methods were compared. Combinations of MPEG-7 texture features of the homogeneous texture descriptor (HTD) and of the edge histogram descriptor (EHD) were used. The transformations PCA or ICA were applied as dimension reduction methods. The model-based classification was done using GMMs as well as Support Vector Machines (SVMs). Experiments were performed with the aim to find a suitable combination. The intensity-invariant version of the HTD combined with the global EHD values as features, the PCA for reducing the dimensionality to 15 and a bimodal GMM was found to perform best. Our proposed technique uses this combination. Here, it is important to have a high recall for detection of frontal faces and it is not critical to have false positives included in the classification results. Thus, the classifier is biased to decide rather for frontal faces.

The *cluster analysis* in the visual domain uses the frontal face images and the time properties of corresponding shots. The clustering is again performed in three steps. The distances between all frontal faces are computed at first. Subsequently, a hierarchical clustering is performed with these distance values. The AP cluster of the visual domain is selected in dependence on temporal properties of each cluster.

High dimensional features are used for the computation of the distances between the frontal face images. The feature vector for each image has a total dimensionality of 212 and consists of 80 local, 5 global, and 65 semi-global EHD features combined with 62 HTD features. With that combination of texture features, three cluster methods (single, complete and average linkage) were verified with the city-block and a fractional distance measure. The fractional distance metric [11] is derived from the Minkowski metric with the fractional order 1/2 and is used for the computation of distance values between the images. The fractional order increases the contrast of the distances between high-dimensional feature vectors in comparison with the city-block distance and hence makes the features of the faces more distinguishable. So the fractional distance together with the average linkage method are used for creating a set of clusters by using dendrogram cutting with a common global threshold.

Selecting the AP cluster from the set of clusters in the visual domain follows the same idea as already used in the audio analysis. Quantifying the temporal structure of the formed clusters, the following measures are used. With M_{v} clusters, the total time $t_{\mathrm{v},k}$ of cluster $\mathcal{V}_k$ is calculated by the summation of duration of shots included in the respective cluster. The index v denotes the visual domain and k is the cluster index. The mean $\mu_{\mathrm{v},k}$ and the variance $\sigma^2_{\mathrm{v},k}$ of the key-frame time points are also considered. The mean is used to determine the closeness $c_{\mathrm{v},k}$ of the center of the broadcast as likewise done in Eq. (1). Then M_{v} vectors

$$\mathbf{v}_k^{\mathrm{T}} = \Bigg(\; c_{\mathrm{v},k} \, , \, \sigma^2_{\mathrm{v},k} / \max_{\forall k}\Big(\sigma^2_{\mathrm{v},k}\Big) \, , \, t_{\mathrm{v},k} / \max_{\forall k}\Big(t_{\mathrm{v},k}\Big) \; \Bigg) \tag{4}$$

are defined. Similarly to Eq. (3), the selected AP cluster has the maximum value of the Euclidean norm of all vectors $\mathbf{v}_k$. With the chosen cluster, the start and end time points of the AP segments are given via the shots of the key-frames, from which the faces are extracted. The segments then are available for the decision fusion.

2.3. Decision fusion

As the results from cluster analysis are available only as hard decisions and there are no confidence values, the decision fusion is used [12]. For the combination of the decisions of the auditory and visual modality, the AND as well as the OR operation are available in order to determine the final AP segments. Furthermore, it is obvious that the AND operation maximizes the precision whereas the OR operation maximizes the recall.

Even though the decision fusion is quite simple, the results can benefit from the OR fusion, especially for topic-oriented navigation. This requires a high precision of the results in each analysis domain. For example, it is possible to detect additional AP segments in the audio signal, which were not detected in the visual domain. That is also possible in the other way around. Furthermore, while the AP is detected continuously in a certain studio segment, shot boundaries caused by background changes in the visual domain indicating also the change of a topic. This can be used to split these studio segments into adjacent AP segments.

3. RESULTS

Our approach was evaluated in terms of the well-known recall (RCL), precision (PRC) and combined F_1 measure. Two different granularities were considered. The measures were determined at the level of I-frames and at the level of AP segment boundaries. A five seconds fuzzy window was used for the evaluation of the segment boundaries.

Our tests involved video material of three broadcast news sequences from public German TV with a total duration of one hour and a resolution of 720 × 576 pixels. Excellent results were achieved using the OR operation in the decision fusion. This improved the completeness of the AP segments in comparison to an exclusive audio or visual analysis. The OR operation reached an average F_1 measure of 0.98 for I-frames and 0.97 for segment boundaries (Tab. 1).

Fusion Type	**Video Material**	**Anchorperson Segments**					
		I-Frames			**Segment Boundaries**		
		PRC	RCL	F_1	PRC	RCL	F_1
AND	A	1.00	0.70	0.82	0.94	0.80	0.86
	B	1.00	0.80	0.89	1.00	0.85	0.92
	C	1.00	0.87	0.93	1.00	1.00	1.00
OR	A	0.98	0.99	0.98	1.00	0.90	0.95
	B	0.98	0.99	0.98	0.96	0.96	0.96
	C	1.00	1.00	1.00	1.00	1.00	1.00

Table 1. Results after decision fusion for OR ground truth

The demonstration program (Fig. 4) detects the AP segments on a current personal computer about seven times faster than real time. During the analysis, intermediate results of the processes are visualized, e.g. face detection or audio segmentation results. After the analysis, the AP segments are shown in form of a diagram per modality and a list for the combined results with extracted key-frames and additional time information.

4. CONCLUSIONS AND FURTHER WORK

In this paper, we have formulated an anchorperson detection technique by separate analysis of the audio and video signals. The overall results are formed by decision fusion. It has been experimentally demonstrated that the proposed technique obtains highly accurate results. Further work should prove the system on a larger amount of common used news test videos. The more consequent usage of the compressed domain data can increase the computational efficiency. In addition, the cluster selection could be modified in such a way that it provides N-best matches and therefore further anchorpersons or other visible speakers could be identified. As byproduct, the information about faces in key-frames and the studio segments can be used for metadata enhancement.

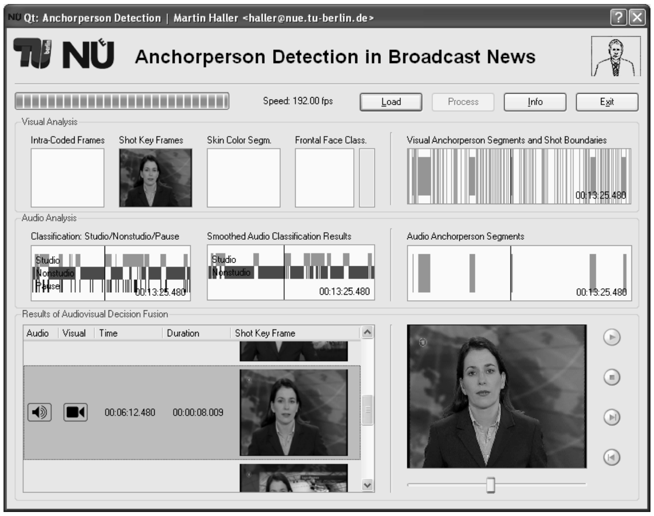
Qt: Anchorperson Detection | Martin Haller <haller@nue.tu-berlin.de>
Anchorperson Detection in Broadcast News
Speed: 192.00 fps
Load
Process
Info
Exit
Visual Analysis
Intra-Coded Frames
Shot Key Frames
Skin Color Segm.
Frontal Face Class.
Visual Anchorperson Segments and Shot Boundaries
00:13:25.480
Audio Analysis
Classification: Studio/Nonstudio/Pause
Studio
Nonstudio
Pause
00:13:25.480
Smoothed Audio Classification Results
Studio
Nonstudio
00:13:25.480
Audio Anchorperson Segments
00:13:25.480
Results of Audiovisual Decision Fusion
Audio
Visual
Time
Duration
Shot Key Frame
00:06:12.480
00:00:08.009

Fig. 4. Results are shown after anchorperson detection

5. ACKNOWLEDGMENT

This research was supported by the European Commission under contract FP6-027026-K-SPACE.

6. REFERENCES

[1] S. W. Smoliar and H.-J. Zhang, "Content-based video indexing and retrieval," *IEEE Multimedia*, vol. 1, no. 2, pp. 62–72, 1994.

[2] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Mag.*, vol. 17, no. 6, pp. 12–36, 2000.

[3] C. G. M. Snoek and M. Worring, "Multimodal video indexing: a review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[4] I. Ide, K. Yamamoto, and H. Tanaka, "Automatic video indexing based on shot classification," in *Lecture Notes in Computer Science*, 1999, vol. 1554, pp. 87–102.

[5] W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating visual, audio and text analysis for news video," in *Proc. IEEE Int. Conf. Image Processing*, 2000, vol. 3, pp. 520–523.

[6] Dong-Jun Lan, Yu-Fei Ma, and Hong-Jiang Zhang, "Multi-level anchorperson detection using multimodal association," in *Proc. Int. Conf. Pattern Recognition*, 2004, vol. 3, pp. 890–893.

[7] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 audio and beyond: audio content indexing and retrieval*, Wiley, 2005.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[9] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, no. 1–2, pp. 111–126, 2000.

[10] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases*, 1998, pp. 290–301.

[11] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proc. Int. Conf. Database Theory*, 2001, pp. 420–434.

[12] C. Sanderson and K. K. Paliwala, "Identity verification using speech and face information," *Digital Signal Processing*, vol. 14, no. 5, pp. 449–480, 2004.