Multi-level Anchorperson Detection Using Multimodal Association

Dong-Jun Lan[†], *Yu-Fei Ma*[‡], *Hong-Jiang Zhang*[‡]
[†]Dept. of Electronic Engineering, Tsinghua University, Beijing, China (100084)
[‡]Microsoft Research Asia, 5F Sigma Center, 49 Zhichun Road, Beijing, China (100080)

Abstract

In contemporary TV news programs, multi-level anchorpersons are often used which indicate the inherent hierarchical structure of news program. However, these diverse anchorperson patterns make the conventional anchorperson detection algorithms failed. In this paper, we propose a robust approach to anchorperson detection by integrating visual modality, auditory modality and human appearance modality into multimodal associated clustering. Based on the structure of clustered multi-level anchorpersons, the ToC (Table-of-Content) of news video can be effectively generated. The effectiveness and robustness of the proposed approach are demonstrated by the experiments on five hours news programs from different TV channels.

1. Introduction

News video is a sort of well structured video, because it has the similar structure of book, such as chapter, section, etc. These structural elements are usually indicated by the anchorpersons at different levels. Therefore, anchorperson detection is key issue in news video content analysis. Although there are many news video analysis algorithms in the literatures, anchorperson detection is still an open issue because the advanced TV technologies have made TV news much more diverse and complicated than before.

There have been many prototypes of news video system, such as MEDUSA system [1], Broadcast News Navigator System [2] and Informedia Project [3]. Anchorperson shot is an important cue to extract the structure of news video, although some of the systems mainly rely on textual or linguistic information such as overlaid captions and transcripts. The template-based anchorperson detection method proposed in [4] assumes that different anchorperson models have the same background. Zhang, *et al.*, construct three models for an anchorperson shot respectively, shot, frame, and region, in order to adapt to various TV news programs [5]. However, it is difficult to construct models for each kind of anchorperson in different news programs. A graph-theoretical cluster algorithm is applied to classify video shots into anchorperson shots and news footage shots in [6]. In fact, as only visual information is used, the shots with similar visual content may be falsely detected as anchorperson shots.

Obviously, these methods cannot handle the complicated cases of contemporary TV news programs. For example, there are often five or more anchorpersons, such as main anchorperson, news reporter for special topic, sports section anchorperson, financial section anchorperson, etc. Also, the background and camera angle may change greatly. Moreover, some anchorperson patterns or special background appear only once in entire news program. In this paper, we proposed a robust approach to detect multiple level anchorpersons in TV news program, using multimodal associated unsupervised clustering. Base on multi-level anchorpersons, ToC (Table-of-Content) of news video is easily generated. Experiments on five hours news program collected from five different TV channels have demonstrated the effectiveness and robustness of the proposed scheme.

The rest of the paper is organized as follows. Section 2 gives a detailed description of multi-level anchorperson detection. In Section 3, the ToC of news video generated by our approach is presented. The evaluation results are given in Section 4. Section 5 concludes the paper.

2. Multi-level anchorperson detection

The proposed system is composed of three modules, candidate selection module, multimodal associated clustering module and final decision module. As shown in Figure 1, SVMs are first used to determine anchorperson shot candidates from entire news video. Then multimodal associated clustering is applied which associates visual, auditory and human appearance modalities. Final decision is made to identify commercials and weather forecast sections, as well as to remove false alarm, such as dialogue section in news program.

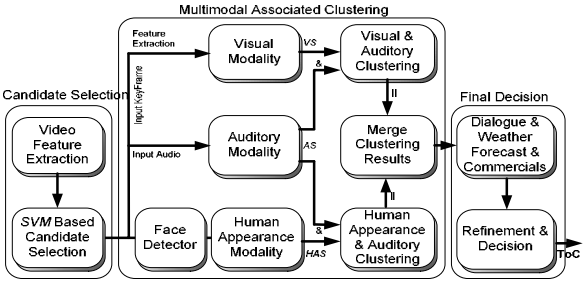

Figure 1. System Overview

2.1. Candidate selection

The task of candidate selection is to identify possible anchorperson shots, or, anchorperson shot candidates. Here, SVMs are employed as two-class classifier for anchorperson candidate selection. The input of SVMs is 14-dimension feature vector x_i, including P_M, P_{NZ}, M_{MAD}, V_{MAD}, M_{ECR}, V_{ECR}, M_{AFL}, V_{AFL}, SCR_i and BFR_i (i = 1, 2, 3).

P_M, P_{NZ}, M_{MAD} and V_{MAD} denote *Percentage of All Types of Camera Motion*, *Percentage of Not Zoom Camera Motion*, *Mean of Frame Difference MAD* and *Variance of Frame Difference MAD*, respectively. They are computed based on the dominant camera motion curve (*MAJ*) and frame difference curve (*MAD*) proposed in [7],

$$P_M(S) = \frac{M_{pan} + M_{tilt} + M_{zoom}}{N-1} * 100\,\% \quad (1)$$

$$P_{NZ}(S) = \frac{M_{pan} + M_{tilt}}{N-1} * 100\,\% \quad (2)$$

$$M_{MAD}(S) = \frac{1}{N-1}\sum_{n=1}^{N-1} MAD_n \quad (3)$$

$$V_{MAD}(S) = \frac{1}{N-1}\sum_{n=1}^{N-1} (MAD_n - M_{MAD}(S))^2 \quad (4)$$

where M_{pan}, M_{tilt}, M_{Zoom} are the numbers of pan, tilt, zoom motions in shot S, N is the total frame number of S and MAD_n is frame difference between frame $n+1$ and n. Similarly, M_{AFL} and V_{AFL} are the mean and variance of *AFL* (Average Frame Luminance). M_{ECR}, V_{ECR} are the mean and variance of *ECR* (Edge Change Rate) defined as,

$$ECR_n = \max\left(\frac{E^{in}_{n,n+1}}{E_{n+1}}, \frac{E^{out}_{n,n+1}}{E_n}\right) \quad (5)$$

where E_n is the total number of edge pixels in frame n, $E^{in}_{n,n+1}$ and $E^{out}_{n,n+1}$ are respectively entering and exiting edge pixels between frame $n+1$ and n.

SCR denotes shot change rate within the neighborhood. We use the minimum *SCR* of left-side and right-side neighborhood of the shot boundary as the value of *SCR*. *BFR* denotes black frame rate. *SCR* and *BFR* are computed in 200, 400 and 600 frames' neighborhood respectively, that is, SCR_i and BFR_i (i = 1, 2, 3).

For given shot i the output of SVMs is class y_i, where $y_i \geq 0$ indicates shot i is anchorperson candidate shot. The training of SVMs is the solution of following optimization,

$$\min_{\omega,b,\xi}\left(\frac{1}{2}\omega^\tau\omega + C\sum_{i=1}^{l}\xi_i\right) \quad (6)$$
$$subject\ \ to\ \ y_i\left(\omega^\tau\phi(x_i)+b\right) \geq 1-\xi_i, \xi_i \geq 0$$

where ω, b, ξ are output parameters, ϕ decides the kernel function of SVMs. Here, RBF (Radial Basis Function) is employed. At this pre-processing stage, as we only select anchorperson candidates instead of accurately identifying anchorpersons, the classification results should not miss any anchorperson shot, but may contain some false alarms. Therefore, we set higher *Weight* for anchorperson candidate class (default 1), that is, the parameter C in (6) for anchorperson candidate class is *Weight*C*.

2.2. Multimodal associated clustering

Multimodal associated clustering is applied to identify anchorperson shots from the candidates by clustering the periodically occurring candidate shots in news program. Visual features, auditory features and human appearance features are first extracted respectively. As anchorperson shots are those having similar visual background, at the same time having similar auditory speech, or those having similar newsreader appearance and speech, although with different visual background, we define multimodal association as a Boolean expression in (7), based on which the associated clustering is conducted:

$$((V\ \&\ A)\ ||\ (HA\ \&\ A)) \quad (7)$$

where *V*, *A*, and *HA* denote that shots have similar visual modality, auditory modality and human appearance modality, respectively. While *&* (*and*) and || (*or*) are the association operations between different modalities.

For visual modality, we extract visual similarity metrics *VS*. Four color and four texture based features are extracted first. That is, one 256-bin *HSV* histogram, three 64-bin *Y, U, V* histograms and four color correlograms with the distance d = 1, 3, 5, 7. In addition, five region-based histograms are also extracted. The histograms of corresponding N regions of two keyframes are compared. The K regions with the largest histogram correlation are used and the similarity of the two keyframes is defined as the sum of the histogram correlation of the selective K regions. In our implementation, N = 16 and K = 8, 10, 12, 14, 16 respectively. To compute similarity of each visual feature, pearson's correlation coefficient is used,

$$Cor(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \quad (8)$$

where the range of *Cor* is from -1 to 1. In this way we get M-dimension *VS* (Visual Similarity Metrics), $M = 13$.

From audio track, we extract auditory features and compute their similarities according to [8]. The *MFCC* (Mel-frequency Cepstrum) is extracted from *FFT* power coefficients because it is measure of the subjective pitch and frequency content of audio signals. Additionally, *Pitch* and *LSP* (Linear Spectral Pairs) are also used. Based on these we get three-dimension *AS* (Audio Similarity Metrics).

We also extract human appearance modality from keyframes to cluster the anchorperson shots with different camera angles or background. Given a face region located by face detection module in [9], the head-shoulder portrait is then identified; from this region human appearance features are extracted. Besides the thirteen visual features defined above, other two features,

dominant color and color moment are also extracted. Dominant color is the color that occupies the largest percentage of human head-shoulder region. The first central moment of color is used to describe the color-spatial distribution of human appearance region. Using correlation coefficient in (8), we then get fifteen-dimension *HAS* (Human Appearance Similarity Metrics).

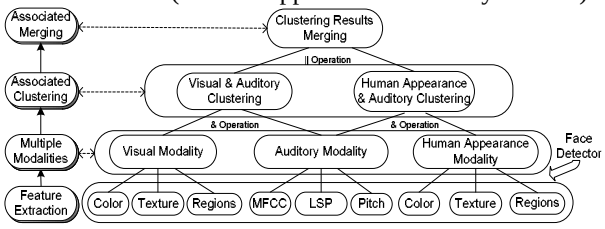

Figure 2. Multimodal Associated Clustering

Figure 2 shows an overview of multimodal associated clustering. For & (*and)* operation of modality association, the weighted similarity summation of different modalities is used as the input of unsupervised clustering. The ‖ operation is implemented by clustering results merging at the top level. During the clustering process, the similarity measure between element and cluster, or two clusters, is defined as the average pairwise distances of elements. Given input element E_i , the winner cluster C_I to which to assign E_i is determined through

$$C_I = \arg\min_n \left(\max_{1 \le n \le N} \left(Sim\left(E_i, C_n\right) \right) \right) \quad (9)$$

where N is the current cluster number. The clustering process can be described as follow,

(1) Initialize: the first element E_1 as initial cluster C_1
(2) For each input element E_i, calculate the winning class C_I according to (9)
(3) Whether E_i is added to C_I or not is decided according to threshold Th_I, which is the variance of the cluster C_I. If E_i is assigned to C_I, Th_I is updated, otherwise, create a new cluster C_{N+1}
(4) Continue (3) until all elements have been assigned a label, then start following refinement
(5) Move: find the nearest C_I for each element E_i and then move E_i into C_I
(6) Merge: if the similarity of two cluster C_i and C_j is greater than Th_d, C_i and C_j are merged
(7) Split: for element E_i^m of cluster C_m, if $\frac{1}{M-1}\sum_{j=1, j\neq i}^{M} Sim\left(E_i^m, E_j^m\right) \le Th_d$, E_i^m is deleted from C_m
(8) Continue (5), (6), (7) until no change occurring or given step arrived, output clustering result.

In this manner synchronized multimodal information are incorporated. Different modalities complement each other and are more robust than single modality.

2.3. Final decision

According to the results of multimodal clustering, the groups of anchorperson, the dialogue segments, weather forecast segments as well as commercial segments are identified. Dialogues segments are those several interview scenes, interviewee and interviewer recursively appear. Weather forecasts are those identical weatherman continuously appear for multiple shots in small duration of news program. Currently, commercial segments are detected according to *BFR* (Black Frame Rate).

In fact, anchorperson shot is not the only type of shots that have multiple matches along the entire news program, because some similar shots appearing more than once will also be clustered. We define four criteria to remove false alarms and get the final results. Given K clusters $C_1,\ldots, C_K$, while cluster C_I having M shots $S_1,\ldots, S_M$, and N frames $F_1,\ldots, F_N$, T_{FI} (Total Frame Interval), T_{SI} (Total Shot Interval)), M_{SI} (Mean Shot Interval), N_{AS} (Number of Shots with Similar Audio) for C_I are defined respectively,

$$T_{FI}(C_I) = \max_{F_i^I \in C_I}\left(F_i^I\right) - \min_{F_j^I \in C_I}\left(F_j^I\right), \quad (10)$$

$$T_{SI}(C_I) = \max_{S_i^I \in C_I}\left(S_i^I\right) - \min_{S_j^I \in C_I}\left(S_j^I\right) \quad (11)$$

$$M_{SI}(C_I) = \frac{1}{N-1}\sum_{j=1}^{N-1}\left(S_{j+1}^I - S_j^I\right) \quad (12)$$

$$N_{AS}(C_I) = \{S_j \mid S_j \in \overline{C}_I; \exists S_i \in C_I, AS(S_i, S_j) \ge Th\} \quad (13)$$

where in (13) $AS(S_i, S_j)$ is auditory similarity between shot S_j and S_i, Th is predefined threshold. Those clusters with smaller T_{FI}, T_{SI}, M_{SI}, or N_{AS} are removed as false alarms.

3. News program ToC generation

Based on the results of multi-level anchorperson detection, ToC (Table-of-Content) of news program is generated to represent the inherent structure of news video. The ToC of news video is a four-level hierarchical structure, as shown in Figure 3.

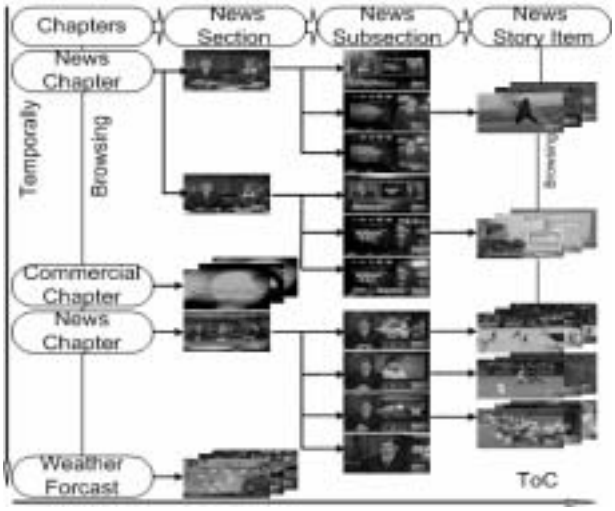

Figure 3. ToC of News Program

The highest level of news program ToC is chapter, including actual news chapter, weather forecast chapter and commercial chapter. The second level is news section indicated by main anchorpersons and the third level is sub-section indicated by reporter hand-off. The last level is news story item which is composed of individual news story shots.

The large clusters whose elements distribute in entire news program correspond to main anchorperson. While anchorpersons for specific sections, for example, sports section, financial section, or news reporter for specific topic, only appear in particular segments of news program, correspond to small clusters. At the third level, the reporter hand-off, such as anchorperson to reporter, reporter's report and reporter to anchorperson again, is clearly presented which is very useful for news video content understanding.

4. Experiments

We test our approach on 7 news programs with the total duration of more than 5 hours, which were collected from 5 different TV channels as shown in Table 1. The ground truth is manually labeled in advance.

Table 1 News videos for evaluation

	Program	*Length*	*Shots*	*Anchor*
I	*ABC*	*29:09*	*425*	*32*
II	*CBS*	*30:01*	*428*	*20*
III	*CNN*	*02:03:46*	*974*	*119*
IV	*Fox*	*38:06*	*441*	*27*
V	*NBC*	*27:05*	*351*	*44*
VI	*NBC*	*30:27*	*419*	*43*
VII	*NBC*	*30:10*	*432*	*20*

The evaluation results of multi-level anchorperson detection algorithm are listed in Table 2.

Table 2 Anchorperson detection evaluation

	Total	Detect	Miss	False	Recall *(%)*	Precision *(%)*
I	*32*	*27*	*5*	*0*	*84.38*	*100.00*
II	*20*	*26*	*0*	*6*	*100.00*	*76.92*
III	*119*	*119*	*10*	*10*	*91.60*	*91.60*
IV	*27*	*23*	*4*	*0*	*85.19*	*100.00*
V	*44*	*39*	*5*	*0*	*88.64*	*100.00*
VI	*43*	*38*	*5*	*0*	*88.37*	*100.00*
VII	*20*	*27*	*0*	*6*	*100.00*	*77.78*
Avg	***305***	***299***	***29***	***22***	***90.49***	***92.64***

The overall Recall is 90.49% and Precision is 92.64%. Comparing with those schemes proposed in [4-6], our approach is more robust, because the anchorpersons at different levels can all be identified, including main anchorperson, news reporter for special topic, sports section anchorperson, financial section anchorperson, etc. At the same time, because of the integration of multimodal association, including visual modality, auditory modality and human appearance modality, our algorithm handle those difficult cases with different background or camera angles, even those appearing only once in entire news program very well.

5. Conclusion

In this paper, we proposed a robust approach to anchorperson detection using multimodal association. Based on multiple-level anchorperson clustering, ToC of news video can be easily generated. The proposed approach overcomes the drawbacks in conventional methods, so that it can handle contemporary TV news program well. The promising results have been obtained in the extensive testing on five hours news videos from different TV channels.

6. References

[1] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, S. J. Young, "Automatic content-based retrieval of broadcast news," *ACM Multimedia'95*, San Francisco, CA, 1995, pp. 35-43.

[2] S. Boykin, A. Merlino, "Improving broadcast news segmentation processing," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, vol. 1, 1999, pp. 744-749.

[3] A. G. Hauptmann, M. J. Witbrock, "Informedia: News-on-Demand multimedia information acquisition and retrieval," *Intelligent Multimedia Information Retrieval*, 1997, pp.213-239.

[4] A. Hanjalic, RL Lagendijk, J. Biemond, "Template-based detection of anchorperson shots in news programs," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, 1998, pp. 148–152.

[5] H.J. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan, "Automatic parsing of news video," in *Proc. Int. Conf. Multimedia Computing and Systems*, 1994, pp. 45–54.

[6] X. Gao, X. Tang, "Unsupervised video shot segmentation and model-free anchorperson detection for news video story parsing", *IEEE Transaction on Circuits, Systems and Video Technology*, vol. 12, no. 9, pp.765-776, Sept., 2002.

[7] D.J. Lan, Y.F. Ma, H.J. Zhang, "A Novel Motion-based Representation For Video Mining," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, 2003.

[8] L. Lu, H.J. Zhang, "Speaker Change Detection and Tracking in Real-Time Broadcasting Analysis," *ACM Multimedia'02*, pp. 602-610, 2002.

[9] S. Z. Li, et al., "Statistical Learning of Multi-View Face Detection," in *Proc. European Conf. on Computer Vision 2002*, Denmark, May, 2002.