A Novel Anchorperson Detection Algorithm Based on Spatio-temporal Slice

Anan Liu[1,2], Sheng Tang[2], Yongdong Zhang[2], Jintao Li[2], Zhaoxuan Yang[1]
1. School of Electronic Engineering,
Tianjin University
Tianjin, 300072, China
2. Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, 100080, China
liuanan@ict.ac.cn

Abstract

For conveniently navigating and editing the news programs, it is very important to segment the video into meaningful units. The effective indexing of news videos can be fulfilled by the anchorperson shot because it is an indicator which denotes the occurrence of upcoming news stories. The paper presents a novel anchorperson detection algorithm based on spatio-temporal slice (STS). With STS pattern analysis, clustering and decision fusion, anchorperson shots can be detected for browsing news video. The large-scale experimental results demonstrate that the algorithm is accurate, robust and effective.

1. Introduction

Nowadays, we can access huge amount of information in visual modality. Consequently, an efficient management of the increasing amount of videos is strongly needed. For content-based video analysis, indexing and retrieval, researchers pay much attention to news video because of the unique structure feature shown in Fig.1 (a). A news video can be seen as the combination of anchorperson shots, news story shots and possible commercial breaks. The anchorperson shot and corresponding news story consist of an integrated news event. The segmentation of the events in a news video can facilitate users locating video content of particular interest.

To segment a news video into meaningful units, anchorperson detection is the most important. In [1], a template matching method, the earliest method for anchorperson detection, is proposed to extract anchor shots. However, the experimental result shows that the algorithm is not robust. To enhance the robustness, an anchor shots detection method integrating visual, auditory and human appearance modalities is presented in [2]. Although stronger robustness can be achieved, it is time-consuming to train the modal for each anchorperson. To solve this problem, Ki Tae PARK et-al presented a modal-independent anchor detection method based on anchor object extraction in [3]. It achieves high accuracy but the performance greatly depends on static background and the preciseness of shot change detection. In conclusion, to detect anchorpersons in news video, there are mainly two difficulties. One results from multiple video sources by different providers. Different anchorpersons and studio background make it difficult to have a universal anchorperson detection method for news videos. The other results from the advanced editing technique. Diversity of shot boundary formats, i.e. cut, wipe, dissolve and so on, makes it difficult to extract integrated story units accurately and automatically. Therefore, the existing anchorperson detection methods cannot perfectly satisfy the application in news video analysis.

To overcome the difficulties mentioned above, we present a novel anchorperson detection algorithm based on spatio-temporal slices (STS). With STS pattern analysis, clustering and decision fusion, we can quickly extract integrated anchorperson shot. Not only does the algorithm can get accurate results with low computational complexity, but also it is robust to different video sources and video editing technique because it neither depends on specific templates or modals nor needs to extract shot boundary.

14th International Conference on Image Analysis and Processing (ICIAP 2007)
0-7695-2877-5/07 $25.00 © 2007 IEEE

The rest of the paper is organized as follows: Section 2 specifically illustrates the anchorperson detection algorithm based on STS. Section 3 provides the experimental results and detailed analysis. In section 4, conclusions and future advanced work are presented.

2. Anchorperson detection algorithm based on STS

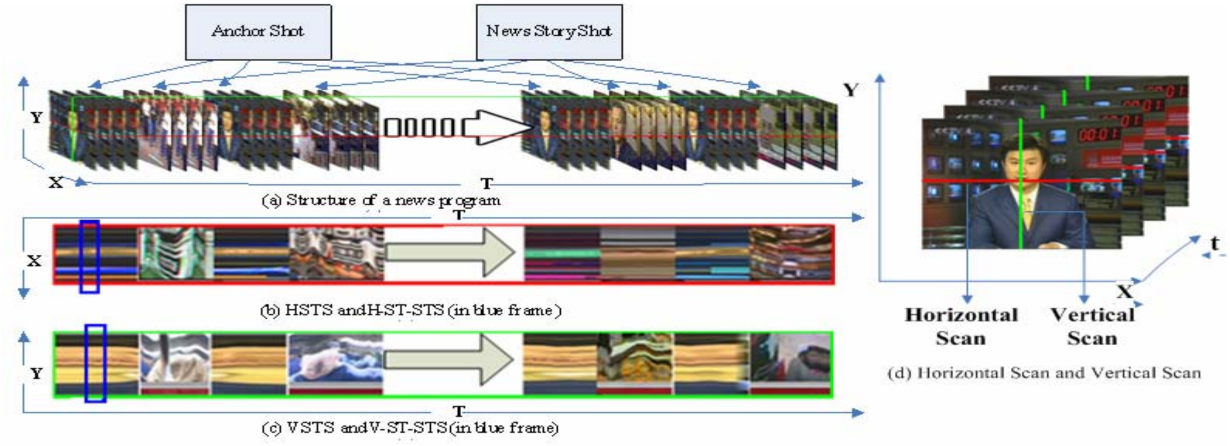
2.1. Definitions and detection rule

For clear introduction and easy understanding of the method, the definitions of some important concepts are elaborated in this section. Besides with the analysis of the characteristics of the anchor shots, the detection rule is presented.

2.1.1. Definitions

(a) ***Spatio-temporal slice (STS)***: Spatio-temporal slice is a collection of scans (a scan is a strip composed of a line of pixels in an image shown in Fig.1.(d)) in the same selected position of every frame of a video as a function of time [4]. There are many kinds of STS with different selection methods of scans. Horizontal STS (HSTS) is formed by horizontal scans and vertical STS (VSTS) is formed by vertical scans shown in Fig.1 (b) (c) (d).

Short-time spatio-temporal slice (ST-STS): STS can be divided into a set of coherent STS clips with the same length in temporal domain. Each STS clip is called short-time spatio-temporal slice. There are also two kinds of ST-STS, namely, horizontal ST-STS (H-ST-STS) and vertical ST-STS (V-ST-STS) shown in Fig.1 (b) and (c).



(a) Structure of a news program

(b) HSTS and H-ST-STS (in blue frame)

(c) VSTS and V-ST-STS (in blue frame)

(d) Horizontal Scan and Vertical Scan

Fig. 1. Video sequence of a news program and corresponding HSTS/VSTS, H-ST-STS/V-ST-STS

Let T be the number of frames in a video, L be the length of each ST-STS and F be a frame of size M*N. 1-D horizontal scans and vertical scans can be represented by the concatenation of pixels in one line as follows:

$$h_i =< P_{F_i}(1,\tfrac{N}{2}),...,P_{F_i}(x,\tfrac{N}{2}),...,P_{F_i}(M,\tfrac{N}{2})>$$
$$v_i =< P_{F_i}(\tfrac{M}{2},1),...,P_{F_i}(\tfrac{M}{2},y),...,P_{F_i}(\tfrac{M}{2},N)> \quad (1)$$

where h_i and v_i denote the horizontal and vertical scans extracted from i^{th} frame (F_i) in the video and $P_{Fi}(x, N/2)$ and $P_{Fi}(M/2, y)$ mean the positions of pixels in i^{th} frame (F_i). H-ST-STS and V-ST-STS can be represented by the concatenation of scans extracted from the adjacent frames as follows:

$$h_j^{ST} =< h_{((j-1)*L/2+1)}, h_{((j-1)*L/2+2)},......,h_{((j-1)*L/2+L)} >$$
$$v_j^{ST} =< v_{((j-1)*L/2+1)}, v_{((j-1)*L/2+2)},......,v_{((j-1)*L/2+L)} > \quad (2)$$

where h_j^{ST} and v_j^{ST} separately mean the j^{th} H-ST-STS and V-ST-STS and each ST-STS overlaps the previous one with L/2 length (L can be exactly divided by 2). Then the STS of one video can be composed of K ST-STSs. K is calculated by

$$K = 2*[T/L]-1 \quad (3)$$

where "[ ]" means the operation that the value of *T/L* is rounded. Then HSTS and VSTS can be represented by :

$$HSTS = <h_1^{ST}, h_2^{ST}, ..., h_K^{ST}>$$
$$VSTS = <v_1^{ST}, v_2^{ST}, ..., v_K^{ST}> \quad (4)$$

(b) ***Anchorperson shot***: Anchorperson shots are visually characterized by studio background and by one or two news readers appearing separately or together, also with some possible variations of camera angle and changes in news icons appearing in a screen corner [1]. Then ***anchor shots group*** is comprised of all the anchor shots in one video.

2.1.2. Detection rule

The anchor shots and groups have the following characteristics [5]:

0-7695-2877-5/07 $25.00 © 2007 IEEE

(a) Anchorperson shots tend to occur periodically in a news video.
(b) Anchorperson shots tend to have great similarity with each other.
(c) Anchorperson groups tend to be larger than most of other groups due to the fact that the anchorperson shots have much similarity in visual features while other shots match well only in their closest neighborhood within the same news story.

Based on the characteristics mentioned above, we detect the anchorperson shots on the rule that the anchorperson shot is the only type of news unit that has multiple matches of its visual content along the entire program [1]. With the rule, it is theoretically feasible that the ST-STS extracted from the anchorperson shots can be easily classified into the same cluster and this cluster definitely has the most elements.

2.2. Anchorperson detection

In this section, we will illustrate anchorperson detection algorithm based on STS, including three modules, namely, STS pattern analysis and feature extraction, ST-STS clustering and decision fusion.

2.2.1. STS pattern analysis and feature extraction

By scrupulously observing STS shown in Fig.1 (b) and (c), two opinions can be obtained. On one hand, the discontinuity of color and texture denotes the appearance of a new shot. With the characteristics mentioned in Section 2.1.2, the anchorperson frames usually have great similarity with each other. Therefore the patterns of H-ST-STS and V-ST-STS extracted from the anchorperson shots, especially the adjacent ones with overlapped section, change slightly. As a result we can classify the similar ST-STS into the same cluster. On the other hand, STS provides rich visual cues along a larger temporal scale while the conventional methods only focus on the neighboring frames. Although there may be changes of studio background, persons' movement and advanced editing technique in some frames, the detection cannot be severely influenced with scans in them because they are only a small part of STS.

It is perceivable that color features and texture features well depict the characteristics of STS. As for color feature of each ST-STS, we equally divide the picture into 4*4 rectangular blocks and extract 32-bin histogram for them to depict the local color feature. Moreover, 3 color moments are also taken into account for global color feature. As for texture feature, we adopt edge histogram descriptor proposed in [6]. Consequently, each ST-STS is represented by a visual feature vector including 515 dimension color feature and 150 dimension texture feature.

2.2.2. ST-STS Clustering

K-mean algorithm is a simple and effective clustering algorithm. It is implemented to classify both H-ST-STS and V-ST-STS represented by the visual feature vectors mentioned above.

After clustering ST-STS, we separately combine coherent H-ST-STS/V-ST-STS in the temporal domain as one element in the clusters. Each element can be mapped to one Shot, a kind of physical unit of video, with the staring frame and the ending frame, namely, the starting time and the ending time. Then the composition of cluster can be seen as:

$$\begin{aligned} Cluster_i^H &= < Shot_{i1}^H, Shot_{i2}^H, ...Shot_{iR}^H > \\ Cluster_j^V &= < Shot_{j1}^V, Shot_{j2}^V, ...Shot_{jS}^V > \end{aligned} \qquad (5)$$

where $Cluster_i^H$ and $Cluster_j^V$ respectively denotes i^{th} *Cluster* in H-ST-STS clustering result and j^{th} *Cluster* in V-ST-STS clustering result and *R/S* means the element number of the H-ST-STS / V-ST-STS cluster.

2.2.3. Decision fusion

Because both H-ST-STS and V-ST-STS clustering results are useful for anchor person detection, they are fused to achieve the more accurate result with both horizontal and vertical information. Depending on the rule stated in Section 2.1.2, we only fuse the $Cluster_i^H$ and $Cluster_j^V$ with the most elements, i.e., ${Cluster_{Max}}^H$ and ${Cluster_{Max}}^V$. The similarity of $\mathrm{Shot_i}^\mathrm{H}$ in ${Cluster_{Max}}^H$ and $\mathrm{Shot_j}^\mathrm{V}$ in ${Cluster_{Max}}^V$ is calculated by

$$Sim < Shot_i^H, Shot_j^V > = \frac{Min(T_{End}^H, T_{End}^V) - Max(T_{Start}^H, T_{Start}^V)}{Max(T_{End}^H, T_{End}^V) - Min(T_{Start}^H, T_{Start}^V)} \qquad (6)$$

where $T_{Start}^H, T_{End}^H, T_{Start}^V, T_{End}^V$ separately denotes the starting time and ending time of $Shot_i^H$ and $Shot_j^V$ and *Min* and *Max* separately mean the operation of choosing the smaller value and bigger value . If the value of similarity is less than 0, it is rectified as 0. In reality, one shot in one cluster can only match another in the other cluster. Then the similarity of ${Cluster_{Max}}^H$ and ${Cluster_{Max}}^V$ is calculated by

$$Sim < Cluster_{Max}^H, Cluster_{Max}^V > = \frac{1}{Min(R,S)} \sum_{i=1}^{R} \sum_{j=1}^{S} Sim < Shot_i^H, Shot_j^V > \qquad (7)$$

The fusion method is listed as follows:

(a) If the similarity between ${Cluster_{Max}}^H$ and ${Cluster_{Max}}^V$ is bigger than the presetted threshold, they are classified into one cluster. In this case, if one shot do not have the corresponding shot in another cluster, it is regarded as the element of the new cluster; otherwise, the combination of them with the earliest starting time and latest ending time is seen as the element of the new cluster. Then the new cluster is the anchor shots group.

(b) If the similarity between the two clusters is not bigger than the presetted threshold, the two clusters cannot be fused. Thus we do not extract anchor shots because the video does not obviously include this kind of segments.

3. Experimental results

To demonstrate that algorithm proposed in the paper is a robust method for different video sources and advanced editing technique, we select 50 news videos for test from the test data of high-level feature extraction in TRECVID 2005 and 2006. The specific information of our test data is shown in Table1. The experimental results shown in Table2 demonstrate that the anchorperson detection algorithm based on STS has high accuracy and strong robustness.

Besides compared to the method mentioned in [7] with high computational complexity of processing adjacent frames, this algorithm save lots of time by processing STS. Experimental results show that for the same video lasting for 5 minutes, the mean time with the novel method is about 3 minutes and the mean time with the method in [7] is about 7.2 minutes which is about 2.4 times of the former one. The comparison shows that the novel algorithm is more effective.

4. Conclusions and future work

The paper presents a novel anchorperson detection algorithm based on STS. The large-scale experimental results demonstrate that the algorithm is accurate, robust and efficient. It can be used to index the news video and facilitate users browsing video content of particular interest.

Moreover, multi-modal video content analysis can integrate the superiority of each modality and achieve more satisfactory results. As the future work, audio and text technique will be implemented on the anchorperson detection. With the multi-modal method, we can not only detect the anchor shots but also identify anchorpersons. Consequently, an ideal news navigation system can be founded for browsing and retrieving broadcast news.

5. Acknowledgements

This work was supported by Beijing Science and Technology Planning Program of China (D0106008040291).

6. References

[1] Hanjalic, A., Lagensijk, R.L., Biemond, J., “Template-based detection of anchorperson shots in news programs,” ICIP Proc., pp.148 – 152, 1998.

[2] Dong-Jun Lan, Yu-Fei Ma, Hong-Jiang Zhang, “Multi-level anchorperson detection using multimodal association,” Proc. of 17th International Conference on the Pattern Recognition, Vol.3, pp: 890- 893, 2004.

[3] Ki Tae PARK, Doo Sun HWANG, Young Shik MOON, “Anchor Frame Detection in News Video Using Anchor Object Extraction,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E88-A, pp: 1525 –1528, 2005.

[4] Chong-Wah Ngo, Ting-Chuen Pong, and Roland T. Chin, “Video Partitioning by Temporal Slice Coherency,” IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, pp.941 - 953 No.8,August 2001.

[5] N. E. O'Connor, S. Marlow et al, “Físchlár: An On-Line System For Indexing And Browsing Broadcast Television Content,” 2001.

[6] DK Park, YS Jeon, CS Won, and S.-J. Park, Efficient use of local edge histogram descriptor, Proc. of the ACM Workshops on Multimedia , Los Angeles, CA, Nov. 2000.

[7] Xi-Dao LUAN, Yu-Xiang XIE et al, “AnchorClu: An Anchorperson Shot Detection Method Based on Clustering,” Proc. of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies, 2005.

Table 1. Test data

Video source	Number	Time(minute)
NTDTV (from 20041101_120001 to 20041109_120100)	10	About 30
LBC (form 20041031_200001 to 20041105_200001)	10	About 30
CCTV (form 20041105_150000 to 20041114_150000 from 20051201_145800 to 20051217_145800)	30	About 10

Table 2. Experimental results

Video source	Labeled shots	Detected shots	False acceptance	False rejection	Precision (%)	Recall (%)
NTDTV	199	199	0	0	100	100
LBC	196	193	1	4	99.48	97.96
CCTV	502	500	0	2	100	99.60