# ANCHORPERSON SHOT DETECTION IN MPEG DOMAIN

*Zhong Ji, Chuntian Zhang, Yuting Su*

School of Electronic and Information Engineering, Tianjin University, Tianjin 300072, P. R. China
ji.zhong@hotmail.com, {zhangct, ytsu} @tju.edu.cn

## ABSTRACT

In this paper, a refined ASD algorithm in MPEG compressed domain is proposed. The new method is expected to outperform the existing strategies based on the following two improvements. One is that an effective face detection method is introduced. It aims to further remove the false alarms in the candidates generated from the previous modules, employing the chrominance DC coefficients. The other is the utilization of a new robust metric, which represents the dissimilarity of the video frames in the unsupervised clustering module. The proposed algorithm has been evaluated by six different TV channels. Compared with the state-of-the-art methods, the new algorithm is effective and computationally efficient.

## 1. INTRODUCTION

News video is one of the most important types in the research of multimedia content analysis. One fundamental step toward effective indexing of news video is to partition it into small, single-story units according to their semantics. Anchorperson shot offers an important cue for this so-called story segmentation task, since a news story is generally composed of an anchorperson shot followed by relevant news footage. Consequently, anchorperson shot detection (ASD) plays a crucial role in news video content analysis.

In recent years, there have been several prior works on this topic. The earlier work by Zhang et al. [1] predefined a set of models of an anchorperson shot and then matched against them with all the shots in a news video to get the potential anchorperson shots. However, this template-matching approach has a severe limitation since it is difficult to construct a general model for different kinds of news video. Gao and Tang [2] proposed a graph-theoretical clustering algorithm to classify video shots into anchorperson shots and news footage shots. Combining the previous two ideas, Anna et al. presented a mutli-stage approach by developing anchorperson shot models in an unsupervised way [3]. Besides visual modality, auditory modality is also integrated in [4] and [5], which is helpful to improve the detection precision.

In fact, most of the above algorithms are operated in pixel domain, where the complicated decoding procedure and some feature extraction and analysis processes (e.g. face detection) are quite time-consuming. However, little work is investigated in MPEG domain except for a few cases [6, 7]. In [6], Wang and Gao proposed a template-based ASD method on MPEG video, which is based on two assumptions: the anchorperson face region is fixed and there exists an unchanging feature region in the background of the anchorperson DC image (DC image is a low-resolution image generated from the original DCT compressed image using only the DC component of each 8 x 8 block). Since human face is a distinct semantic feature for anchorperson shot, Avrithis et al. [7] applied it to identify anchorperson shots directly in MPEG domain, where shots with one or two face close-ups are classified as single or double anchorperson shots.

In this paper, we propose a new ASD algorithm in MPEG compressed domain, which improves the effectiveness and execution speed of ASD mainly in two ways. First, an improved face detection algorithm in MPEG domain, especially suitable for ASD, is presented. Performed before the clustering module, it may reduce the false alarms by removing candidates containing no face. Second, a novel metric is introduced in the clustering module, which is proved to be very effective.

The paper is organized as follows. Section 2 gives a detailed description of the proposed ASD algorithm. Experimental results are presented in section 3. Section 4 concludes the paper.

## 2. THE PROPOSED ALGORITHM

### 2.1. Candidate Selection

On the basis of the characteristics of anchorperson shots, three steps are applied in this module to select the candidates, which mainly aim at reducing the computational burden of the following modules. In the first step, based on the results of the shot boundary detection on MPEG-coded news video, those shots with duration over two seconds are selected since anchorperson shots last rarely less than two seconds. And then, motion vectors for the P-macroblocks

This work is supported by 863 Project (No. 2006AA01Z407).

are used to choose those shots with low visual activity, because anchorpersons in the programs usually have no severe movement. Finally, for each remained shots, a key frame is selected as candidate for further investigation. Here we choose the last I-frame in each candidate shot.

### 2. 2. Face Detection

Human face is one of the most important semantic features for anchorperson shots. And candidates containing no face should be considered as false alarms. In this subsection, making use of the skin color, size and shape of faces, we introduce an effective face detection technique to remove this kind of false alarm.

The face in anchorperson key frames has lots of advantages for detection. For example, there are generally only one or two frontal human faces in the center of the image, without any occlusion. Moreover, the size of an anchorperson's face is large enough for face recognition. We may name the face in such a condition as 'distinct face', which is not only easier to detect, but also an important difference between anchorperson shots and news footage shots. However, video data is so massive that the heavy computational burden restricts the use of face detection technique. In [3], the authors considered that face detection is too time-consuming for practical application. Face detection in compressed domain makes it possible to use human face feature for fast ASD. The technique has been investigated in recent years, and experiments have proved its effectiveness and rapidness [8, 9].

Using the DC coefficients of Cb and Cr blocks, we present a fast face detection algorithm for ASD. Since the lighting condition in TV studio is quiet well, it does not have to take lighting compensation technique. In the first stage, an elliptical model, first introduced in [10], is applied for skin-tone detection on MPEG video. It is designed for face detection in color images, thus a linear transformation for the chrominance DC coefficients is needed for the coherence of the data range, which can be written as:

$$C_x' = \frac{C_x + 1024}{8} \qquad (1)$$

where $x$ denotes $b$ or $r$, $C_x$ denotes the DC values of Cb or Cr blocks, and $C_x'$ is the adjusted value of $C_x$. The elliptical model can be written as:

$$\frac{(x-e_x)^2}{a^2} + \frac{(y-e_y)^2}{b^2} <= 1 \qquad (2)$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} C_b' - c_x \\ C_r' - c_y \end{bmatrix} \qquad (3)$$

where $c_x$ = 109.38, $c_y$ = 152.02, $\theta$ = 2.53 radians, $e_x$ = 1.6, $e_y$ =2.41, a = 25.39 and b = 14.03.

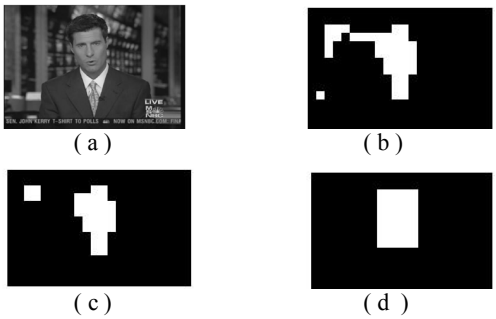


**Fig. 1**. Face detection results of every stage: (a) the original video frame, (b) the result of skin-tone detection, (c) the result of morphological operation and size filtering, (d) the result of shape constraint.

The pair (x, y) is leaded by the transformed chrominance DC coefficients $C_b'$ and $C_r'$, and if this pair is inside the ellipse, it indicates that the color corresponds to the skin color. In fact, the skin-tone detection is implemented in MPEG macroblock level, which reduces the spatial resolution of the video frames by 16 in both horizontal and vertical direction (for 4:2:0 chrominance format). This may influence the skin-tone detection in two aspects. One is that the small face is hard to detect, the other is that the detected face region is not well aligned with the actual face region. However, these problems have no influence for ASD since we only care about the presence or absence of the face in the candidate key frames.

In the second stage, morphological opening and closing with a structural element of $N$*$N$ ($N$ = 3) are applied in the binary skin-tone image to obtain smoothed homogeneous areas of connected pixels. Connected component labeling is then performed, and small areas less than eight pixels are removed.

Besides human faces, there are natural scenes with colors similar to skin colors in candidate key frames. Finally, to eliminate some of these false alarms, the binary template matching method in [8] is employed for shape constraint on the connected region generated from the stage two. The range of aspect ratio of the rectangle is set to [0.9, 2].

It has to be emphasized that the face detection result makes the upper bound of ASD. To ensure all the anchorperson key frames pass this module, the parameters in this subsection are all carefully set. An example for the results of every stage is illustrated in figure 1.

### 2. 3. Clustering Analysis

The input of this module is generally composed of anchorperson, reporter, interviewee key frames and some other key frames satisfying the first two modules. The aim of this module is to cluster the candidates by clustering techniques. A key step in a clustering method is to select a distance measure. In this module, we introduce a novel

metric in MPEG domain for further enhancing the execution speed. First, four features are extracted: the transformed luminance DC values $Y_{dc}'$ (transformation equation is similar to equation 1), the transformed chrominance DC values $C_b'$ and $C_r'$, and $C_d'$, which represents the ratio of the corresponding pixels of $C_b'$ and $C_r'$, that is, $C_d' = C_b'/ C_r'$. Then four distances, namely $D_y$, $D_b$, $D_r$, and $D_d$, between the input key frames are computed. For the $Y_{dc}'$ feature, the histogram distance $D_y$ between pairwise luminance DC images is defined as:

$$D_y(n) = \frac{1}{P \times Q}\left(\sum_{k=1}^{L} \left|H_u(k) - H_v(k)\right|\right) \qquad (4)$$

where $H_u(k)$ and $H_v(k)$ denote the gray-level histograms in the luminance DC images, $P \times Q$ is the size of a luminance DC image, $L$ is the possible gray levels, here we take $L$=256. While for the last three features, the spatial distances $D_b$, $D_r$, and $D_d$ are defined as:

$$D_x(n) = \frac{1}{M \times N}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} \left|I_u(i,j) - I_v(i,j)\right|\right) \qquad (5)$$

where $x$ denotes $b$, $r$, or $d$, $I_u$ and $I_v$ denote the chrominance DC images of $C_b'$, $C_r'$ or the generated images of $C_d'$ in a candidate I-frame, $M \times N$ denotes the size of a chrominance DC image.

For convenience of processing, we normalize the four distances into the interval [0, 1] respectively, then a final metric $D_f$ is defined as:

$$D_f(n) = \sqrt{a_1 D_b^2 + a_2 D_r^2 + a_3 D_d^2 + a_4 D_y^2} \qquad (6)$$

where $a_1 \sim a_4$ are the weight for each distance (experimental results have indicated a suitable ratio of 1:1:2:1).

By using the metric $D_f$, a hierarchical agglomerative algorithm is employed and a correspondent dendrogram is generated. However, it is difficult to determine the cluster numbers of the anchorperson key frames because of the diversity of contemporary news programs. Therefore, it is hard to apply a fixed threshold to select the anchorperson clusters from the clustering results.

To determine the adaptive threshold $Tc$, we employ the Fuzzy C-Means algorithm to cluster the linkage distances into $m$ clusters. The max distance value in the first cluster is chosen as the adaptive threshold (the first cluster is the one contains the smaller linkage distances than other clusters). It is reasonable for the way of $Tc$ selection since the linkage distances in the same anchorperson cluster are low, but relatively high between different clusters. In this paper, we take $m$=3. An example of the clustering dendrogram and the selection of adaptive threshold is illustrated in figure 2. We assume that an anchorperson appears at least twice in a news program. Then the clusters with the size greater or equal to two are selected as candidate clusters.

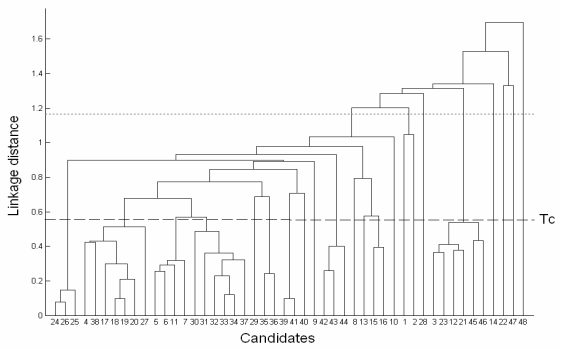


**Fig. 2**. Hierarchical clustering dendrogram and the adaptive threshold.

### 2.4. Final Decision

Actually, anchorperson clusters are not the only type in the candidate clusters, because some other kinds of shots appearing more than once will also be clustered. Thus, similar to the criteria proposed in [4], we apply three rules to remove the false alarms, which are:

1) The maximum temporal interval of the key frames in a cluster, named cluster lifetime ($C_{IT}$), should be larger than 50 seconds.
2) The maximum shot interval ($M_{SI}$) in a cluster should be larger than 19.
3) The average shot interval ($A_{SI}$) in a cluster should be larger than 16.

At last, those clusters with smaller $C_{IT}$, or $M_{SI}$, or $A_{SI}$ are removed as false alarms, and only the clusters satisfying all the three rules are identified as anchorperson clusters.

## 3. EXPERIMENTAL RESULTS

We test our approach on 18 MPEG-1 coded news videos from 6 different TV channels, which include 1 Lebanese channel, 2 U.S. ones, and 3 Chinese ones with the total duration of more than 9 hours. This is a comprehensive database, and the constitution is similar to the database provided in the TREC Video Retrieval Evaluation 2005. The detailed information of these news videos is summarized in Table 1 (three videos from each channel).

We employ the well-known criterion, i.e. precision (Pre.), recall (Rec.) and F-measure, to evaluate the performance of the proposed algorithm, and the experimental results are given in Table 2. Finally, we compare our algorithm with two unsupervised algorithms proposed in [2] and [3] respectively on the same dataset, as shown in Table 3. Our algorithm gives better performance than the algorithm in [2], and speedup of 1.5~2 times is observed. Under the similar performance of the algorithm proposed in [3], the runtime of ours is only its 20%~35%.

The high speed mainly lies in the face detector and savings in decoding.

There are mainly two types of errors. The first one is due to the great change of background and camera angle between anchorperson shots, which may cause much difference in the visual content. Consequently, this kind of anchorperson shot will be missed in our ASD algorithm. Figure 3 shows a missed anchorperson key frame and a detected anchorperson key frame in a complete MSNBC news program by comparison. The second type of error results in the appearing of the news footage shots with similar visual content in a news video. Figure 4 shows a pair of key frames which are falsely identified as anchorperson key frames in a BJTV news program (2003-06-24), appearing at the time of 13′28″ and 19′35″, with the shot number of 70 and 129 respectively.

**Tab. 1**. Detailed information about the news video programs

| Channel | Length | Shot | Anchorperson |
|---|---|---|---|
| CCTV | 01:30:00 | 1078 | 29 |
| BJTV | 01:16:42 | 669 | 42 |
| TJTV | 01:22:33 | 807 | 33 |
| MSNBC | 01:21:56 | 966 | 55 |
| CNN | 01:23:51 | 1123 | 38 |
| LBC | 02:15:36 | 1215 | 61 |
| Total | 09:10:58 | 5858 | 258 |

**Tab. 2**. Results of the proposed algorithm

| | Hit | Miss | False | Rec. (%) | Pre. (%) | F (%) |
|---|---|---|---|---|---|---|
| CCTV | 29 | 0 | 2 | 100 | 93.55 | 96.67 |
| BJTV | 42 | 0 | 2 | 100 | 95.45 | 97.67 |
| TJTV | 33 | 0 | 1 | 100 | 97.06 | 98.51 |
| MSNBC | 50 | 5 | 3 | 90.91 | 94.34 | 92.59 |
| CNN | 35 | 3 | 3 | 92.11 | 92.11 | 92.11 |
| LBC | 61 | 0 | 1 | 100 | 98.39 | 99.19 |
| **Average** | **-** | **-** | **-** | **96.90** | **95.42** | **96.15** |

**Tab. 3**. Performance Comparison

| | Rec. (%) | Pre. (%) | F (%) |
|---|---|---|---|
| Our algorithm | 96.90 | 95.42 | 96.15 |
| Gao and Tang [2] | 91.09 | 89.35 | 90.13 |
| Anna et al. [3] | 97.29 | 96.91 | 97.10 |

## 4. CONCLUSIONS AND FUTURE WORKS

In this work, an improved ASD algorithm in MPEG domain is proposed. Its contributions lie in the two ideas of the effective face detection and the new dissimilarity metric for clustering. Experimental results show its high performance over a comprehensive database. Future work will apply audio information to compensate for the limitation of the algorithm. Moreover, this algorithm will also be integrated into the technique of news video story segmentation and video semantic analysis.

(a) (b)

**Fig. 3**. A missed anchorperson key frame (a), and a detected anchorperson key frame (b) in a complete news video.

(a) (b)

**Fig. 4**. Two falsely detected key frames of anchorperson shots in a complete news video.

## 5. REFERENCES


[1] H.J. Zhang, Y Gong, S.W. Smoliar, and SY Tan. "Automatic Parsing of News Video," Proceedings of the International Conference on Multimedia Computing and Systems, Boston, pp. 45-54, 1994.

[2] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," IEEE Transactions on Circuits and Systems for Video, pp. 765-776, 2002.

[3] L.D. Anna, G.Marrazzo, G.Percannella, C.Sansone, and M.Vento, "A multi-stage approach for anchor shot detection," Joint IAPR International Workshops SSPR 2006 and SPR 2006, Hong Kong, pp. 773-82, 2006.

[4] D.J. Lan, Y.F. Ma, and H.J. Zhang, "Multi-level anchorperson detection using multimodal association", Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, pp. 890-893, 2004.

[5] H.Martin, H.G Kim and S.Thomas, "Audiovisual Anchorperson Detection for Topic-oriented Navigation in Broadcast News", IEEE 7th International Conference on Multimedia & Expo, Toronto, Canada, pp. 1817-1820, 2006.

[6] W.Q. Wang and W.Gao, "A Fast Anchor Shot Detection Algorithm on Compressed Video", IEEE Pacific Rim Conference on Multimedia, Beijing, pp. 873-878, 2001.

[7] Y. Avrithis, N. Tsapatsoulis, and S. Kollias, "Broadcast news parsing using visual cues: A robust face detection approach," 2000 IEEE International Conference on Multimedia and Expo, New York, pp. 1469-1472, 2000.

[8] H.L.Wang and S.F.Chang. "A highly efficient system for automatic face region detection in MPEG video," IEEE Transactions on Circuits and Systems for Video, pp. 615-628, 1997.

[9] T.S. Chua, Y.S. Zhao, and M.S. Kankanhalli, "Detection of human faces in a compressed domain for video stratification," The Visual Computer, pp.1-18, 2002.

[10] R.L.Hsu, M. Abdel-Mottaleb, and A.K. Jain, "Face detection in color images," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 696-706, 2002.