# Exploiting Visual-Audio-Textual Characteristics for Automatic TV Commercial Block Detection and Segmentation

Nan Liu, Yao Zhao, *Member, IEEE*, Zhenfeng Zhu, and Hanqing Lu, *Senior Member, IEEE*

***Abstract*—Automatic TV commercial block detection (CBD) and commercial block segmentation (CBS) are two key components of a smart commercial digesting system. In this paper, we focus our research on CBD and CBS by the means of collaborative exploitation of visual-audio-textual characteristics embedded in commercials. Rather than utilizing exclusively visual-audio characteristics like most previous works, an abundance of textual characteristics associated with commercials are fully exploited. Additionally, Tri-AdaBoost, an interactive ensemble learning manner, is proposed to form a consolidated semantic fusion across visual, audio, and textual characteristics. In order to segment a detected commercial block into multiple individual commercials, additional informative descriptors including textual characteristics are introduced to boost the robustness in the detection of frame marked with product information (FMPI). Together with the characteristics of audio spectral variation pointer and silent position, FMPI can provide a kind of complementary representation architecture to model the similarity of intra-commercial and the dissimilarity of inter-commercial. Experiments are conducted on a large video dataset from both China central television (CCTV) channels and TRECVID'05, and promising experimental results show the effectiveness of the proposed scheme.**

***Index Terms*—Commercial detection, commercial segmentation, multi-modal fusion, text detection, video analysis.**

## I. Introduction

As one of the most effective, pervasive, and popular means of promoting products or services, TV commercials have become an inescapable part of modern life, significantly influencing our work habits and other aspects of life. In essence, a commercial can be interpreted as a special TV program which attempts to communicate up-to-date "product" information to a tremendous number of consumers simultaneously and generate sustained appeal in their minds even long after the span of the commercial campaign. Therefore, tens of thousands of commercials are produced and broadcasted on many TV channels to promote a variety of new commodities or services.

Meanwhile, benefiting from the rapid development of multimedia acquisition and storage technologies, people can conveniently record more and more commercials through various multimedia devices for time-shifted easy access and consumption if necessary. However, owing to the insufficiency of effective video content analysis techniques, the explosive growth of recorded commercials results in critical demands for the actual applications of smart commercial digesting systems (CDS) for different user groups, such as TV viewers, media professionals, governing bodies, and advertisement companies. For the TV viewers who are likely to use the digital TV set-top boxes to record broadcast videos, it is deeply desirable to design a powerful CDS to help them remove commercial blocks from the general programs to maintain a normal watching mode. Furthermore, the CDS may assist media professionals in quickly retrieving, browsing, and indexing daily updated commercials for the purpose of information acquirement. As the crux of an effective CDS, commercial block detection and segmentation have drawn lots of attention in recent years and comparative efforts have been devoted to these areas.

### A. Commercial Block Detection (CBD)

Commercial block detection, which automatically detects commercial blocks from broadcast videos based on some intrinsic commercial characteristics, has become an indispensable component of a smart CDS. Generally, two key challenges related to CBD are: 1) designing the unique semantic descriptors to reflect the intrinsic characteristics of commercials versus general programs; and 2) developing an effective discrimination model based on the exploitation of these descriptors. Some previous studies treat CBD as a heuristic rule-based problem that a series of broadcast editing rules, such as the occurrence of black/silent frames among individual commercials [1], [2], the absence of TV station logos [3] or subtitles [4] in advertising time, and their combinations [5], are required to be available prior to performing detection. This approach, however, heavily depends on the specified rules and would fail in some countries, especially Asia, where few of these rules are used to indicate the start/end positions of commercial blocks. Furthermore, whereas some recognition-based methods [6]–[9] achieve excellent performance on CBD resorting to the robust video fingerprint and efficient indexing structure, their disadvantage

Manuscript received October 31, 2010; revised March 22, 2011 and June 06, 2011; accepted June 14, 2011. Date of publication June 23, 2011; date of current version September 16, 2011. This work was supported in part by 973 Program (No. 2012CB316401), National Science Foundation of China (No. 61025013, No. 61172129), Sino-Singapore JRP (No. 2010DFA11010), Beijing Natural Science Foundation (No. 4112043), and Fundamental Research Funds for the Central Universities (No.2009JBZ006-3). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhu Liu.

N. Liu, Y. Zhao, and Z. Zhu are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: 05112073@bjtu.edu.cn; yzhao@bjtu.edu.cn; zhfzhu@bjtu.edu.cn).

H. Lu is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: luhq@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2160334

1520-9210/$26.00 © 2011 IEEE

is fairly obvious in that they are unable to cope well with those unregistered commercials.

Aiming at alleviating these aforementioned problems, a more recent attempt, called learning-based method [10]–[17], has been developed by exploring various semantic characteristics associated with commercials versus general programs. Prior to automatic online commercial detection, a large number of training samples are utilized to learn a discrimination model. For instance, Hua *et al.* [13] and Zhang *et al.* [14] extracted some unique context-based audio-visual features from each shot considering the temporal information along with its contextual parts. Moreover, Mizutani *et al.* [15] fused audio/visual/temporal-based local features of commercials in the context of their global temporal characteristics to detect commercial blocks. To the best of our knowledge, few studies have paid attention to the exploitation of textual information appearing in the commercial frames, which is one of the most distinct characteristics to differ commercials from general programs. Although some plain cases of its applications, e.g., simply employing the locality information of text occurrence, have been proposed [15], [16], an in-depth research on mining this kind of intrinsic characteristic for characterizing commercials was rarely explored.

### *B. Commercial Block Segmentation (CBS)*

On the basis of CBD, commercial block segmentation refers to automatically segmenting the detected commercial block into several individual commercials, utilizing some diverse multimodal cues which are capable of demonstrating the essential differences between the intra-commercial and inter-commercial. However, it is definitely non-trivial to discover these cues for some countries, especially Asia, since there is seldom a clear semantic unit, such as the occurrence of black frames among the individual commercials, to tell where to locate the actual boundaries. On this point, how to extract them has posed a great challenge on the successful CBS scheme.

For the sake of the great difficulty in presenting some effective descriptors to characterize these intrinsic cues for CBS, only a few studies [18]–[20] have concentrated on this research. To determine the boundaries for each individual commercial, two kinds of mid-level descriptors, named audio scene change indicator (ASCI) and frame marked with product information (FMPI), respectively, were proposed by Duan *et al.* [18], [19] to extract the audio-visual discriminatory characteristics. Furthermore, Wang *et al.* [20] extended FMPI and ASCI to the task of segmenting multiple types of TV programs including commercial, news program, and sitcom, and presented a descriptor of textual content similarity (TCS) to model the intra-program similarity and inter-program dissimilarity in terms of automatic speech recognition (ASR) or machine translation (MT) transcripts.

However, the descriptor of TCS may not be suitable for CBS due to the following reasons. Firstly, the extensive use of background music in commercials may degrade the performance of ASR, causing numerous false alarms in the transcripts [21]. Secondly, the quality of MT transcripts is occasionally not satisfactory due to the bottleneck of multi-linguistic machine translation technique. Despite the fact that promising experimental results have been reported, there is still room left for exploiting some essential characteristics, such as textual information appearing in the commercial frames and other high-level audio spectral variation properties, which have not been fully considered but account for much of CBS.

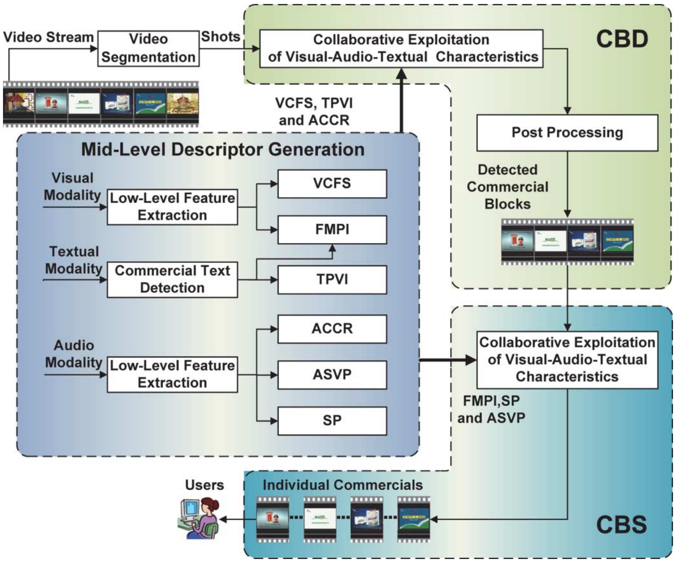

Video Stream
Video Segmentation
Shots
Collaborative Exploitation of Visual-Audio-Textual Characteristics
CBD
VCFS, TPVI and ACCR
Post Processing
Detected Commercial Blocks
Mid-Level Descriptor Generation
Visual Modality
Low-Level Feature Extraction
VCFS
FMPI
Textual Modality
Commercial Text Detection
TPVI
Audio Modality
ACCR
Low-Level Feature Extraction
ASVP
SP
Collaborative Exploitation of Visual-Audio-Textual Characteristics
FMPI,SP and ASVP
Users
Individual Commercials
CBS

Fig. 1. Proposed unified solution for CBD and CBS based on the collaborative exploitation of visual-audio-textual characteristics.

### *C. Our Solution for CBD and CBS*

The main goal of our research is to provide a unified solution for CBD and CBS by collaboratively exploiting the visual-audio-textual characteristics. The framework is illustrated in Fig. 1, demonstrating an integral scheme to robustly locate the overlay texts in commercials, detect commercial blocks from broadcast videos, and segment them into multiple individual commercials.

To address the aforementioned issues, the following points highlight several contributions of this paper:

- The extensive use of overlay texts in commercials is a type of unique characteristic for commercials versus general programs. However, their appearances are much more diversified than the close-captions in general programs, resulting in the great difficulty of automatically locating them. To pave the way for in-depth textual characteristics analysis for CBD and CBS, we present an enhanced co-training-based commercial text detection approach by interactively exploiting the intrinsic correlation of multiple texture representation spaces.
- Aside from exclusively employing the commonly used visual-audio characteristics in CBD, some intrinsic textual characteristics associated with commercials but rarely presented in general programs are fully exploited via analyzing the spatio-temporal properties of overlay texts in commercials. Moreover, Tri-AdaBoost, an interactive ensemble learning approach, is proposed to form a consolidated semantic fusion across the concurrent visual, audio, and textual characteristics.
- In order to segment the detected commercial block into multiple individual commercials, additional informative

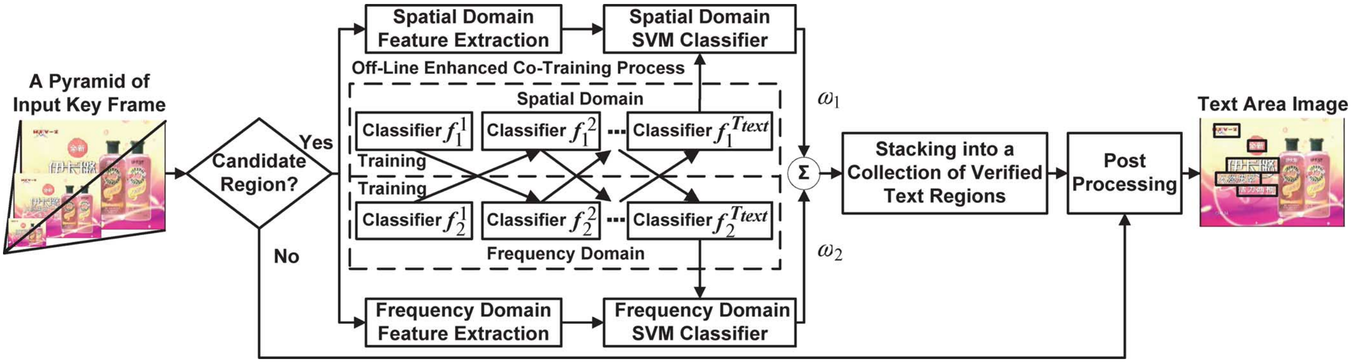

A Pyramid of Input Key Frame
Candidate Region?
Yes
No
Spatial Domain Feature Extraction
Spatial Domain SVM Classifier
Off-Line Enhanced Co-Training Process
Spatial Domain
Classifier $f_1^1$
Classifier $f_1^2$
Classifier $f_1^{T_{text}}$
Training
Training
Classifier $f_2^1$
Classifier $f_2^2$
Classifier $f_2^{T_{text}}$
Frequency Domain
Frequency Domain Feature Extraction
Frequency Domain SVM Classifier
$\omega_1$
$\omega_2$
Σ
Stacking into a Collection of Verified Text Regions
Post Processing
Text Area Image

Fig. 2. Proposed CTD scheme based on the enhanced co-training strategy.

descriptors including textual characteristics are introduced to boost the robustness of the FMPI detection proposed in [18] and [19]. Together with the characteristics of audio spectral variation pointer (ASVP) and silent position (SP), FMPI provides a kind of complementary representation architecture to describe the intrinsic characteristics of inter-commercial versus intra-commercial.

- Comprehensive experiments are conducted on a sizeable video data collection from China central television (CCTV) channels and TRECVID'05, indicating that the proposed unified solution for CBD and CBS is very convincing.

The remainder of this paper is organized as follows: Section II introduces the proposed method for commercial text detection. In Sections III and IV, we present the intrinsic visual-audio-textual characteristics associated with commercials as well as the means of collaborative exploitation of them for CBD and CBS, respectively. Section V presents the experimental results and performance analysis. Section VI concludes this paper.

## II. Commercial Text Detection

The use of the overlay texts in the salient areas of commercial frames is an unequaled semantic characteristic of commercials versus general programs. For the purpose of utilizing this kind of textual cue, the locality information of commercial texts needs to be acquired prior to forming the textual descriptor. But, the traditional text detection methods like the heuristic means [22] cannot be served as an effective solution for commercial text detection (CTD) due to the complex and diverse appearances of these texts. For instance, various kinds of text blocks appearing in commercials always hold different colors, font sizes, and slantwise directions and embed in much complex backgrounds. To circumvent these problems, we propose a CTD scheme as shown in Fig. 2 based on an enhanced co-training strategy.

### A. *Overview of the Proposed Scheme*

We pose CTD as a supervised texture classification problem. To boost the ability of discrimination of the text areas from complex backgrounds, an enhanced co-training strategy is introduced by exploiting the intrinsic correlation between two conditional independent texture representation spaces, i.e., frequency domain by wavelet transform and spatial domain based on gray-level co-occurrence.

As illustrated in Fig. 2, the proposed scheme utilizes a small sliding window with the size of 16 × 16 to scan the input frame. Then, the fused SVM classifier with enhanced co-training strategy is applied to classify the pixels located in the window into text or non-text area based on the analysis of their texture properties. Specifically, to detect the texts with various sizes, we generate a pyramid of frames from the original frame by gradually changing the resolution at each level. Then, the extracted text blocks of each level are refined based on some constraint conditions [23], [24], and further projected to the original scale to form the final detection result.

### B. *Enhanced Co-Training Strategy*

The co-training strategy [25] is a kind of algorithm that combines the multi-view and semi-supervised learning into one unified framework. Only based on small amount of labeled as well as a vast amount of unlabeled data, it provides an effective way to construct a robust classifier from two conditionally independent views of the data, which deliver diverse and complementary information for the sample. That is, two initial classifiers are separately built from each view and then updated incrementally in each iteration using the unlabeled examples with the highest prediction confidence in each view. Once the pre-defined iteration condition is satisfied, the weighted linear combination of the output classifiers will give a consolidated decision for the multi-view representation spaces.

Although the successful applications of the co-training algorithm have been reported in several aspects such as multi-modal fusion and web page classification, there still exist some disadvantages that may degrade the performance of the trained classifiers. That is, noisy samples will be brought into the training set due to the inevitable false alarms existed in instance labeling process in each iteration. In the case of CTD, the existence of a large number of indistinguishable samples, which are hard to be correctly classified as text or background areas by classifiers built in co-training process, may lead to bringing more noisy instances into training data collection. To alleviate this problem, we integrated the Bootstrap [26] algorithm into the co-training strategy. In other words, we select those misclassified negative samples, which are considered to be the most informative ones, as the incrementally labeled negative instances in each iteration, whereas no positive samples are selected in the iteration process. Moreover, the new collected samples are collaboratively provided to other views for the purpose of interactive learning. The

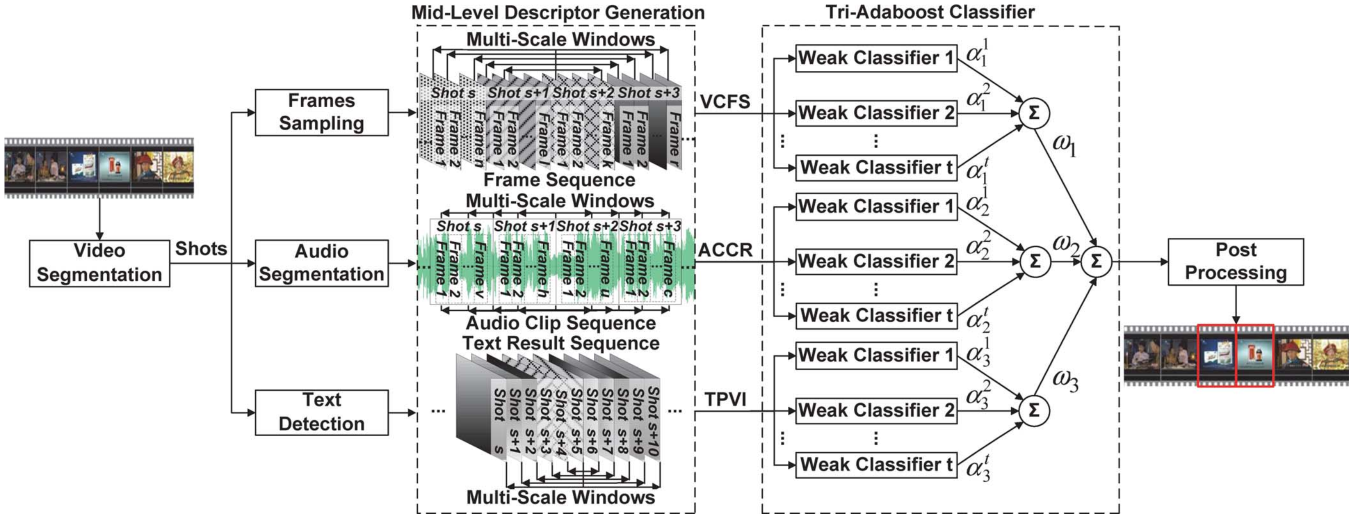


Fig. 3. Proposed CBD scheme based on the collaborative exploitation of visual, audio, and textual descriptors.

more detailed elaboration on the enhanced co-training strategy can be referred to Algorithm I.

ALGORITHM I: ENHANCED CO-TRAINING STRATEGY FOR CTD

```
Input: a set of labeled samples $L_1^0$ and $L_2^0$ with the subscripts denoting frequency domain and spatial domain, respectively, unlabeled sample set (without positive instances) $U$, the validation set $V$, and the maximum iteration number $T_{text}$.
Enhanced Co-Training Procedure:
1) Learn the classifiers $f_1^0$ and $f_2^0$ based on $L_1^0$ and $L_2^0$, respectively.
2) For $t = 1, 2, \ldots, T_{text}$
      2.1) Label $U$ by the classifiers $f_1^{t-1}$ and $f_2^{t-1}$ and collect the mislabeled samples $E_1^t$ and $E_2^t$ classified as positive instances, respectively.
      2.2) Update the training sets $L_1^t = L_1^{t-1} \cup E_2^t$ and $L_2^t = L_2^{t-1} \cup E_1^t$.
      2.3) Learn the classifiers $f_1^t$ and $f_2^t$ based on $L_1^t$ and $L_2^t$, respectively.
3) Calculate the precisions of $f_1^{T_{text}}$ and $f_2^{T_{text}}$ (denoted by $p_1$ and $p_2$, respectively) based on $V$.
4) Output the final combined classifier $f = \omega_1 f_1^{T_{text}} + \omega_2 f_2^{T_{text}}$ with weights $\omega_1 = (p_1 - 0.5)/(p_1 + p_2 - 1)$ and $\omega_2 = 1 - \omega_1$.
```

## III. COMMERCIAL BLOCK DETECTION

The proposed CBD method based on the collaborative exploitation of visual-audio-textual characteristics is illustrated in Fig. 3. The video stream is first segmented into a sequence of shots, from which three kinds of mid-level descriptors [i.e., visual change on frame sequence (VCFS), audio content consistency representation (ACCR), and text pattern variation indicator (TPVI)] are extracted to describe the intrinsic characteristics of commercials versus general programs. Then, to determine whether or not these shots belong to commercial, the Tri-AdaBoost strategy with consideration of exploiting the complementary semantics across visual, audio, and textual characteristics is investigated to form a consolidated semantic fusion of them. Ultimately, the postprocessing module is triggered out to further reduce the false alarms.

### A. Visual-Audio-Textual Characteristics Analysis

To make use of the intrinsic visual, audio, and textual characteristics of commercials versus general programs, a variety of mid-level descriptors are extracted from each shot by exploiting various spatio-temporal properties.

*1) Visual Change on Frame Sequence (VCFS):* Commercial can be interpreted as a sort of special TV program that attempts to communicate some commodity or service information over a considerably limited duration. As a result, the spatial and temporal variations in visual contents of commercials tend to be more drastic than those in general programs. Thus, accounting for these latent semantics, we propose a type of mid-level descriptor, VCFS, to delineate the local and global visual variations for each shot and its contexts.

As shown in Fig. 3, a series of key frames are equally sampled from each shot with 30-frame interval. Aiming at the construction of a salient descriptor, various local and global properties are exploited with $N_v$ frame-level multi-scale sliding windows. With respect to the local information, each frame is partitioned into $h \times v$ blocks and then a set of local properties, including HSV histogram (6 dimensions), edge change ratio [13] (2 dimensions), and gray-scale frame difference [13] (2 dimensions), are extracted from each block to form the local representation vector $V_l(t)$ of each key frame. In addition, the 18-dimensional global HSV histogram is adopted to construct the global vector $V_g(t)$. Then, the first- and second-order statistical moments of the variation on $V(t) = \{V_l(t),\ V_g(t)\}$ across multiple frames $V(t+u)$ contained in a sliding window with a certain scale are given as

$$C_1^v(t) = \frac{1}{2W_v} \sum_{u=-W_v}^{W_v} [V(t) - V(t+u)]$$

$$C_2^v(t) = \frac{1}{2W_v} \sum_{u=-W_v}^{W_v} [V(t) - V(t+u) - C_1^v(t)]^2 \quad (1)$$

where $W_v$ is half the size of the sliding window. For each shot, the VCFS descriptor is defined as the mean of $C_1^v(t)$ and $C_2^v(t)$ over all the $R$ key frames within a shot, and we have

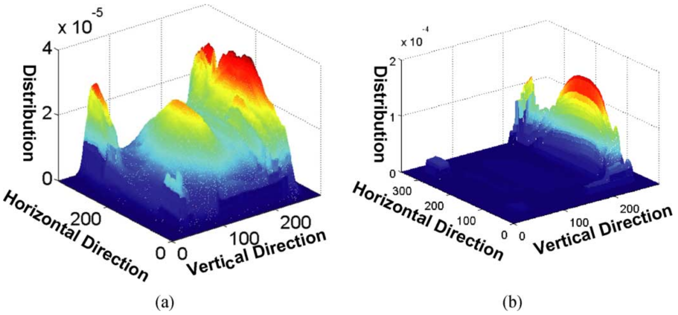
$$C^v = \left[ \frac{1}{R} \sum_{t=1}^{R} C_1^v(t), \quad \frac{1}{R} \sum_{t=1}^{R} C_2^v(t) \right]. \tag{2}$$

Thus, by integrating both local and global visual variation information, a 432-dimensional VCFS feature vector is obtained, given the empirical parameter setting as $h = v = 3$, $N_v = 4$, and $W_v \in \{2, 3, 4, 5\}$.

*2) Audio Content Consistency Representation (ACCR):* The audio modality of commercials is another informative cue to differing them from general programs. That is, many commercials feature songs or melodies that generate sustained appeal, which may remain in the minds of TV viewers even long after the span of the commercial campaign. Although some entertainment programs show similar characteristics, the transfer frequency among different kinds of voice (e.g., male to female, music to speech) is considerably drastic. Thus, we use ACCR for each shot to characterize the spectral-temporal audio variations.

For each visual shot, the audio stream is first segmented into a sequence of 20-ms-long non-overlapping frames and then every 50 frames are combined into a 1-s length of non-overlapping audio clip, from which a collection of features $A(t)$ such as 60-dimensional timbre texture features [27] (e.g., spectral centroid, flux, MFCC), 28-dimensional psychoacoustic features [28] (e.g., spectral flatness, sharpness), and other 20-dimensional low-level features [28] (e.g., bandwidth, fundamental frequency), are extracted to provide a semantic interpretation of the audio contents. Similar to the VCFS, the first- and second-order statistical moments $C_1^a(t)$ and $C_2^a(t)$ of the variation on $A(t)$ are calculated according to (1). Then, as in (2), the 432-dimensional ACCR vector $C^a$ [the mean of $C_1^a(t)$ and $C_2^a(t)$ over all audio clips in a shot] can be formed. In this case, $R$ denotes the total number of audio clips within a shot and $W_a$, half the size of sliding window, is empirically set to be {5, 10, 15, 20}.

*3) Text Pattern Variation Indicator (TPVI):* As we know, commercial is one of the most important media forms to convey commodity, service provision, or brand information to consumers. For the purpose of reinforcing impression of the promoted products through the advertisements, a large number of text blocks, such as brand names and catch-phrases, are presented in the salient areas for a rather limited time to highlight their names or functions. But these texts are extremely unwonted in the majority of general programs, except some close-captions that appear in the bottom of the frames as shown in Fig. 4. Even though there are a certain number of texts to appear in the central areas in some news programs, their duration is also much longer than that in commercials because the TV viewers need sufficient time to catch their meaning along with the contextual contents of the news. As a result, the occurrence frequency of text blocks can be reasonably taken to form an effective descriptor to discriminate commercials from general programs. In addition, the variation patterns of text blocks in commercials are usually more complex than those in general programs, in terms of the occurrence location, font size, and orientation (see Fig. 5). However, to the best of our knowledge, few works have concentrated on the in-depth research on this type of textual information in CBD, except some simple applications [15], [16].



(a)

(b)

Fig. 4. Statistical distribution of text area positions for commercial and general program. (a) Commercial. (b) General program.

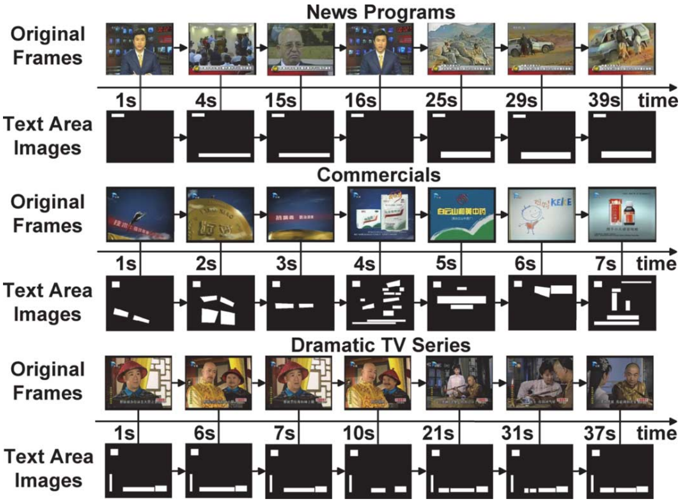


Fig. 5. Different variation patterns of text blocks in commercials, news programs, and dramatic TV series.

To make insight into this kind of textual characteristic, a novel textual descriptor, named TPVI, is proposed to extract the occurrence frequency and variation patterns of the overlay texts in commercials. To construct TPVI, we first employ our proposed CTD scheme to obtain a binary text area image $I(x, y, t)$ for each key frame (see Fig. 6). Note that the key frame here is simply chosen as the middle frame of each shot rather than the multiple key frames as mentioned in VCFS. Then, based on the statistical analysis of the set of binary images as shown in Figs. 4 and 5, five kinds of significant features are introduced, employing $N_t$ shot-level multi-scale sliding windows to represent the frequency and variation patterns of the overlay texts in commercials.

*Ratio of Text Area (RTA, 62 Dimensions):* The weighted ratio of the text areas to the whole video frame can be taken as an important indicator of the quantity of text occurrence, which is given by

$$P(t) = \frac{1}{M \times N} \sum_{x=1}^{N} \sum_{y=1}^{M} I(x, y, t) e^{-2 \max(|x - x_c|/N, |y - y_c|/M)} \tag{3}$$

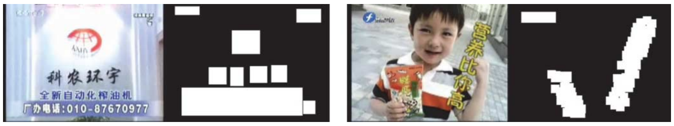


Fig. 6. Examples of text area images of commercial key frames.

where $M$, $N$, and $(x_c,\ y_c)$ are the size and center of $I(x,y,t)$, respectively. Then, we utilize the RTA, which consists of the temporal density ($td$) and the variance in unit time ($tv$) of the weighted ratio of text blocks appeared in a sliding window, to reflect the text occurrence frequency characteristic. They are defined as

$$td = \frac{\sum_{t=-W_t}^{W_t} P(t)}{\sum_{t=-W_t}^{W_t} len(t)}$$

$$tv = \frac{\sum_{t=-W_t}^{W_t} \left[P(t) - \frac{1}{2W_t}\sum_{t=-W_t}^{W_t} P(t)\right]^2}{\sum_{t=-W_t}^{W_t} len(t)} \tag{4}$$

where $W_t$ is half the size of the sliding window and $len(t)$ is the length of each shot within the window.

*Local Text Area Indicator (LTAI, 320 Dimensions):* Considering the local distribution information of text areas, we partition $I(x,y,t)$ into $r \times c$ blocks. Then LTAI can be defined as the ratio-based temporal density and variance in unit time of each block over multi-scale sliding windows.

*Text Block Frequency (TBF, 30 Dimensions):* In essence, the TBF is the same indicator as RTA but with different granular representation; in other words, the quantity of text blocks in a key frame is used to substitute for $P(t)$ in (4).

*Text Orientation Histogram (TOH, 33 Dimensions):* The moment of inertia [29] is first applied to calculate the orientation of each text block. Then we employ a histogram with three bins, which correspond to horizontal, vertical, and slantwise directions, respectively, to delineate the orientation distribution of the text blocks appeared in each key frame within a sliding window. Subsequently, the TOH is characterized as the temporal density and the variance in unit time for each bin over the multiple histograms.

*Randomness of Text Occurrence (RTO, 10 Dimensions):* As clearly shown in Fig. 5, the occurrence patterns of text blocks in commercials are revealed to be more random compared with those in general programs. Thus, the randomness of text occurrence can be described as

$$R = \frac{1}{M \times N} \times \sum_{u=-W_t}^{W_t} \sum_{x=1}^{N} \sum_{y=1}^{M} \mathrm{sgn}[I(x,y,t) - I(x,y,t+u)] e^{-\alpha|u|} \tag{5}$$

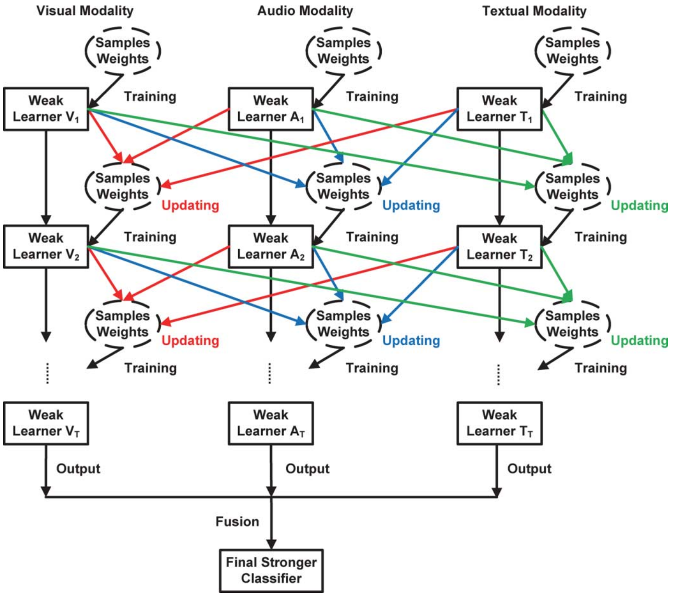


Fig. 7. Interactive training process of Tri-AdaBoost across visual, audio, and textual characteristics.

where $\mathrm{sgn}(x)$ denotes the binary sign function and $\alpha$ is a scale parameter with the constraint of $\sum_{u=1}^{W_t} e^{-\alpha|u|} = 1$.

Finally, based on the five kinds of features described above, we obtain a combined 455-dimensional TPVI descriptor with the empirical parameter setting as $r = c = 4$ and $N_t = 10$. Particularly, two types of sliding windows are used for the key frame sequence. The first one is based on the total number $W_t$ of the key frames contained in the sliding window and $W_t \in \{2,3,4,5\}$ in our experiments; the other one is simply based on the temporal duration from 5 s to 30 s with 5-s interval. The goal of employing the second type of sliding window is to avoid the effect of the possible absence of text blocks in some commercial shots.

## B. Tri-AdaBoost Fusion Strategy

As one of the most popular ensemble learning algorithms, AdaBoost [30] is designed to sequentially select a set of weak learners to form a strong classifier and maintain a distribution or set of weights over the training set. All weights are initially set to be uniform, but in each iteration, the weights of the incorrectly (correctly) classified examples are increased (decreased) by a penalty item so that the weak learner is forced to put more attention on the "hard" examples in the training set.

Considering the concurrent visual, audio, and textual characteristics associated with commercials, the proposed Tri-AdaBoost as in Fig. 7 refers to the interactive ensemble learning process across multiple modalities to form a consolidated semantic fusion by interactively exploiting the complementary semantics embedded in these characteristics, which is the key difference from the traditional AdaBoost strategy. The pseudo-code for Tri-AdaBoost is shown in Algorithm II. It is noteworthy that Zhou *et al.* [31] have presented a similar Tri-Training strategy for semi-supervised learning, but their work concentrated mainly on selecting unlabeled instances in an interactive voting manner.

ALGORITHM II: TRI-ADABOOST FOR COLLABORATIVELY EXPLOITING THE VISUAL-AUDIO-TEXTUAL CHARACTERISTICS

```
Input: a set of training samples $S = \{(x_{ji}, y_i)\}_{i=1,\ldots,m}$
and validation set $V = \{(x_{jl}, y_l)\}_{l=1,\ldots,n}$, where
$j = 1, \ldots, d$ denotes the different modalities and
$y_i \in \{-1, 1\}$. $T$ denotes the maximum iteration number.
Tri-AdaBoost Procedure:
  1) Initialize the weights of training samples for each
     modality by $D_{ji}^0 = 1/m$.
  2) Do for each iteration $t = 1, 2, \ldots, T$
        2.1) Do for each modality $j = 1, 2, \ldots, d$
        — Utilize the WeakLearn to train the weak learner $h_j^t$.
        — Calculate the error rate of $h_j^t$ by
          $\varepsilon_j^t = (1/2) \sum_{i=1}^{m} D_{ji}^{t-1} |h_j^t(x_{ji}) - y_i|$.
        — Set the coefficient by $\alpha_j^t = (1/2) \ln[(1 - \varepsilon_j^t)/\varepsilon_j^t]$.
        2.2) Do for each modality $j = 1, 2, \ldots, d$
        — Update the weights by $D_{ji}^t = D_{ji}^{t-1} \exp\{-s_{ji}^t \alpha_j^t y_i h_j^t(x_{ji})\}$, where
          $s_{ji}^t = 2^{r_{ji}^t - d}$,

              $$r_{ji}^t = \underset{k=1,\ldots,d}{Vote} \, [y_i h_k^t(x_{ki}) == y_i h_j^t(x_{ji})]$$

          and $Vote$ is a counting function.
        — Normalize the weights $D_{ji}^t = D_{ji}^t / Z_j^t$, where
          $Z_j^t$ is a normalization factor.
  3) Do for each modality $j = 1, 2, \ldots, d$
        3.1) Construct the ensemble of $h_j^t$

     $$H_j^T = \sum_{t=1}^{T} \left( \frac{\alpha_j^t}{A_j} \right) h_j^t(x_j), \quad \text{where} \quad A_j = \sum_{t=1}^{T} \alpha_j^t$$

        3.2) Calculate the error rate of $H_j^T$ based on $V$

          $$E_j = \frac{1}{2n} \sum_{l=1}^{n} |H_j^T(x_{jl}) - y_l|$$

  4) Output $H = \sum_{j=1}^{d} \omega_j H_j^T$, where $\omega_j = E_j^{-1} / \sum_{k=1}^{d} E_k^{-1}$.
```

In order to convey the complementary semantics across multiple modalities into the ensemble learning process, we introduce a scale factor $s_{ji}^t$ into the penalty item of the $i$th training sample on the $j$th modality for the weight $D_{ji}^t$. That is, the penalty degree of the weight is controlled by the scale factor $s_{ji}^t$ which is determined by the number $r_{ji}^t$ of all weak learners $h_k^t$ ($k = 1, 2, 3$) achieving the agreement with $h_j^t$ at the $t$th iteration. With exploitation of complementary semantics, we can select the "hardest" samples that are misclassified by $h_j^t$ ($j = 1$, 2, 3) from the training set and place bigger weights on them. Thus, the new trained weak learners will be forced to focus on the "hardest" examples in the next iteration to achieve better generalization ability. Note that the *WeakLearn* mentioned in Algorithm II denotes the selection strategy for the optimal weak learner that can minimize the error rate $\varepsilon_j^t$ based on $D_{ji}^{t-1}$.

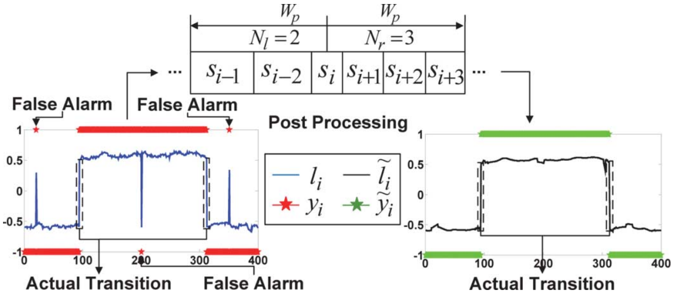


Fig. 8. Postprocessing for refinement of Tri-AdaBoost-based commercial shot classification.

### C. Postprocessing

Although a variety of semantic mid-level descriptors and an effective fusion strategy are proposed in our scheme, the false alarms are still inevitable due to the influence of the commercial shots without typical advertisement characteristics. Nevertheless, the nature that a certain number of commercials are broadcasted sequentially to form a commercial block provides an effective solution for us to refine the classification results by maintaining the temporal consistency of the occurrence of commercial shots in these blocks.

Fig. 8 illustrates the postprocessing for refinement of Tri-AdaBoost-based commercial shot classification. Suppose that $x_i$ denotes the unified representation of visual-audio-textual descriptors for each testing shot $s_i$, the classification likelihood score $l_i \in [-1, 1]$ and corresponding label $y_i$ are given by $H(x_i)$ and $\text{sgn}(l_i)$, respectively. To further reduce the potential false alarms in a sequence of shots $S = \{s_i | i = 1, \ldots, n\}$, the postprocessing procedure for $s_i$ is triggered by calculating the average score $\tilde{l}_i$ as

$$\tilde{l}_i = \frac{1}{(N_r + N_l)} \left( \sum_{t=1}^{N_l} l_{i-t} + \sum_{t=1}^{N_r} l_{i+t} \right) \tag{6}$$

where $N_l$ and $N_r$ are the total number of the shots contained in the left and right half side of the sliding window with $W_p$ seconds. Then the re-evaluated label $\tilde{y}_i$ for $s_i$ is determined by $\text{sgn}(\tilde{l}_i)$. Particularly, to avoid eliminating the actual transition between commercials and general programs, a constraint condition $\text{sgn}(\sum_{t=-N_l}^{-1} y_{i+t}) == \text{sgn}(\sum_{t=1}^{N_r} y_{i+t})$ is considered for $s_i$ prior to calculating $\tilde{l}_i$. If this constraint condition is not satisfied, $s_i$ is directly thrown into a candidate sequence $C = \{s_{i-u}, \ldots, s_i, \ldots, s_{i+v}\}$ comprised of its all successive shots that also violate the constraint condition. Then we can determine the actual transition position $[p^*, p^* + 1]$ in the candidate sequence based on $p^* = \arg\max_p(|l_p - l_{p+1}|)$, where $p \in [i - u, i + v - 1]$. For each shot $s_j$, $j \in [i - u, p^*]$, the $\tilde{l}_j$ can be given by $(1/(p^* - i + u + 1)) \sum_{k=i-u}^{p^*} l_k$ and $\tilde{l}_j = (1/(i + v - p^*)) \sum_{k=p^*+1}^{i+v} l_k$ for $j \in [p^* + 1, i + v]$.

## IV. COMMERCIAL BLOCK SEGMENTATION

Given the commercial blocks obtained through CBD, they need to be further segmented into multiple individual commercials. Aimed at providing a solid basis for further individual

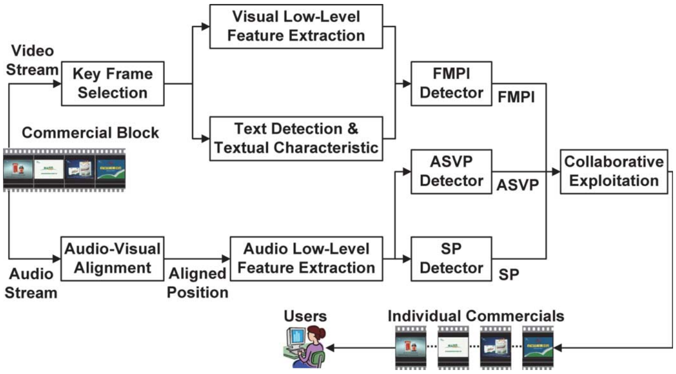
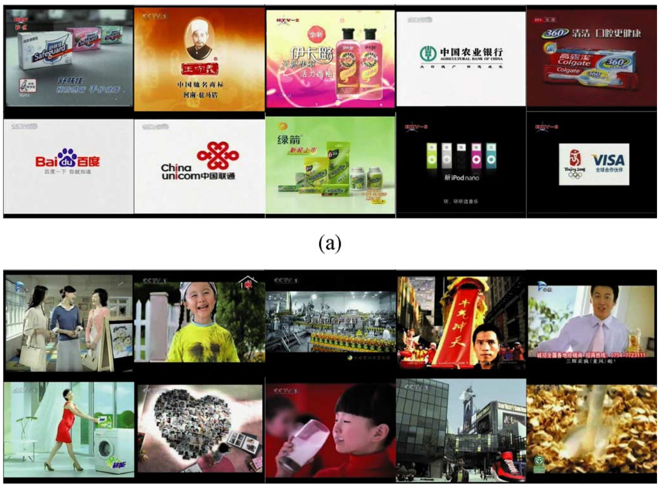


Fig. 9. Proposed CBS scheme based on the collaborative exploitation of the intrinsic visual-audio-textual characteristics.



(a)

(b)

Fig. 10. FMPI and general commercial frames. (a) Some examples of FMPI frames in commercials. (b) Several instances of general commercial frames.

commercial retrieval, the collaborative exploitation of the intrinsic visual-audio-textual characteristics for CBS is proposed as in Fig. 9.

### A. Visual-Audio-Textual Characteristics Analysis

To take advantage of these intrinsic characteristics for modeling the similarity of intra-commercial and the dissimilarity of inter-commercial, we propose a variety of unique semantic descriptors by collaboratively exploiting the intrinsic visual, audio, and textual semantic cues.

*1) FMPI Frame Detection:* FMPI frame is a sort of special frame to highlight the promoted “product” information in commercials by leveraging the close-up on products or company logos accompanying with simple even monochromatic background and large number of overlay texts for reinforcing the TV viewers’ appeal of the names or features of these products. To understand the FMPI frame better, we show in Fig. 10(a) several examples of FMPI frames, which demonstrate the explicitly different appearances from those in general commercial frames [see Fig. 10(b)].

Indubitably, the presence of FMPI frames, which generally appears around the end of individual commercials, can be reasonably taken to form an effective descriptor for CBS [18], [19], which indicates a series of potential positions for the boundaries of individual commercials. To utilize this kind of semantic information, Duan *et al.* [18], [19] resorted to the combination of texture, edge, and color features to represent an FMPI frame. Although their method achieved promising detection results, it only concentrated on some visual features. As one of the most distinct characteristics, the textual semantic information was not considered in their work. Accordingly, in order to reinforce the discrimination ability of FMPI frames, we propose an enhanced FMPI representation by exploiting the intrinsic visual and textual characteristics.

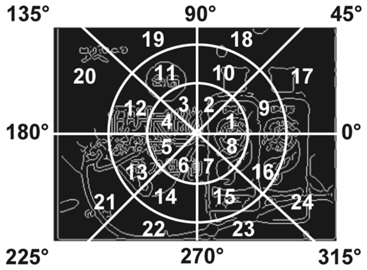


Fig. 11. Illustration of polar coordinate-based frame partition strategy for the 24-bin polar-like histogram.

*Text Descriptor:* It is worth noting that the statistical information of the overlay texts in FMPI frames is one of the most significant characteristics, which has potential ability to improve the performance of FMPI detection, but it was not fully considered in [18] and [19]. As shown in Fig. 10, the overlay texts always appear in the central area of FMPI frames. Thus, their occurrence positions can provide us an explicit semantic cue to discriminate FMPI frames from general ones. To utilize this kind of textual information, we first resort to the binary text area image generated by our proposed CTD method for each key frame to obtain the locality information of these texts. Then, the weighted ratio of the text areas to the whole video frame [see (3)], the quantity of text blocks, as well as the mean and variance of the distance and angle of each text block to the center of the video frame are extracted to describe the global distribution information. Regarding the local property, we partition $I(x,y,t)$ into $4 \times 4$ blocks and employ the ratio of the text areas to each block to form the spatial representation of text areas.

*Edge Descriptor:* An edge direction histogram of 8 bins is employed to record the global edge directions quantized at $\pi/4$ interval. Moreover, to reinforce the representation of edge locality information [18], [19], the edge image is divided into $3 \times 8$ regions according to the distance and angle of the pixels to the center of the video frame (see Fig. 11). For each region, the edge density is extracted to form a 24-bin polar-like histogram to reflect the spatial distribution of edge points.

*Corner Descriptor:* The corner descriptor, which was also neglected in [18] and [19], is a type of intrinsic semantic cue for characterizing the details of FMPI frames. Specifically, the 24-bin polar-like histogram is employed based on the corner detection results.

*Color Descriptor:* We exploit the dominant color features as presented in [18] and [19] to describe the color coherence information of the frame as well. In addition, the first $m$ maximum bins of the histogram $H_{sort}$ with $\sum_{i=1}^{m} H_{sort}(i) \geq 0.9$ is selected as a compact representation to describe the global color distribution information.

*Texture Descriptor:* The slight difference in contrast to the texture descriptor mentioned in [18] and [19] is that, in addition to utilizing the mean magnitudes of Gabor transform coefficients, the magnitudes variance of them are also considered to describe the homogeneity of texture.

Given the above five kinds of intrinsic textual and visual characteristic descriptions for FMPI frames, we can obtain a combined 266-dimensional FMPI descriptor $F_{FMPI}$. We straightforwardly apply the SVM classifier to form a final decision $p(F_{FMPI})$ for the $F_{FMPI}$ generated from each shot.

*2) Audio Spectral Variation Pointer (ASVP):* As we know, in order to generate more impressive appeal in the TV viewers' mind from the huge collection of advertisements, different commercials generally hold many distinct audio characteristics, such as the use of diverse background music, sound effects, and voice-over narrations. These intrinsic characteristics directly lead to the drastic spectral variations between two connection commercials, but keep consistent in each commercial. Consequently, this kind of audio spectral variation (ASV) characteristic associated with the transitions between two adjacent commercials can provide us another cue to segment commercial block into multiple individual commercials.

To determine the audio content change positions, the audio descriptor ASCI has been proposed in [18] and [19] resorting to hidden Markov model (HMM) based on some simple audio features. However, these plain audio representations cannot exactly describe the complicated spectral variations in commercials, e.g., the transfer between different background music. Consequently, we propose a new descriptor, named ASVP, to delineate the spectral content variations around the boundary for each individual commercial. In order to construct the salient audio representation, we utilize a collection of high-level spectral features, i.e., $A(t)$ in ACCR, which are widely used in music information retrieval [27], [28], to describe the audio spectral contents. We employ these features only to represent the spectral change between adjacent shots, not aiming at reflecting the change frequency in a sliding window.

Due to the video production rules and accumulated delay brought by video acquirement devices, there commonly exists an offset between the ASV and its associated video scene change at most TV commercial boundaries. Thus, the audio-visual alignment [18], [19] needs to be implemented beforehand for ASVP to find the most likely ASV position within the neighborhood of a shot change point. In spite of that the HMM decision-based audio-visual alignment has been introduced by [18] and [19], its computational burden is obviously heavy due to the use of two competitive HMMs for every candidate position. To reduce the computational complexity, we propose a simple but efficacious audio-visual alignment method by seeking the maximum spectral variation position around the shot boundary.

For each visual shot, the audio stream is segmented into a sequence of 20-ms-long frames with an overlapping interval of 10 ms. Then, a set of multi-scale sliding windows based on the temporal duration from 200 ms to 1 s with 200-ms interval are employed for each frame to seek the most likely ASV position $t^* \in [shot\_end, shot\_end + 1.5]$, which is determined by

$$t^* = \arg\max_{t} \sum_{i=1}^{n} \sum_{j}^{m} \|A_i(j, t+w) - A_i(j, t-w)\|_2 \quad (7)$$

where $A_i(j, t \pm w)$ are the $j$th dimension of the spectral features extracted from the right or left side of the $i$th sliding window at frame $t$ with the size $2w$. Whereas, $m$ and $n$ are the total number of the dimensions of spectral features and multi-scale sliding windows, respectively.

Given the alignment position $t^*$, we further utilize a 1-s-long symmetric window to characterize the spectral variations $F_{ASV} = \|A_{sw}(t^* + 0.5) - A_{sw}(t^* - 0.5)\|_2$ around the most likely ASV position. The ASVP is defined as the prediction probability $p(F_{ASV})$ of a pre-built SVM classifier.

*3) Silent Position Detection:* The occurrence of silent audio frames around the commercial boundaries can be reasonably taken as another essential characteristic for CBS. That is, the silent frames always appear at the end of commercials in most cases owing to the audio-visual asynchrony, although they are extremely undesirable around the intra-commercial shot boundaries, because the extensive use of music throughout commercial reduces the possibility of silence occurring. To construct SP, the silence detection is performed on the audio frame sequence around each shot boundary by comparing the zero crossing rate (ZCR) and short time energy (STE) of each frame with the pre-defined thresholds. Once ZCR and STE are both lower than the pre-defined thresholds, the frame is classified as silence. If the shot contains at least one silent frame around the boundary, the descriptor of SP $p(F_{SP})$ for it will set to be 1; otherwise, $p(F_{SP}) = 0$.

### *B. Collaborative Exploitation of Visual-Audio-Textual Characteristics for CBS*

In some cases, we should note that FMPI, ASVP, and SP might not simultaneously appear in the last shot of an individual commercial, but may gradually occur in a certain range around the end of commercial. Thus, in order to collaboratively exploit these characteristics, we employ a symmetrical sliding window to explore the intrinsic correlation among FMPI, ASVP, and SP for each shot and its contexts. That is, besides utilizing $p(F_{FMPI})$, $p(F_{ASV})$, and $p(F_{SP})$ of each shot, the mean and variance of these three characteristics of its neighborhood shots, which are contained in both the left and right sides of the window, are, respectively, extracted to form a 15-dimensional combined descriptor. Then we simply exploit the SVM classifier to form a consolidated semantic decision for the combined descriptor. Particularly, the size of the temporal duration-based sliding window is empirically set to be 4.5 s in our experiments according to the fact that the minimum length of individual commercials generally might not be less than 5 s.

## V. Experimental Results and Analysis

Owing to the lack of benchmark dataset for CBD and CBS, commercial data from TRECVID'05 have been used for per-

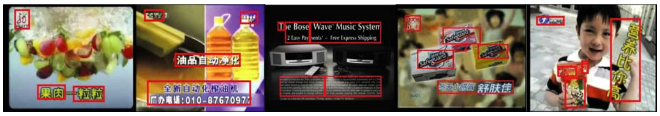


Fig. 12. Some results of the proposed CTD.

TABLE I
CTD PERFORMANCE COMPARISONS

| Methods | Totally Detected | Correctly Detected | $P_t$ | $R_t$ |
|---|---|---|---|---|
| The method in [24] | 3321 | 2126 | 64.02% | 79.18% |
| The proposed method | 2959 | **2396** | **80.97%** | **89.24%** |

formance evaluation [18]–[21]. Although TRECVID'05 is a good open video corpus for video search task, it only comprises small number of commercials, which leads to the limitation on making fair performance comparisons. In order to make the experiment analysis more convincing, we have collected from CCTV channels 20.3 h of video data with around 1.8 million frames, which contain 634 individual commercials covering 509 different ones. To the best of our knowledge, it might be the biggest dataset to date for the research on commercial analysis. Particularly, the quantity of different commercials is more than two times as compared to the one in [18]–[21] (499 individual commercials covering 191 different ones).

### A. *Performance of CTD*

To evaluate the performance of the proposed CTD method, benchmark comparisons with [24] are carried out on a dataset of 600 key frames containing 2685 text blocks. Two metrics are used to validate the CTD:

$$P_t = \frac{\#correctly\ detected\ text\ blocks}{\#\ detected\ text\ blocks}$$
$$R_t = \frac{\#\ correctly\ detected\ text\ blocks}{\#\ actual\ text\ blocks}. \quad (8)$$

As clearly shown in Table I, the proposed scheme is far better than [24] due to the interactive exploitation of the intrinsic correlation between different representation spaces. Several examples of CTD results are shown in Fig. 12; it is obvious that the proposed scheme is capable of handling some complicated cases like the multilingual texts and slantwise directions of text blocks.

### B. *Empirical Evaluation on CBD*

To verify the performance of the proposed CBD scheme, we select 8.6 h of videos containing 8723 shots as training set, 3.4 h of videos (3706 shots) for validation, and the remainder (8.3 h with 7424 shots) as testing set. Specifically, there are a total of 2359 commercial shots and 5065 general program shots in the testing set. The diverse types of the general program include news, sports, dramatic TV series, and entertainment programs. The measures *Precision*, *Recall*, *Accuracy*[13]–[16] and $F1$ [18], [19] are utilized to evaluate the performance.

*1) Performance Analysis of Tri-AdaBoost:* To obtain the optimal value for the iteration number $T$ in Tri-AdaBoost, a validation set with 1565 positive and 2141 negative samples is used to make an observation on the error rate with the variation of $T$. It is clear from Fig. 13 that the error rates for both AdaBoost and Tri-AdaBoost tend to be stable around $T = 200$. Thus, considering the tradeoff between the classification accuracy and computational burden in practical application, we set $T = 200$ in the following experiments. Meanwhile, with $T = 200$, the corresponding weights $\omega_j$'s for $H_j^T$'s mentioned in Algorithm II are 0.30 (visual), 0.43 (audio), and 0.27 (textual), respectively.

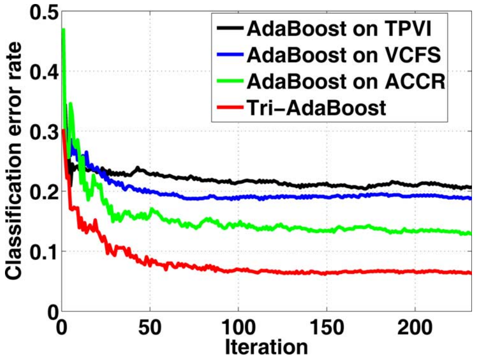


Fig. 13. Error rates of *Tri-AdaBoost* and *AdaBoost* with iteration $T$ on the validation set.

TABLE II
PERFORMANCE COMPARISONS FOR ADABOOST AND TRI-ADABOOST

| Classification Methods | *F1* | *Precision* | *Recall* | *Accuracy* |
|---|---|---|---|---|
| AdaBoost on VCFS | 75.41% | 76.53% | 74.32% | 80.81% |
| AdaBoost on ACCR | 84.14% | 85.13% | 83.18% | 86.72% |
| AdaBoost on TPVI | 75.28% | 65.03% | 89.36% | 79.03% |
| Tri-AdaBoost | **92.06%** | **88.88%** | **95.47%** | **94.92%** |

Table II shows the performance comparisons for AdaBoost and Tri-AdaBoost on the test set. It is evident that the *Tri*-AdaBoost convincingly outperforms all original AdaBoost methods on individual characteristics, taking advantage of the intrinsic semantic correlation across visual, audio, and textual characteristics. Moreover, we can see that TPVI achieves better performance in *Recall* compared with VCFS and ACCR, demonstrating the effectiveness of the essential characteristics of the texts appearing in commercials. But it fails in *Precision* and *Accuracy* due to the performance degeneration of CTD with some false alarms occurred in the text area images of general programs.

Since the types of general programs vary greatly, we also perform an experiment on the effectiveness of the proposed visual, audio, and textual descriptors and the collaborative exploitation of them, discriminating commercials from all different kinds of general programs. It can be seen from Table III that the proposed scheme achieves promising performance on the first three types of general programs. But for entertainment programs, TPVI almost completely fails to make the correct responses. It is more likely that entertainment programs take on many more similar characteristics for commercials, such as drastic camera motion and the use of music and slogan. In addition, the imperfect performance of CTD is another factor leading to the serious degeneration of TPVI.

*2) Performance Evaluation on Proposed CBD:* A felicitous sliding window size $W_p$ in the postprocessing has much to do

TABLE III
*PRECISION* EVALUATION ON ADABOOST AND TRI-ADABOOST FOR DIVERSE GENERAL PROGRAMS

| Types | Total Number | AdaBoost on VCFS | AdaBoost on ACCR | AdaBoost on TPVI | Tri-AdaBoost |
|---|---|---|---|---|---|
| Dramatic TV series | 2663 | 87.57% | 93.54% | 86.07% | **97.26%** |
| News | 1218 | 92.86% | 95.16% | 93.68% | **96.80%** |
| Sport | 536 | 95.34% | 97.57% | 94.03% | **98.13%** |
| Entertainment | 877 | 88.03% | **88.71%** | 25.43% | 81.76% |

with the performance of CBD. Larger $W_p$ will lead to the involvement of more unwanted noise in temporal analysis; likewise, smaller $W_p$ will cause less contextual information to be collected. Fig. 14 shows the curve of *Accuracy* versus $W_p$ on the validation set, from which we can find a turning point around $W_p = 40$ s, and thus, we set $W_p = 40$ s unless special specification. Moreover, the performance comparisons of our proposed approach with the shared commercial detection software Comskip [1][5] are given in Table IV. We can explicitly observe that the proposed method convincingly outperforms Comskip since some more intrinsic characteristics associated with commercials have been exploited. In addition, the necessity of the postprocessing is also validated as shown in Table IV with the evident performance gains.

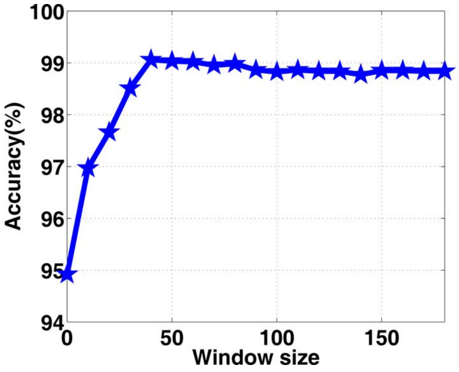


Fig. 14. *Accuracy* versus window size $W_p$ on the validation set.

TABLE IV
PERFORMANCE COMPARISONS WITH COMSKIP [5]

| Methods | *F1* | *Precision* | *Recall* | *Accuracy* |
|---|---|---|---|---|
| Comskip[1] [5] | 43.23% | **98.37%** | 27.70% | 82.95% |
| The proposed method without post processing | 92.06% | 88.88% | 95.47% | 94.92% |
| The proposed method with post processing | **98.49%** | 97.79% | **99.19%** | **99.06%** |

### C. Empirical Results of CBS

To evaluate the performance of CBS, a series of experiments are conducted on 2.48-h commercial blocks. Specifically, we select 1.05-h commercial blocks containing 270 individual commercials as training samples and the remainder (composed of 364 individual ones) as a testing set.

*1) Performance Evaluation on FMPI Frame Detection:* We manually collect 5915 frames comprising 1980 FMPI frames as well as 3935 non-FMPI frames from commercial shots to verify the performance on the FMPI frame detection. It is worth noting that the random bi-partition strategy is adopted to build the training and testing set. Moreover, LIBSVM [2] is employed to learn a classifier with the RBF kernel, and the optimal parameters $C$ and $\gamma$ are obtained by cross-validation. Fig. 15 shows the impact of different descriptors on the discrimination ability of FMPI frames. It is clarified that the performance of the "Text" descriptor is unsatisfactory compared with other descriptors, mainly stemming from some inevitable noisy text areas brought by the automatic CTD method. Moreover, we can observe that the combination of descriptors shows the significant performance improvements that benefit from the collaborative exploitation of the intrinsic visual and textual semantic cues.

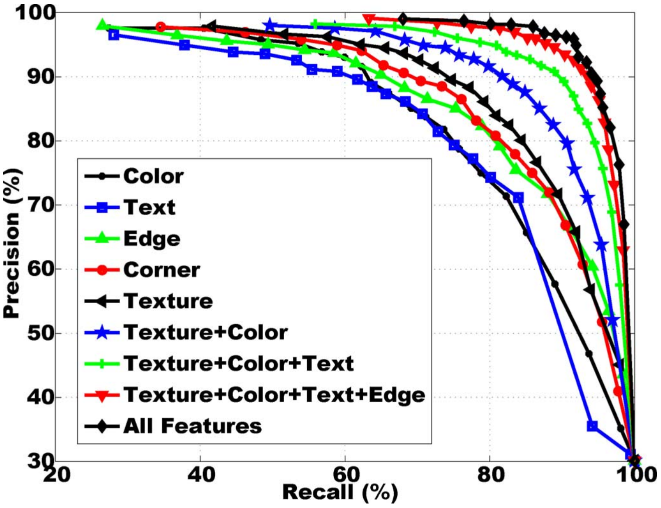


Fig. 15. Performance of combining different descriptors in FMPI detection.

TABLE V
PERFORMANCE COMPARISONS OF THE PROPOSED FMPI DETECTION METHOD WITH [18] AND [19]

| Approaches | *F1* | *Precision* | *Recall* | *Accuracy* |
|---|---|---|---|---|
| Approach in [18]-[19] | 91.06% | 94.34% | 87.99% | 94.79% |
| Our approach | **93.63%** | **95.77%** | **91.58%** | **96.25%** |

We also compare the proposed FMPI construction method to the one [3] in [18] and [19] based on the optimal *Accuracy* performance. The results are shown in Table V, from which we can find an evident gain of our approach with 1.5% improvement on *Accuracy*, owing to the use of the intrinsic textual characteristics and more complementary visual cues, which were not considered in [18] and [19].

*2) Results of ASVP Classification:* In order to construct a robust ASVP classifier for CBS, 2046 ASV and 5963 non-ASV samples are manually collected from the huge collection of commercial shots. Half of them are randomly selected as the training set and the rest as the testing set. To evaluate the impact

[1] In this case, all kinds of available detection methods, including black frame, logo, scene change, fuzzy logic, closed captions, aspect ratio, and silence, were utilized in the experiments with default parameter setting.

[2] The source code is available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[3] The source code is available: http://nlpr-web.ia.ac.cn/english/iva/homepage/jqwang/Demos. htm

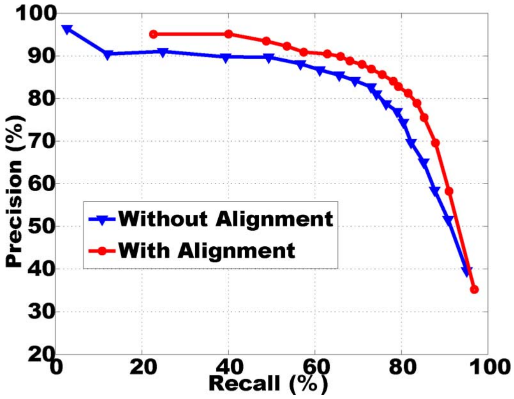


Fig. 16. ASVP detection with and without audio-visual alignment.

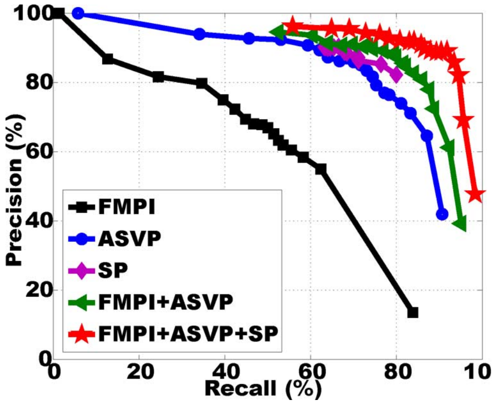


Fig. 17. *Precision* versus *Recall* of different characteristics and their combinations for CBS.

TABLE VI
PERFORMANCE COMPARISONS OF THE PROPOSED ASVP DETECTION WITHOUT AND WITH ALIGNMENT

| Alignment | *F1* | *Precision* | *Recall* | *Accuracy* |
|---|---|---|---|---|
| Without | 77.40% | 82.69% | 72.75% | 89.14% |
| With | **79.47%** | **86.72%** | **73.34%** | **90.31%** |

TABLE VII
PERFORMANCE COMPARISONS OF INDIVIDUAL CHARACTERISTICS AND THEIR COMBINATION FOR CBS

| Characteristics | *F1* | *Precision* | *Recall* | *Accuracy* |
|---|---|---|---|---|
| FMPI | 57.49% | 63.16% | 52.75% | 92.65% |
| ASVP | 77.62% | 82.41% | 73.35% | 96.01% |
| SP | 79.82% | 86.03% | 74.45% | 96.45% |
| Combination | **88.11%** | **89.74%** | **86.54%** | **97.80%** |

of the audio-visual alignment process on the performance of ASVP classifier, we test our proposed means with or without the alignment process on the above data collection and obtain two *Precision-Recall* curves, shown in Fig. 16. We can explicitly observe that the proposed audio-visual alignment process has effectively pruned a large number of incorrect ASV positions and significantly improved the classification results. Moreover, *Accuracy*-based optimal performance comparisons of the proposed ASVP detection without and with alignment are given in Table VI. It is clear that the 1.2% improvement on *Accuracy* with the alignment demonstrates the effectiveness and the necessity of audio-visual alignment.

*3) Performance of CBS:* Based on the above evaluations on the effectiveness of the proposed construction methods for FMPI and ASVP, we can obtain a series of reliable characteristics for CBS. In order to further analyze the contributions of different characteristics and their combinations to our proposed CBS scheme, some experiments are conducted on the 2.48-h-long commercial block dataset. As illustrated in Fig. 17, we can intuitively find that the FMPI does not achieve desirable results. A possible explanation is that FMPI frames occur not only around the end of commercials but also in other portions to timely highlight product information. If we rely only on the characteristic of FMPI to segment commercial blocks, some false alarms will be inevitable. Noticeably, SP achieves the best performance. Reduction of false alarms in CBS is partially due to the robustness of the detection approach. Moreover, the combinations of different characteristics show the significant gains benefiting from the collaborative exploitation of the intrinsic visual-audio-textual characteristics.

Table VII highlights the *Accuracy*-based optimal performance of each characteristic and their combination. It is clarified by Table VII that our CBS method has achieved promising performance with *Accuracy* of 97.80%, which evidently outperforms all individual characteristics due to the exploitation of the collaborative relations among them.

## VI. CONCLUSION

In this paper, we have presented a unified solution for commercial block detection and segmentation based on the collaborative exploitation of the intrinsic visual-audio-textual characteristics. In addition to utilizing the widely applied audio-visual descriptors, an abundance of textual characteristics associated with TV commercials are fully exploited to describe the diverse intrinsic characteristics for CBD and CBS, respectively. Moreover, we introduce an interactive ensemble learning method, i.e., Tri-AdaBoost, to form a consolidated semantic fusion across visual, audio, and textual characteristics. The promising experimental results on large video data collections have shown the effectiveness of the proposed scheme.

## ACKNOWLEDGMENT

The authors would like to thank Dr. C. Hepp from Arizona State University for manuscript proofreading.

## REFERENCES

[1] R. Lienhart, C. Kuhmiinch, and W. Effelsberg, "On the detection and recognition of television commercials," in *Proc. ICMCS'97*, 1997, pp. 509–516.

[2] D. A. Sadlier, S. Marlow, N. O'Connor, and N. Murphy, "Automatic TV advertisement detection from MPEG bitstream," *Pattern Recognit.*, vol. 35, no. 12, pp. 2719–2726, Dec. 2002.

[3] A. Albiol, M. J. Ch, F. A. Albiol, and L. Torres, "Detection of TV commercials," in *Proc. ICASSP'04*, May 2004, vol. 3, pp. 541–544.

[4] Y. P. Huang, L. W. Hsu, and F. E. Sandnes, "An intelligent subtitle detection model for locating television commercials," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 2, pp. 485–492, Apr. 2007.

[5] Comskip. [Online]. Available: http://www.kaashoek.com/comskip/.

[6] M. Covell, S. Baluja, and M. Fink, "Advertisement detection and replacement using acoustic and visual repetition," in *Proc. IWMSP'06*, Oct. 2006, pp. 461–466.

[7] J. M. Gauch and A. Shivadas, "Finding and identifying unknown commercials using repeated video sequence detection," *J. CVIU*, vol. 103, no. 1, pp. 80–88, Jul. 2006.

[8] H. T. Shen, X. F. Zhou, Z. Huang, J. Shao, and X. M. Zhou, "UQLIPS: a real-time near-duplicate video clip detection system," in *Proc. ICVLDB'07*, Sep. 2007, pp. 1374–1377.

[9] N. Liu, Y. Zhao, and Z. F. Zhu, "Commercial recognition in TV streams using coarse-to-fine matching strategy," in *Proc. PCM'10, Part 1*, Sep. 2010, pp. 296–307.

[10] L. Agnihotri, N. Dimitrova, T. McGee, S. Jeannin, D. Schaffer, and J. Nesvadba, "Evolvable visual commercial detector," in *Proc. CVPR'03*, Jul. 2003, vol. 2, pp. 79–84.

[11] P. Duygulu, M. Y. Chen, and A. Hauptmann, "Comparison and combination of two novel commercial detection methods," in *Proc. ICME'04*, Jun. 2004, vol. 2, pp. 1267–1270.

[12] T. Y. Liu, T. Qin, and H. J. Zhang, "Time-constraint boost for TV commercials detection," in *Proc ICIP'04*, Oct. 2004, vol. 3, pp. 1617–1620.

[13] X. S. Hua, L. Lu, and H. J. Zhang, "Robust learning-based TV commercial detection," in *Proc. ICME'05*, Jul. 2005, pp. 149–152.

[14] L. Zhang, Z. F. Zhu, and Y. Zhao, "Robust commercial detection system," in *Proc. ICME'07*, Jul. 2007, pp. 587–590.

[15] M. Mizutani, S. Ebadollahi, and S. F. Chang, "Commercial detection in heterogeneous video streams using fused multi-modal and temporal features," in *Proc. ICASSP'05*, Mar. 2005, vol. 2, pp. 157–160.

[16] M. Li, Y. Cai, M. Wang, and Y. X. Li, "TV commercial detection based on shot change and text extraction," in *Proc. CISP'09*, Oct. 2009, pp. 1–5.

[17] N. Liu, Y. Zhao, Z. F. Zhu, and H. Q. Lu, "Multi-modal characteristics analysis and fusion for TV commercial detection," in *Proc. ICME'10*, Jul. 2010, pp. 831–836.

[18] L. Y. Duan, J. Q. Wang, Y. T. Zheng, J. S. Jin, H. Q. Lu, and C. S. Xu, "Segmentation, categorization, and identification of commercial clips from TV streams using multimodal analysis," in *Proc. MM'06*, Oct. 2006, pp. 201–210.

[19] L. Y. Duan, Y. T. Zheng, J. Q. Wang, H. Q. Lu, and J. S. Jin, "Digesting commercial clips from TV streams," *IEEE Multimedia*, vol. 15, no. 1, pp. 28–41, Jan.-Mar. 2008.

[20] J. Q. Wang, L. Y. Duan, Q. S. Liu, H. Q. Lu, and J. S. Jin, "A multimodal scheme for program segmentation and representation in broadcast video streams," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 393–408, Apr. 2008.

[21] Y. T. Zheng, L. Y. Duan, Q. Tian, and J. S. Jin, "TV commercial classification by using multi-modal textual information," in *Proc. ICME'06*, Jul. 2006, pp. 497–500.

[22] M. R. Lyu, J. Q. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 243–255, Feb. 2005.

[23] K. I. Kim, K. C. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1631–1639, Dec. 2003.

[24] H. P. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 147–156, Jan. 2000.

[25] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. IWCLT'98*, Jul. 1998, pp. 92–100.

[26] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, Jan. 1998.

[27] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[28] B. S. Ong, "Towards automatic music structural analysis: Identifying characteristic within song excerpts in popular music," Master's thesis, Univ. Pompeu Fabra, Barcelona, Spain, 2005.

[29] H. P. Li and D. Doermann, "A video text detection system based on automated training," in *Proc. ICPR'00*, Sep. 2000, pp. 223–226.

[30] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[31] Z. H. Zhou and M. Li, "Tri-training: exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

**Nan Liu** received the B.E. degree in biomedical engineering from Beijing Jiaotong University, Beijing, China, in 2005. Currently, he is pursuing the Ph.D. degree in the Institute of Information Science, Beijing Jiaotong University.

His research interest is in pattern recognition, multimedia retrieval, and multimedia content analysis.

**Yao Zhao** (M'05) received the B.E. degree from Fuzhou University, Fuzhou, China, in 1989 and M.E. degree from Southeast University, Nanjing, China, in 1992, both in radio engineering department. He received the Ph.D. degree in the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 1996.

Currently, he is a Professor and Director of the Institute of Information Science, Beijing Jiaotong University. His research interest includes image/video coding, digital watermarking, and content-based multimedia retrieval.

Dr. Zhao was the recipient of National Science Foundation of China for Distinguished Young Scholars in 2010.

**Zhenfeng Zhu** received the Ph.D. degree in pattern recognition and intelligence system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently an Associate Professor of the Institute of Information Science, Beijing Jiaotong University. He has been a visiting scholar at the Department of Computer Science and Engineering, Arizona State University, during 2010. His research interests include image and video understanding, computer vision, and machine learning.

**Hanqing Lu** (SM'06) received the B.S. and M.S. degrees from the Department of Computer Science and Department of Electric Engineering in Harbin Institute of Technology, Harbin, China, in 1982 and 1985, respectively, and the Ph.D. degree from the Department of Electronic and Information Science in Huazhong University of Science and Technology, Hubei, China.

He is a Professor of the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Currently, his research interests include web multimedia search, personalized video customization, and activity analysis. He has over 200 papers published in academic journals or conferences.