Segmentation, Categorization, and Identification of Commercial Clips from TV Streams Using Multimodal Analysis

Ling-Yu DUAN[1], Jinqiao WANG[2], Yantao ZHENG[1], Jesse S. JIN[3], Hanqing LU[2], Changsheng XU[1]
[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{lingyu, stuytz, xucs}@i2r.a-star.edu.sg
[2]Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
{jqwang, luhq}@nlpr.ia.ac.cn
[3]The School of Design, Communication, and Information Technology, University of Newcastle, Australia.
{Jesse.Jin}@newcastle.edu.au

ABSTRACT

TV advertising is ubiquitous, perseverant, and economically vital. Millions of people's living and working habits are affected by TV commercials. In this paper, we present a multimodal ("visual + audio + text") commercial video digest scheme to segment individual commercials and carry out semantic content analysis within a detected commercial segment from TV streams.

Two challenging issues are addressed. Firstly, we propose a multimodal approach to robustly detect the boundaries of individual commercials. Secondly, we attempt to classify a commercial with respect to advertised products/services. For the first, the boundary detection of individual commercials is reduced to the problem of binary classification of shot boundaries via the mid-level features derived from two concepts: Image Frames Marked with Product Information (FMPI) and Audio Scene Change Indicator (ASCI). Moreover, the accurate individual boundary enables us to perform commercial identification by clip matching via a spatial-temporal signature. For the second, commercial classification is formulated as the task of text categorization by expanding sparse texts from ASR/OCR with external knowledge. Our boundary detection has achieved a good result of F1 = 93.7% on the dataset comprising 499 individual commercials from TRECVID'05 video corpus. Commercial classification has obtained a promising accuracy of 80.9% on 141 distinct ones. Based on these achievements, various applications such as an intelligent digital TV set-top box can be accomplished to enhance the TV viewer's capabilities in monitoring and managing commercials from TV streams.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods, indexing methods.*

General Terms

Algorithms, Design, Experimentation

MM'06, October 23–27, 2006, Santa Barbara, California, USA.

Keywords

TV commercial, multimodal analysis, semantics, segmentation, video classification, mid-level features, text categorization.

1. INTRODUCTION

Advertising is an organized method of communicating information about a product or service which a company or individual wants to promote to the people. It is a paid announcement that is conveyed through words, pictures, music, and action in a medium (e.g., newspaper, magazine, broadcast channels, etc.). Although the costs of creating, producing, and airing a TV commercial are staggering, television is one of the most cost-effective media. Its advantages are impact, credibility, selectivity, and flexibility [1]. In the world of satellite and cable television, TV commercials have become indispensable for most clients. Many cable channels fill in 10 to 12 minutes of a 30-minute serial with commercials.

In just one day, a TV viewer may be exposed to hundreds of commercials. Over a year, it can be tens of thousands. With the advance of digital video recording and playback systems, much previous work [2] – [4] has focused on automatically locating a commercial disposed within a video stream towards "commercial skip" type of applications. When a copy of the program is created for viewing at a later time, many users are not interested in the content of commercials or promotions that are interposed within the television program. Automated commercial detection techniques can replace a user's manual skipping operation. Such work deals with a series of consecutive commercials as a whole block.

Today, TV commercials are produced for 30 or 60 seconds, spending millions of US dollars. One 30-second commercial in prime time can easily cost up to 120,000 US dollars [1]. Millions of people are reached by commercials which modify their living and work habits, if not immediately, at least later. Hence, we make an attempt in this paper to carry out structural and semantic content analysis within a detected commercial segment itself, thereby breaking traditional viewpoints towards TV commercials.

Referring to the illustrative paradigm in Fig.1, we summarize four points to explain the motivations and potential applications of TV commercial video segmentation, categorization and identification. **Firstly**, as the advertisers spend much money, it is necessary to verify their commercials are broadcasted as contracted. A system for TV commercial monitoring is desired. A preliminary stage is

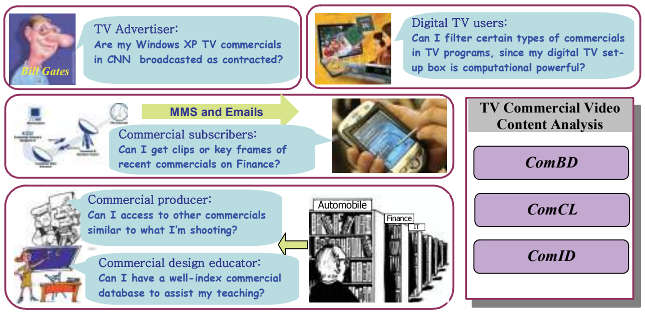

TV Advertiser:
Are my Windows XP TV commercials in CNN broadcasted as contracted?
Bill Gates
Digital TV users:
Can I filter certain types of commercials in TV programs, since my digital TV set-up box is computational powerful?
MMS and Emails
Commercial subscribers:
Can I get clips or key frames of recent commercials on Finance?
TV Commercial Video Content Analysis
ComBD
ComCL
ComID
Commercial producer:
Can I access to other commercials similar to what I'm shooting?
Commercial design educator:
Can I have a well-index commercial database to assist my teaching?
Automobile
Finance
IT

Fig. 1 An application paradigm of TV commercial segmentation (ComBD), categorization (ComCL), and identification (ComID)

to determine the boundaries of individual commercials. *Accurate boundaries are useful for effective clip-level video matching and subsequent statistics of real duration in TV broadcast*. **Secondly**, research shows that most people do not mind TV advertising in general, although they dislike certain commercials; they do not like to be yelled at or treated rudely; they want to be respected[1]. With the advance of digital TV set-top boxes in terms of powerful processors, large hard disks and internet access, it is desirable to furnish consumers with *a TV commercial management system*, which detects commercial segments, determine the boundaries of individual commercials, identify & track new commercials, and summarize the commercials within a period by removing repeated instances. Given a decent interface, this system may change a TV viewer's passive position. *A user can apply positive actions (e.g., search, browse, etc.) to the commercial video archive*. As advertising in the mass media is basically incidental to consumers' use of the media, this system indirectly *improve the reachability of TV commercials*. **Thirdly**, all advertisements deal with one of three concepts: ideas, products, and services [1]. TV commercial classification with respect to the advertised products or services (e.g., automobile, finance, etc.) helps to fulfill the *commercial filtering towards personalized consumer services*. For example, an MMS message (containing key frames or adapted video) on the commercials of interest to a registered user can be sent to her/his mobile device or email box. **Fourthly**, the technology of TV commercials has changed much; they are almost always edited on a computer; the appearance all starts with MTV and MTV commercials are more visual, more quickly paced, use more camera movement, and often combine multiple looks, such as black and white with color, or stills with quick cuts [1]. Accordingly, *a TV commercial archive system* including browse, classification, and search may *inspire the creation of a good commercial*. Marketing companies may even utilize it to *observe competitors' behaviors*.

Aiming at above applications, we have to address two challenging tasks: individual commercials' boundary detection (ComBD) and commercial classification in terms of advertised products/services (ComCL). The first is the problem of video parsing; the latter is that of semantic video indexing. In TV streams, a commercial block consists of a series of individual commercials (spots). Each spot may be dealt with as a semantic scene. The process of detecting such scene transitions within a block is referred to as commercial video parsing. In a classified advertisement (often found in most newspapers), one can easily find information useful to determine if the advertised item is to be bought. Accordingly, semantic commercial video indexing is meant to accomplish such classified TV advertisement through video content analysis techniques. Since an advertising campaign concerns many topics such as babies, cars, entertainment, fashion, food, money, sports, and so on, we will choose some representative categories of products or services to explore the solution via multimodal analysis.

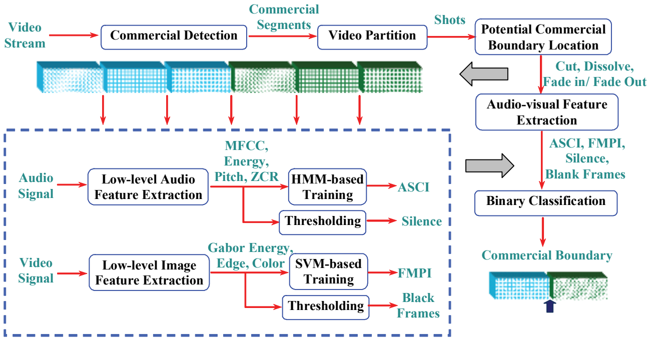


Fig.2 Framework for detecting commercials' boundaries

Once individuals' boundaries are determined, various video clip matching methods can be used to identify commercials (ComID). One issue lies in a compact and robust signature for representing commercial video content. The other issue is to accelerate the clip search in a large database. Compared to ComBD and ComCL, ComID can be easily addressed by existing or modified methods.

The rest of this paper is organized as follows. In Section 2, we present our approach to ComBD, while in Section 3 we present our approach to ComCL. Challenges and our methods' unique features are discussed as well. In Section 4, we brief our previous work on ComID. ComBD is a significant stage for ComCL and ComID. Experimental results and discussions are presented in Section 5. Finally, in Section 6, we briefly present conclusions.

2. INDIVIDUAL COMMERCIALS' BOUNDARY DETECTION (ComBD)

We first discuss the challenges of ComBD from the viewpoint of scene transition detection. An overall framework is then provided. The rest focuses on the extraction of intermediate or mid-level features and their fusion. Production knowledge is emphasized.

2.1 Challenges

The term *scene transition detection* (STD) is used to differentiate commonly known *scene change detection* that aims to detect shot boundaries by visual primitives. Generally, a scene or a story unit is composed of a series of "interrelated shots that are unified by location or dramatic incident" [9]. STD aims to detect scenes on the basis of computable audiovisual characteristics and production rules. There has been lots of prior work on STD concentrating on sitcoms, movies [5] – [9], or broadcast news video [10] [11].

Instead of a single shot, a scene is often treated as an elementary and meaningful unit for effective browse, navigation, and search in video programs where rough scene boundaries suffice for organizing video content. Rather than exactly locating scene boundaries, most work dealt with STD via the aggregation of consecutive shots. Clearly, a scene lies in the marriage of video structure and semantics. By investigating the temporal consistency of audiovisual contents, only an approximation for the actual scene should be expected. That is, exact scene boundaries cannot be secured. In particular, commercial videos are featured by dramatic changes in lighting, chromatic composition, and tempo (determined by shot length, motion, zoom, sound, etc.) amongst shots, and by creative stories. It makes existing STD methods less effective for ComBD, as video shots lack uniform agglomeration within a commercial.

2.2 Solution and Framework

Our approach reduces the problem of commercial STD to that of a binary classification of True Scene Changes versus False Ones at candidate positions consisting of video shot change points. It is reasonably assumed that a TV commercial scene transition always comes with a shot change (i.e., cuts, fade-in/-out, and dissolves). Multimodal features are extracted within a symmetric window at each candidate point. Different or multi-scale window sizes may be optionally applied to different kinds of features. A supervised learning is subsequently applied to fuse multimodal features. Particularly, two concepts of Audio Scene Change Indicator (ASCI) and Image Frames Marked with Product Information (FMPI) are proposed to characterize computational video contents (structural or semantic) of interest to signify the boundaries of an individual commercial. As argued above, it is infeasible to decipher a commercial video's temporal arrangement via a predefined set of shot classes. The role of mid-level features is to condense high-dimensional low-level features by using adequate classifiers to generate as many useful concepts as possible that are supported by commercial video production rules or knowledge. The framework is illustrated in Fig. 2.

General commercial detection is a preliminary stage of significance. Many approaches were proposed [2] – [4], [12] – [14]. An accuracy of 92% on a heterogeneous dataset was reported in [2]. Basically our implementation resorts to the detection and tracking of TV logos as TV logos are often removed during commercials. Satisfactory results of F1 = 97.76% ~ 99.80% were achieved on opaque, semi-transparent, and animated TV logos from 8 TV channels such as NBC, CNN, MSNBC, etc (More details in [15]).

Accurate cuts and fade-in/-out is significant in our scheme. As the experimental videos are all in MPEG-1 format, we employ the compressed domain approach in [16] to determine cuts. In terms of parameter tuning, we prefer a higher recall of cuts. Fade-in/-out is determined by detecting monochrome frames and detecting gradual transitions simply via the twin comparison method (TCM) [17], as the fade-in/-out between two successive spots is in a short duration (often less than 8 frames) and TCM can work well for short gradual transition [43]. In experiments detected cuts and fade-in/-out have covered about 98% true individuals' boundaries.

2.3 Intermediate (or Mid-level) Features

2.3.1 Image Frame Marked with Product Information (FMPI)

2.3.1.1 Using an FMPI frame to locate the most probable boundary candidates

FMPI is used to describe those images containing visual information explicitly illustrating an advertised product or service. The visual information is expressed in the combination of three ways: text, computer graphics, and frames from a live footage of real things and people. Fig.3 lists the examples of FMPI frames. The textual section may consist of brand name, store name, address, telephone number, and cost, etc. Alongside the textual section, a drawing or a photo of a product might be placed with computer graphics techniques. As graphics create a more or less abstract, symbolic, or "unreal" universe in which incredible things happen, live footage of real things or people is usually combined with graphics to solve the problem of impersonality.

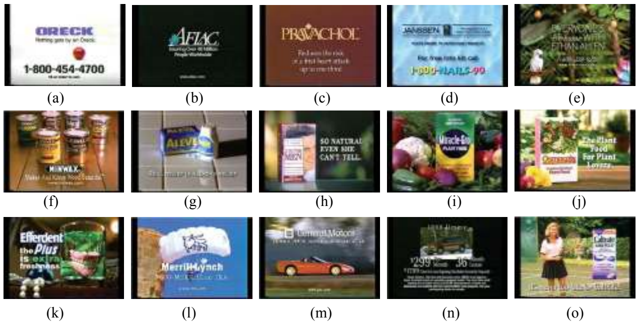


Fig.3 Image frames marked with product information (FMPI)

Let us investigate those examples. Fig. 3 (a)–(e) are the simplest yet most prevalent ones. For Fig. 3 (f)–(j), the product is projected into the foreground, usually in crisp, clear magnification. For Fig. 3 (k)–(o), the FMPI frames are yielded by the superimposed text bars, graphics, and live footage. From the image recognition point of view, Fig. 3 (a)–(e) produce a fairly uniform pattern; for Fig. 3 (f)–(j), the pattern variability mainly derives from the layout and the appearance of a product; Fig. 3 (k)–(o) present more diverse patterns due to unexpected real things.

The spatial relationship between the presence of FMPI frames and individual commercials' boundaries is revealed by the production rules. For the convenience of description, we define the video shot containing at least one FMPI frame as an FMPI shot. Firstly, in many commercial videos, one or two FMPI shots are utilized to highlight the offer at the end of a commercial. It is sometimes hard to see what precisely is on offer in a commercial. An FMPI shot is hence a useful "prop". Secondly, an FMPI shot might be irregularly interposed in the course of some commercials, as our memories are served by of course endless repetition. Occasionally, an FMPI shot may appear at the beginning of a commercial.

Therefore, an FMPI shot can be considered as an indicator, which is able to determine a much smaller set of commercial boundary candidates from large amounts of video shot changes.

2.3.1.2 Constructing an FMPI recognizer

Referring to Fig.3, an FMPI frame can be dealt with as a kind of document image involving graphics (e.g., corporate symbols, logos), images (e.g., product, setting, and props), and text (e.g., brand name, headline or captions, and contact information).

We rely on the combination of texture, edge, and color features to represent an FMPI frame. As the layout is a significant factor in distinguishing an FMPI frame, it is beneficial to incorporate the spatial information. One common approach is to divide an image into subregions and impose positional constraints on the image comparison (called "image partitioning"). Dominant colors are used to construct an approximation of color distributions. They can be easily identified from color histograms. Texture analysis is widely employed in document image processing to discriminate the primary components of page layout, text, line-drawings, background, etc. [18]. Since Gabor filters [19] exhibit optimal location properties in the spatial domain as well as in the frequency domain, they are used to capture rich texture in FMPI frames. Edge is a useful complement of texture when an FMPI frame produces stand-alone edges as a contour of an object, as texture relies on a collection of similar edges. Combined features are shown to yield better results than that using single feature. See Section 5.1.2.

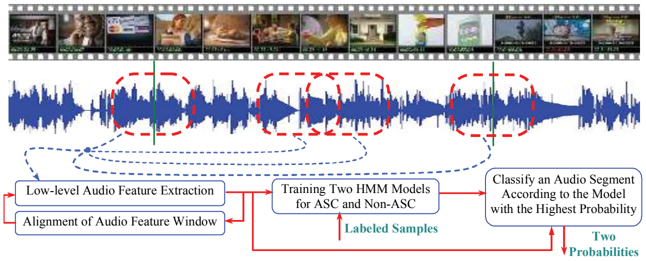

Low-level Audio Feature Extraction
Alignment of Audio Feature Window
Training Two HMM Models for ASC and Non-ASC
Classify an Audio Segment According to the Model with the Highest Probability
Labeled Samples
Two Probabilities

Fig.4 Training an Audio Scene Change Indicator (ASCI)

Our feature extraction procedure is described as below. Let Γ be an $n \times m$ image. *LUV* color space is used. The colors in Γ are uniformly quantized into $3 \cdot l$ bins, each channel being assigned l bins. To extract local features, Γ is partitioned into $r \cdot c$ sub-images equally. Within each sub-image, the first p maximum bin values are selected as dominant color features. Note that the bin values are meant to represent the spatial coherency of color, irrespective of concrete color values. Based on Canny edges [20], we sum up edge pixels within each sub-image, thereby yielding edge density features with $r \cdot c$ dimensions. A set of 2-dimensional Gabor filters are employed for texture. Within each sub-image, the mean μ_{sk} of the magnitudes of transform coefficients is used. For S scales and K orientations, texture features of $r \cdot c \cdot S \cdot K$ dimensions are finally constructed using μ_{sk}. In terms of global features, color and edge are taken into account. Similarly the first q maximum bin values are selected from each channel. Edges are broadly grouped into h categories of orientation by using the angle quantizer as: $A_i = \left[\left\lfloor \frac{180}{h} \right\rfloor \cdot i, \left\lfloor \frac{180}{h} \right\rfloor \cdot (i+1)\right), i = 0 \cdots h-1$. By combining local features and global ones, we obtain the feature of $(3 \cdot p \cdot r \cdot c + r \cdot c + S \cdot K \cdot r \cdot c + 3 \cdot q + h)$ dimensions. Our experiments use parameters as: $r = c = 4$, $p = 1$, $q = 3$, $S = 1$, $K = 4$, and $h = 4$. Note our Gabor filters use one center frequency (one scale) and four equidistant orientations. Finally, we construct the feature of 141 dimensions (128 local features and 13 global ones).

Subsequently, SVMs is utilized to train the FMPI recognizer. Advantages of SVMs consist of: a) working well for data with a large number of features, and b) containing fewer parameters. Our implementation resorts to the *C*-Support Vector Classification (C-SVC) [21]. Aiming to determine an FMPI shot, FMPI recognition may be applied to key frames only. Motion is utilized to identify key frames. We simply use the average intensity of motion vectors from B- and P- frames to measure motion within a shot and to select key frames at the local minima of motion [22]. Directing recognition to key frames has two merits: 1) reducing computation cost; 2) avoiding distracting frames due to animation effects.

2.3.2 Audio Scene Change Indicator (ASCI)

2.3.2.1 Using an ASCI indicator to characterize audio changes occurring at commercial boundaries

The most common TV commercial is a combination of continuous music, sound effects, voice-over narration, and storytelling video. It is easy to imagine different TV commercials exhibit dissimilar audio characteristics. A proper modeling of audio scene changes (ASC) can facilitate the identification of commercial boundaries. An audio scene is often modeled as a collection of sound sources and the scene is further assumed to be dominated by a few of these sources [5]. ASC is said to occur when the majority of the dominant sources in sound change. It is more or less complicated and sensitive to determine the ASC transition pattern in terms of acoustic classes [23] because of model-based methods' weakness: large amounts of samples required and the subjectivity of classes labeling. An alternative is to examine the distance metric between two windows based on audio features. Metric-based methods are straightforward. A quantitative indicator is produced. Yet human knowledge is not incorporated by labeling training data or others.

Given an audio segment (say 4 seconds) at a candidate boundary, ASCI provides a probabilistic representation of ASC. As shown in Fig. 4, HMM is utilized to train two models for "explaining" two dynamic patterns of ASC and Non-ASC. An unknown segment is classified by the model that has the highest posterior probability. ASCI has two features. Firstly, the labeling of ASC/Non-ASC is simpler and can capture the sense of hearing when one is viewing commercial videos. Secondly, as shown in Fig.2, two probability values yielded by the ASCI, as intermediate features, can be fused with others; that is, their subjectivity would not seriously affect the final target. Moreover, a metric-based method is introduced to align the audio feature window. Experiments have shown that the integration of model-based and metric-based yields better results.

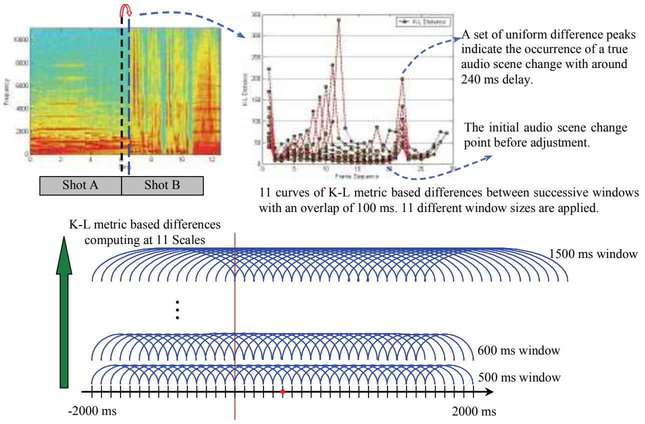
Like FMPI, ASCI is an indicator but cannot secure true boundaries due to dynamic audio characteristics inherent to commercials. Fusing multimodal features, e.g. "ASCI + FMPI', is our solution.

2.3.2.2 Utilizing HMM to train recognizers

A mixture Gaussian HMM (left-to-right) is utilized to train ASC/Non-ASC recognizers. Diagonal covariance matrix is used to estimate the mixture Gaussian distribution. Suppose we have two HMM models for representing ASC and Non-ASC, two likelihood values of an observation sequence are generated by the forward-backward algorithm. HTK toolkit [24] is utilized.

Currently our ASCI considers 43-dimensional audio features comprising Mel-frequency cepstral coefficients (MFCCs) and its first and second derivates (36 features), mean and variance of short time energy log measure (STE) (2 features), mean and variance of short-time zero-crossing rate (ZCR) (2 features), short-time fundamental frequency (or Pitch) (1 feature), mean of the spectrum flux (SF) (1 feature), and harmonic degree (HD) (1 feature). See [23], [25] for details on these features. Given an audio signal, we segment it into a series of successive 20 ms analysis frames by shifting the sliding window of 20 ms with an interval of 10 ms. Features are computed for each analysis frame. Within each analysis frame we compute STE, ZCR, SF, and Harmonic peaks once every 50 samples at an input sampling rate of 22, 050 samples per sec where the size of sliding window is set to 100 samples. Means and variances of STE and ZCR are calculated for 7 values from 7 overlapping frames while the mean of SF is calculated for 6 values from 7 neighbor frames. HD is the ratio of the number of frames having harmonic peaks to the frame number 7. Pitch and MFCCs are computed directly from each frame.

The reasons for audio features are briefed below. MFCCs furnish a more efficient representation of speech spectra. STE provides a basis for discriminating between voiced speech components and unvoiced speech components, speech and music, audible sounds



A set of uniform difference peaks indicate the occurrence of a true audio scene change with around 240 ms delay.

The initial audio scene change point before adjustment.

11 curves of K-L metric based differences between successive windows with an overlap of 100 ms. 11 different window sizes are applied.

Fig.5 Kullback-Leibler based alignment of audio feature windows

and silence. For ZCR, music produces much lower variances and amplitudes than speech does. ZCR is also useful for distinguishing environmental sounds. Pitch determines the harmonic property. Voiced speech components are harmonic while unvoiced speech components are non-harmonic. Sounds from most musical instruments are harmonic while most environmental sounds are non-harmonic. In general, the SF values of speech are higher than music but less than those of environmental sounds [23] [25].

2.3.2.3 Aligning the audio feature window

Referring to Fig.4, the alignment problem has to be addressed for two reasons. Firstly, at most TV commercial boundaries there is an offset of ±0.25 sec ~ ±1.0 sec between an ASC and its associated video scene change. Secondly, due to video production, a mixed soundtrack does not necessarily synchronize a video track; thus a symmetric window exactly at shot transitions cannot secure the extraction of most effective features well matching the ASC nearby. This is supported by the statistics in news program and commercials. Around 95% offsets lie in the range of ±0.25 sec ~ ±1.0 sec wherein the offsets of ±0.25 sec occupy around 85%.

An alignment procedure seeks to locate the most likely ASC point within the neighborhood of a shot change as illustrated in Fig. 5. Let W_i and W_j be two audio analysis windows, and their difference denoted by $d(W_i,W_j)$. By utilizing Kullback-Leibler (K-L) distance metric [26], the difference can be written as

$$d(W_i,W_j)=\int_x \left[p_i(x)-p_j(x)\right]\ln p_i(x)\big/p_j(x)dx$$

where $p_i(x)$ and $p_j(x)$ denote the probability distribution functions (*pdf*) estimated by the features extracted from W_i and W_j. One-scale is considered firstly. Let W_i, $i=1,2,\ldots,N$ be a series of analysis windows with an overlap of INT ms. We then form the sequence $\{D_i\}_{i=1,2,\ldots,N-1}$ as $D_i=d(W_i,W_{i+1})$. An ASC from W_l to W_{l+1} is declared if D_l is the maximum within a symmetric window of WS ms. Window size is critical to good modeling. The difference curves in Fig. 5 have indicated different change peaks in the cases of different window sizes. Since one does not know a priori what sound one is analyzing, multi-scale computing is used. We first make use of multiple window sizes $\{Win_{scale}\}_{scale=1,\ldots,S}$ to yield a cluster of difference value series which is denoted by $\{\text{Distance}_{\text{scale}}\}_{scale=1,\ldots,S}=\left\{D_{i,scale}\middle|i=1,\ldots,N_{scale}\right\}_{scale=1,\ldots,S}$; each series of $\text{Distance}_{\text{scale}}$ is then normalized to [0,1] through dividing difference values $D_{i,scale}$ by the maximum of each series $Max\left(\text{Distance}_{\text{scale}}\right)$; the most likely ASC point ω is finally determined by locating the highest accumulated values.

The probability $p(\omega_\lambda)$ of the candidate window position ω_λ being an ASC point is calculated as:

$$p(\omega_\lambda)=\frac{1}{S}\sum_{scale=1}^{S}\left(\frac{\text{Distance}_{\text{scale}}(\lambda)}{\underset{1\le\lambda\le M}{Max}\ (\text{Distance}_{\text{scale}}(\lambda))}\right)$$

$$\omega=\arg\underset{\lambda}{Max}\left(p(\omega_\lambda)\right),\ \lambda=1,\ldots,M$$

where M denotes the total number of candidate window positions, ω denotes the window corresponding to an ASC point.

Based on offset statistics, the shift of adjusted change point is confined to the range of [-500ms, 500ms], i.e., $WS=1000$. We extract and arrange audio features within the adjusted 4-second feature windows. 11 Scales are employed, i.e., $S=11$, where the window sizes $Win_{i=1,\ldots,11}=500+100\cdot(i-1)$ ms. At all scales, the overlap interval is set to $INT=100$ ms. A single Gaussian *pdf* is used. 20 ms sliding window with an interval of 10 ms is applied.

2.3.3 Silence & Black Frames

Spots may be separated by a short break of several black frames and/or audio-depression occurrences in some TV channels [4]. In heterogeneous video streams, it is useful to incorporate Silence & Black Frames to form a complete feature set towards robustness.

Silence is detected by examining the audio energy level. The short-time energy function is measured every 10 ms and smoothed using an 8-frame FIR filter. The smoothing implicitly imposes a minimum length constraint on the silence period. A threshold is applied, and the segment that has its energy below the threshold is decided as Silence. A black frame is detected by evaluating the mean and the variance of intensity values for a frame. A threshold method is applied. A series of consecutive black frames (say 8) indicate the presence of Black Frames.

2.4 Feature Fusion

The features of FMPI, ASCI, Silence and Black Frames, extracted from a temporal window at a candidate boundary, are fused with a binary classifier as indicated in Fig. 2. Our implementation relies on a SVMs classifier to accomplish this fusion. To evaluate the effectiveness of FMPI and ASCI, we conduct the fusion on five experimental feature combinations of “FMPI”, “Black Frames”, “FMPI + ASCI”, “ASCI + Silence + Black Frames”, and “FMPI + ASCI + Silence + Black Frames”.

ASCI yields two probability values $p(\text{ASC})$ and $p(\text{Non-ASC})$. Silence and Black Frames also yield two values $p(\text{Silence})$ and $p(\text{Black Frames})$ to indicate their presence within a temporal window of 4 sec (left- and right- 2 sec). For FMPI, $2\cdot n$ neighbor video shots at a candidate boundary (Left n shots, Right n shots) produce $2\cdot n$ values $\left\{p_i(\text{FMPI})\right\}_{i=1,\ldots,2n}$ to indicate the presence of FMPI shots. The complete feature is $2\cdot n+4$ dimensional. $n=2$.

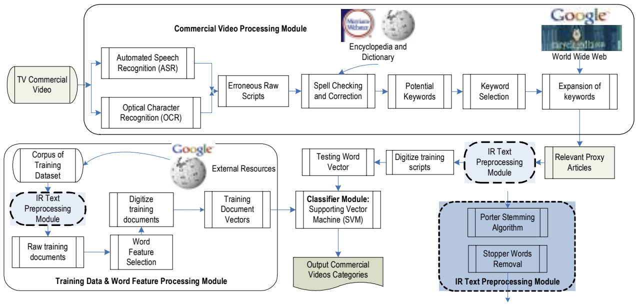

Commercial Video Processing Module
Encyclopedia and Dictionary
World Wide Web
TV Commercial Video
Automated Speech Recognition (ASR)
Optical Character Recognition (OCR)
Erroneous Raw Scripts
Spell Checking and Correction
Potential Keywords
Keyword Selection
Expansion of keywords
Corpus of Training Dataset
External Resources
IR Text Preprocessing Module
Digitize training documents
Training Document Vectors
Raw training documents
Word Feature Selection
Training Data & Word Feature Processing Module
Testing Word Vector
Digitize training scripts
IR Text Preprocessing Module
Relevant Proxy Articles
Classifier Module: Supporting Vector Machine (SVM)
Output Commercial Videos Categories
Porter Stemming Algorithm
Stopper Words Removal
IR Text Preprocessing Module

Fig.6 Framework of TV commercial video classification

The fusion procedure purely relies on SVMs and does not involve any feature selection, weighting, or rules. This simplicity derives from useful concepts well abstracting production knowledge. We are considering other linear classifiers and linear fusion scheme.

3. COMMERCIAL CLASSIFICATION BY TEXT CATEGORIZATION (ComCL)

We first discuss the challenges of ComCL from the viewpoint of semantic video indexing. An overall framework is provided. Then ComCL is formulated as the problem of text categorization.

3.1 Challenges

From TRECVID experiences, the bottleneck of an ideal retrieval lies in the amount of usable semantics obtained [26]. Compared to news or sports, TV commercial videos are essentially creative in terms of copywriting and production techniques. Intermediate visual features (e.g., semantic visual template [27], semantic shot categories [28]) or specialized concept detectors [29] are unable to model intrinsic semantics of commercials using audiovisual features. More recently, lots of research attempts to design a moderate set of semantic concepts towards knowledge network [29], [30]. However, [26], [30] have revealed a significant challenge for extracting high-level features due to the semantic gap.

3.2 Solution and Framework

Alternatively, textual resources are becoming a useful channel for event detection [32], [33] and high-level retrieval [26], [31]. The textual sources can be ASR/OCR script [26], closed caption [33], and web-casting text [32]. The use of textual information has two obvious advantages: clear linkages with semantics, and available text-based external knowledge databases (e.g. WordNet, dictionaries, cyclopedia, and topic-wise document corpora [34]).

Hence, we resort to textual resources for addressing commercial classification with respect to products or services. To the best of our knowledge, no previous work tried to understand what a TV commercial is offering. One related work [42] studied commercial production classification (practical, playful, utopic, and critical). By using text, our approach transforms the problem of semantic video classification to that of automated text categorization [35]. See Fig. 6. It is assumed that ASR/OCR can deliver useful textual hints about advertised products/services. Firstly, we parse the deficient scripts of ASR/OCR to extract keywords, by which search is carried out to retrieve semantically informative articles from internet. The commercial category information is enriched by the document representation of retrieved articles. Secondly, we utilize topic-wise documents from public corpora like Reuters-21578 [34] or from other external sources like internet. Text categorizers are finally trained to determine the commercial category.

(a) on the seasonal **allergies** please S S S T U S S S S S SS SS centered on one **side effects** are generally mild I mean you for us S S S S

(b) Singulair relieves broad range of seasonal allergy symptoms. Singulair will not replace fast-acting inhalers for such symptoms. Continue to take your other … … Ask your doctor about once a day Singulair. Asthma control can help you breathe easier.

(c) Is It a Medication Allergy or a Side Effect? Medications can sometimes cause unpleasant, unwanted effects beyond their intended ones. Although people … … problems fall into the category of side effects or idiosyncratic reactions (an unusual response that is peculiar to that person) to medication.

Fig.7 Singular TV commercial (a) ASR transcript, (b) manually recorded speech transcript, and (c) article searched from Web

[Our OCR utilizes FineReader 8.0 (http://www.abbyy.com/), a commercial OCR system. As the simulation is conducted on TRECVID'05 video corpus, we use the available ASR scripts by MS Speech Recognition Engine.]

3.3 Illustrative Examples

Ideally, ASR and OCR should provide useful texts aiming to classify TV commercials. Unfortunately, the ASR performance would be corrupted by the "noise" background music.

Fig. 7 (a) shows an example of ASR transcript of TV commercial video of *Signulair*, which is a medicine curing asthma and allergy. Fig. 7 (b) presents a part of the manually recorded speech transcript of this commercial. The transcript comparison between Fig. 7(a) and Fig. 7(b) indicates that background music impedes ASR from delivering a semantically meaningful and coherent message describing the advertised commodity. Fortunately, the output of ASR/OCR often contains words related to the advertised commodity's category, like <Allergy> and <side effect> highlighted in Fig. 7(a) and the circled words in Fig. 8.

Since our ultimate purpose is to classify a TV commercial into its advertised commodity's category, e.g., *Singular* commercial into healthcare, it is preferred but not necessary to exploit the actual speech transcript in text categorization. Alternatively, other relevant articles that fall into the same category can serve as the proxy of a TV commercial in the context of classification. For example, the article in Fig. 7(c) is obtained by Google search with keywords <allergy> and <side effect>. Obviously, this article can be classified into the healthcare category.

3.4 Formulation of the proposed approach

Our approach preprocesses the output transcripts of ASR and OCR in TV commercial $TVCom_i$ with spell checking to generate corrected transcript S_i. It then extracts a list L_i of nouns and noun phrases from S_i with the natural language processor. Keywords $K_i(kw_{i1},\ldots,kw_{i\ell})$ are selected by applying the steps as below:

i) Check S_i against a predefined dictionary of brand names;

ii) If the brand name occurs in S_i, it will be selected as the only keyword kw_i and searched on the online encyclopedia Wikipedia (http://en.wikipedia.org/wiki);

iii) Otherwise, from L_i, the n nouns and nouns phrases with largest font size from OCR and the last m from ASR are heuristically selected as keywords to search via a Web Search Engine.

Google search engine is utilized since its superior performance assures the searched articles' relevancy. Among returned articles, the one with the highest relevancy rating is selected as d_i, which we denote as the proxy article of $TVCom_i$. By exploiting d_i, TV

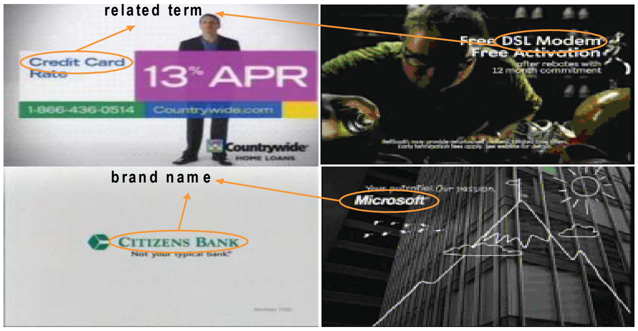


Fig.8 Examples of key frames contain significant semantics

commercial video classification is reduced to the problem of text categorization [35]. That is to approximate a classifier function $\Phi : D \times C \rightarrow \{T, F\}$ to assign a Boolean value to each pair $(d_i, c_i) \in D \times C$, where D is the domain of proxy article d_i and C is the set of predefined commercial category c_i. A value T assigned to (d_i, c_i) indicates the proxy article d_i under c_i, while a value F assigned to (d_i, c_i) means d_i not under c_i.

3.5 Function Modules

3.5.1 IR text preprocessing module

This module functions as a vocabulary term normalization process involving two steps: the Porter Stemming Algorithm (PSA) and the Stop Word Removal Algorithm (SWRA). PSA is to remove the common morphological and inflexional endings from words in English so that different word forms are all mapped to the same token. SWRA is to eliminate words of little or no semantic significance, such as "the", "you", "can", etc. Both testing and training documents go through this module before any other process.

3.5.2 Commercial video module

This module aims to expand the deficient and less-informative transcripts from ASR and OCR with relevant proxy articles.

For each incoming $TVCOM_i$, the module firstly extracts the raw semantic information via ASR and OCR. The accuracy of OCR depends on the resolution of characters in an image. It is empirically observed that the text of large size contains more significant information than small one does. As shown in Fig. 8, it is easy for OCR to recognize the text of large size "Free DSL Modem, Free Activation", which contains more category related information than the small and hard-recognized text "after rebates with 12 months commitment" does. It is the reason why the n nouns and noun phrases with largest font size from OCR are selected to form keywords. Subsequently, the spell checking and correction are applied to the transcripts. The misspelled vocabulary terms are corrected and the terms not found in dictionaries are removed. Both English dictionary and encyclopedia are used as the ground truth for spell checking, as a normal English dictionary may not include non-vocabulary terms like brand names. With the corrected transcript S_i, the proxy article d_i is obtained using the steps stated in Sec. 3.4. The testing vector is generated from d_i.

3.5.3 Training data & word feature module

This module is to generate the training dataset and word feature space for text categorization. Topic-wise document corpora are constructed from available public IR corpora or related articles manually collected from WWW as the training dataset. In this way, the training corpus can possess large amount of training documents and wide coverage of topics. Currently the categorized Reuters-21578 [34] and 20 Newsgroup [36] corpora are combined to construct the training dataset. The topics of these corpora may not exactly match the categories of TV commercials. Our solution is to choose the topics that are related to the commercial category and combine them to jointly construct the training dataset for representing the category. For example, the documents on the topics of "earn", "money", and "trade" in Reuters-21578 are merged together to form the training set for the finance category.

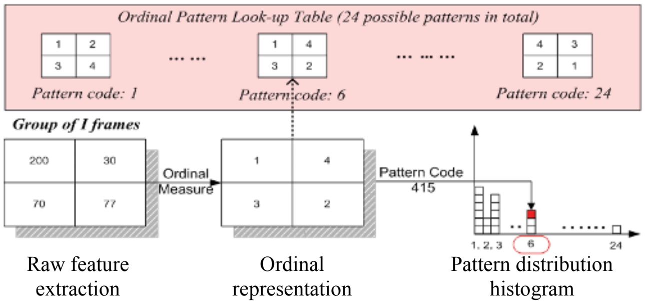


Fig.9 Ordinal feature description

Next document frequency technique is employed to select word features on training dataset. The document frequency $DF(w_i)$ measures the number of documents in which a term w_i occurs. If $DF(w_i)$ exceeds a predetermined threshold, w_i is selected as a feature; otherwise, w_i is removed from the feature space. For each document, the number of occurrences of term w_i is taken as the feature value $tf(w_i)$. Finally, each document vector is normalized to eliminate the influence of different document lengths.

3.5.4 Classifier module

The Classifier Module aims to perform text categorization of proxy articles, and furthermore, determine the categories of respective TV commercials. In [35], various text categorization techniques have been reviewed and SVMs is reported to deliver consistently outstanding performance. Thereby, SVMs is used as the classifier in our implementation. [37] presented the promising characteristics of SVMs to theoretically demonstrate its suitability for text categorization task: 1) capability to handle high dimensional input spaces, say 10,000 dimensions; 2) capability to tackle sparse document vectors. More details can be found in [37].

4. COMMERCIAL IDENTIFING (ComID)

Given the boundaries of individual TV commercials (see Sec. 2), ComID is reduced to video clip matching. To address color distortion and different versions (long or short), we propose a group-of-frames (GoF) based compact signature for characterizing dynamic spatio-temporal patterns inherent to commercial videos.

Our signature combines ordinal feature and color feature. As shown in Fig. 9, each frame is represented by a reduced image of size 2×2. For each Y, Cb or Cr channel, we calculate the average for each of the 4 sub-images. The ordinal measure process [38] is applied. Given a GoF, for each channel $c = Y, Cb, Cr$, the ordinal pattern distribution (OPD) histogram H_c^{opd} is formed. For the

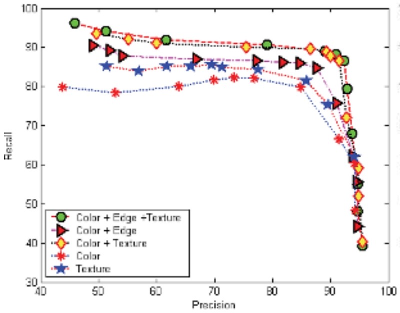


Fig.10 FMPI results yielded by using different features and SVMs parameters

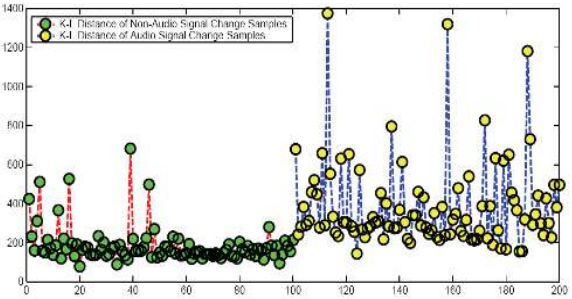


Fig.11 A series of K-L distances calculated from 200 ASC samples and 200 Non-ASC samples.

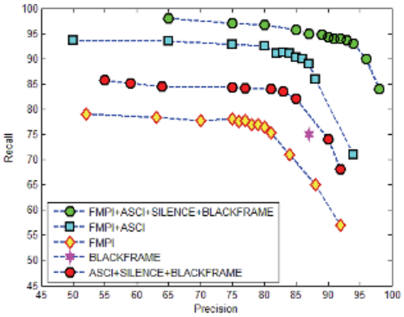


Fig.12 ComBD results yielded by using different features and SVMs parameters

color feature, the cumulative color distribution (*CCD*) histogram H_c^{ccd} is defined. Given the query clip Q and the incoming clip S, the integrated similarity is experimentally defined as the reciprocal of linear combination of the average distance of *OPD* and the minimum distance of *CCD* in three channels:

$$D^{opd}(H_Q,H_S)=\frac{1}{3}\sum_{c=Y,Cb,Cr}D(H_c^{opd}(Q),H_c^{opd}(S))$$

$$D^{ccd}(H_Q,H_S)=\underset{c=Y,Cb,Cr}{Min}\{D(H_c^{ccd}(Q),H_c^{ccd}(S))\}$$

$$Similariy(H_Q,H_S)=\frac{1}{w\times D^{opd}+(1-w)\times D^{ccd}}$$

where Euclidean distance $D(\cdot,\cdot)$ is used, w denotes the weight.

To accelerate video clip search in a large video collection, active sequential search and mrkd-tree based search were compared in [39]. The first is a temporal pruning technique [40] aiming to improve the linear scanning speed while the latter is an index structure for efficient query from the database point of view.

5. EXPERIMENTS AND DISCUSSIONS

In this section, we present the simulation results of ComBD and ComCL in detail. For the first, the comparisons between four different combinations of intermediate features are presented, together with the performance evaluation of key concepts FMPI and ASCI. For the latter, the experiments involve four representative commercial categories: Automobile, Finance, Healthcare, and IT. In terms of ComID, the GoF signature is evaluated.

5.1 ComBD

5.1.1 TV commercial video database

We have built a TV commercial video database, which comprises 499 individual commercials covering 191 different ones. These commercials have extensively covered three concepts: Ideas (e.g., vehicle safety), Products (e.g., vehicles, food items), and Services (e.g., banking, insurance). All commercial clips are collected from TRECVID'05 corpus. It is a heterogeneous video dataset of 169 hours of video taken from 6 different sources: LBC, CCTV4, NTDTV, CNN, NBC, and MSNBC. With a commercial detector, we can quickly locate commercial segments, each comprising a series of commercials. Those segments are cut and recompressed.

TABLE I EXPERIMENTAL RESULT ON ASCI RECOGNIZER

	Alignment (YES or NO)	Precision (%)	Recall (%)	F1 (%)	Overall Accuracy of ASC and Non-ASC (%)
K-L	NO	72.8	76.6	74.6	79.8
K-L	YES	76.7	81.8	79.2	84.0
HMM	NO	76.1	80.5	78.2	83.6
HMM	YES	79.5	84.9	82.1	87.9

5.1.2 FMPI classification results

Our FMPI recognizer has achieved a promising accuracy up to F1 = 89.6% (recall = 88.25%, precision = 91.00%) over 4632 images comprising 1046 FMPI frames and 2987 Non-FMPI frames manually culled from the commercial video database. This accuracy is yielded by averaging the results of ten runs formed by conducting 10 times of random half-and-half training/testing partition. LIBSVM [41] is utilized to accomplish C-SVC learning. Radial basis function (RBF), $\exp(-\gamma\|x_i-x_j\|^2),\gamma>0$, is used. We are required to tune four parameters, i.e., γ, penalty C, class weight w_i, and tolerance e. w_i is for weighted SVMs to deal with unbalanced data, which set the cost C of class i to $w_i\times C$. e sets the tolerance of termination criterion. Class weights are set as $w_1=5$ for FMPI class and $w_0=1$ for Non-FMPI class. e is set to 0.0001. γ is tuned between 0.1 and 10 while C is tuned between 0.1 and 1. An optimal pair of $(\gamma,C)=(0.6,0.7)$ is set.

In order to evaluate the effects of low-level visual features on the performance, a set of recall/precision curves are yielded by using different visual features and different pairs of (γ,C) as shown in Fig. 10. Two curves of "Color" and "Texture" have demonstrated the individual capability of color and texture features. Texture features play a comparatively important role. Combining color and texture features significantly improves performance. Comparing "Color + Texture" with "Color + Edge" indicates that edge is less effective than texture. However, the performance is further improved more or less by fusing Color, Texture, and Edge as illustrated by the curve of "Color + Edge + Texture". Edge is a useful complement of texture.

5.1.3 ASCI classification results

To intuitively indicate low-level audio features' effectiveness, we calculate the K-L distances from a small set of ASC and Non-ASC samples as illustrated in Fig. 11. The duration of each sample is 2 seconds. Two *pdfs* (one Gaussian) are computed for the left- and right-side windows of one second. The sampling rate of 20 ms unit with a 10 ms overlap is applied. As shown in Fig. 11, two clusters of different K-L distances are delineated clearly.

Although the K-L distance can explicitly and quantitatively measure ASC, the temporal context is not utilized. HMM is a powerful model to characterize the temporally non-stationary but learnable and regular patterns for the audio signal. In experiments, performance comparisons are conducted between a K-L based method and an HMM-based one, between before- and after- alignment.

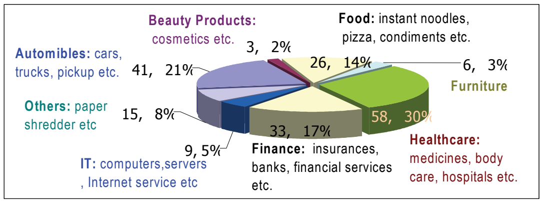


Fig.13 TV commercial category distribution

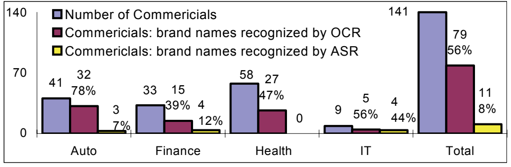


Fig.14 Commercials with brand names recognized by ASR & OCR

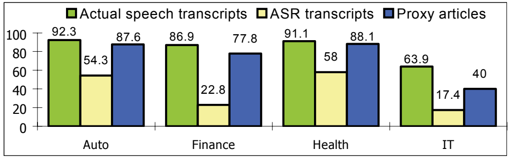


Fig.15 F1 values of three classifications based on different sources

Table I lists the results. The dataset comprises 2394 Non-ASC samples and 1932 ASC samples. A half-and-half training/testing partition is applied. The HMM structure of 8 hidden states and 12 mixture components is used. With an alignment process, the F1 of ASC and the overall accuracy are increased by 3.9% — 4.6%. Against K-L distance metric, HMM can improve the F1 of ASC and the overall accuracy by 2.9% — 4.2%. The alignment plays an important role. The overall accuracy of ASC and Non-ASC is of interest as those two probabilities jointly contribute to a boundary classifier. A promising overall accuracy of 87.9% has been achieved by HMM along with an alignment process.

5.1.4 *ComBD classification results*

Our experimental dataset produces 498 true boundaries and 2050 false ones. For unbalanced data, class weights are set as $w_1 = 5$ for true boundary class and $w_0 = 1$ for false boundary class.

Fig. 12 illustrates the simulation results of ComBD. A promising accuracy of F1 = 89.22% (recall/precision = 91.00%/87.50%) is achieved via half-and-half training/testing with FMPI and ASCI only. The performance has provided a basis for a reliable ComBD system as FMPI and ASCI are independent of post-editing effects. A further F1 improvement from 89.22% to 93.70% is obtained by fusing FMPI, ASCI, Silence, and Black Frames whereas using Black Frames yields a poor result of F1 = 81.0% (recall/precision = 87.00%/75.80%). The inferior result of "ASCI+Silence+Black Frames" clearly shows the improvement by introducing FMPI.

For comparison purpose, we attempt to use audio features only to detect commercial boundaries. The time series of low-level audio features originally extracted and arranged for training ASC/Non-ASC recognizers are fed into two HMM models for representing true and false boundaries. An accuracy of F1 = 74% is achieved. It is less than the accuracy obtained by using FMPI features only. However, combining FMPI and ASCI makes a big difference.

The performance may vary with different streams. However, a heterogeneous dataset has been employed for a fair evaluation. The use of only Black Frames (as suggested in [4] [14]) would produce even worse result (<< 81.0%) if they were not used as a delimiting flag, easily omitted by TV stations, to separate spots.

TABLE II EXPERIMENTAL RESULT ON TV COMMERCIAL CLASSIFICATION

	Auto	Finance	Health	IT	Count	Recall (%)
(a) Classification with manually recorded speech transcripts						
Auto	**38**	2	0	1	41	90.2
Finance	1	**28**	2	2	33	84.8
Health	3	1	**50**	4	58	86.2
IT	0	1	3	**5**	9	55.6
Sum	42	37	59	8	141	
Precision (%)	94.5	89.2	96.6	75.0		**85.8***
(b) Classification with ASR transcripts						
Auto	**19**	22	0	0	41	46.3
Finance	9	**9**	13	2	33	27.3
Health	2	14	**31**	11	58	53.5
IT	1	1	5	**2**	9	22.2
Sum	29	46	49	14	141	
Precision (%)	65.5	19.6	63.3	14.3		**43.3***
(c) Classification with proxy articles						
Auto	**35**	3	2	1	41	85.4
Finance	3	**25**	2	3	33	75.8
Health	3	1	**50**	4	58	86.2
IT	0	3	2	**4**	9	44.4
Sum	40	35	55	11	141	
Precision (%)	90.0	80.0	90.1	36.4		**80.9***

* OVERALL CLASSIFICATION ACCURACY

All ground truths (i.e., FMPI shots, ASC/Non-ASC at candidate points, and commercial boundaries) are manually labeled. Since two layers of training/testing (intermediate features and boundary classifier) are executed, a data partition scheme is taken as below. In order to deliver intermediate features to the boundary classifier, we employ two separate training/testing video datasets of equal size for two layers of computation like TRECVID [26], [30]. That is, we use one common dataset for training FMPI, ASCI, and boundary classifiers wherein one third of this dataset is allocated for validation; subsequently, the trained FMPI, ASCI, and boundary classifiers are applied to the other common dataset for testing. This scheme assures the accumulated negative effects of missed or false alarms from intermediate features can be finally incorporated and help fairly evaluate the overall performance of ComBD.

5.2 ComCL

5.2.1 *Commercial data observations and parameter*

From the commercial video database, we extracted 191 distinct English ones. By their advertised products or services, the 191 TV commercials are distributed in 8 categories, as shown in Fig. 13. Our experiments involve 4 categories: Automobile, Healthcare, IT , and Finance. Though they do not exclusively cover all commercials, they count up to 141 and 74% of total commercials. Also, the training documents for the 4 categories are accessible. Hence, they should be able to show how effective our approach is. For each category, we have collected 1,000 training documents from Reuters-21578 and 20 Newsgroup. Altogether the training documents amount to 4,000. At the phase of feature selection, the document frequency threshold is set to 2, and 9107 word features are selected. Prior to SVMs training, these 4,000 documents were evaluated by a 3-fold cross-validation to examine their integrity and qualification as training data. The cross-validation accuracy reached up to 96.9%, where RBF kernel was used and SVMs parameters C and γ were determined to be 8,000 and 0.0005.

The statistics show that, on average, ASR and OCR can provide 2.8 and 2.3 potential keywords for each automobile commercial, 4.5 and 2 for finance, 6.4 and 2.5 for healthcare, and 5.7 and 2.3 for IT, respectively. We empirically set both keyword selection parameters n and m to be 2. The recognition of brand names

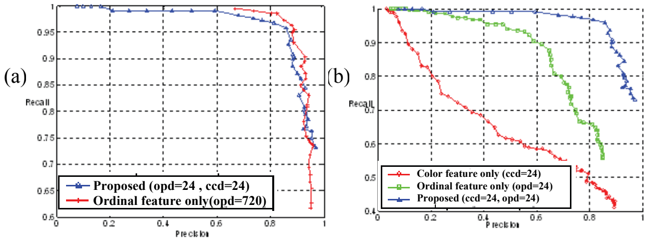


Fig.16 Performance of our signature versus ordinal features

plays an important role, as brand names are the best keyword candidates. Fig. 14 presents the number of commercials, in which ASR/OCR recognized their brand names successfully. It shows that OCR can recognize brand names in a considerable amount of commercials, especially in automobile ones. Overall, OCR can recognize brand names of 56% of total commercials.

5.2.2 Results evaluation and discussion

The classification based on manually recorded speech transcripts of commercials is firstly performed. As Table II (a) shows, except IT, all other categories achieve satisfactory results and the overall classification accuracy reaches up to 85.8%. The reason for lower accuracy in IT category lies in the mismatch of topic definition between the training documents and the testing commercials. In the training data, IT category mainly covers computer hardware and software. However, in the testing data, it includes other IT products, like printers and photocopy machines. ASR transcripts are also applied to text categorization. As Table II (b) shows, the ASR transcripts deliver bad results in all categories. Table II (c) lists the results by proxy articles. Compared with ASR transcripts, proxy articles have improved the performance drastically and the overall classification accuracy is increased from 43.3% to 80.9%. Fig. 15 displays the F1 values of classifications by three sources. For most categories, proxy articles deliver slightly lower accuracies than manually recorded speech transcripts. The accuracy differences imply that the errors in keyword selection and proxy article acquisition do occur, and however, they do not necessarily provoke serious degrades on the final performance.

5.3 ComID

To evaluate our signature's robustness, we have tried to identify 84 commercial clips in a 10.5 hour video collection. Given their exact boundaries, we have achieved 100% accuracy for matching amongst commercials. Moreover, we conduct the sliding window based matching plus the active search technique [40] over video streams. As shown in Fig. 16(a), our signature obtains comparable results with [38] whose feature has 3×720 = 2160 dimension. Our feature is 6×24 = 144 dimensional, 15 times smaller than that of [38]. And from Fig. 16 (b), compared with using ordinal or color features only, using combined features delivers better results. See [39] for details on the computational cost of feature extraction and fast search via active search and mrkd-tree index structure.

6. CONCLUSIONS

A TV commercial digest scheme of "ComBD+ComCL+ComID" is suggested to enrich the interaction between television and audiences. We will simulate a digital TV set-top box for live testing. A few open issues remain: 1) exploring the impact of production formats (e.g. demonstration, product alone, spokesperson, etc. [1]) on ComBD; 2) evaluating the role of shot coherence in ComBD; 3) seeking a systematic approach to accurate keywords selection; 4) utilizing computable visual concepts on scenes & object categories to classify commercials lack of textual semantics; and 5) investigating other text categorization approaches (e.g., Cluster by Committee, etc.) to address more commercial categories.

7. REFERENCES

[1] J.V. Vilanilam and A.K. Varghese, A*dvertising basics! A resource guide for beginners*. Response Books, New Delhi, 2004.

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

[3] L. Agnihotri, etc., "Evolvable visual commercial detector," *Proc. CVPR' 03*, Madison, Wisconsin, USA, pp. 79-84, vol.2.

[4] R. Lienhart, C. Kuhmunch, and W. Effelsberg, "On the detection and recognition of television commercials," *Proc. ICMCS'97*, Ottawa, Canada, pp. 509-516.

[5] H. Sundaram and S.-F. Chang, "Computable scenes and structures in films," *IEEE Tran. Multimedia*, 4(4):482-491, 2002.

[6] J. R. Kender and B.L. Yeo, "Video scene segmentation via continuous video coherence," *Proc. CVPR'98*, CA, USA, pp.367-373.

[7] M. Yeung and B.L. Yeo, "Time-constrained clustering for segmentation of video into story units," *Proc. ICPR'96*, Vienna, Austria, pp.375-380.

[8] A. Hanjalic, R.L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE Tran. CSVT*, 9(4):580-588, 1999.

[9] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Scene determination based on video and audio features," *Proc. ICMCS'99*, pp.685-690, vol.1.

[10] A. G. Hauptmann and M. J. Witbrock, "Story segmentation and detection of commercials in broadcast news video," *Proc. Conf. Advances in Digital Libraries*, Santa Barbara, 1998.

[11] L. Chaisorn, etc., "A two-level multi-modal approach for story segmentation of large news video corpus," *Proc. TREC Video Retrieval Evaluation*, Gaithersburg, MD, USA, 2003.

[12] X.-S. Hua, L. Lu, and H.-J. Zhang, "Robust learning-based TV commercial detection," *Proc. ICME'05*, Amsterdam, Netherlands, pp.149-152.

[13] A. Albiol, M.J.C. Fulla, A. Albiol, and L. Torres, "Commercials detection using HMMs," *Proc. Int. Workshop Image Analysis for Multimedia Interactive Services*, Portugal, 2004.

[14] S. Marlow, etc., "Audio and video processing for automatic TV advertisement detection," *Proc. Conf. Irish Signals and Systems*, Ireland, 2001.

[15] J. Wang, etc. "A robust method for TV logo tracking in video streams," in *Proc. ICME'06*.

[16] K. Matsumoto, etc., "Shot boundary determination and low-level feature extraction experiments for TRECVID 2005," *Proc. TREC Video Retrieval Evaluation'05*, Gaithersburg.

[17] H. Zhang, A. Kankanhalli, S.W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, 1(1):10-28, 1993.

[18] A. K. Jain, S.K. Bhattacharjee, and Y. Chen, "On texture in document images," *Proc. CVPR'92*, Champaign, IL, USA, pp.677-680.

[19] B.S. Manjunath and W.Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Tran. PAMI*, 18(8):837-842, 1996.

[20] J. Canny, "A computational approach to edge detection," *IEEE Tran. PAMI*, 8(6):679-698, 1986.

[21] V. Vapnik, *The nature of statistical learning theory*. Springer-Verlag, 1995.

[22] W. Wolf, "Key frame selection by motion analysis," *Proc. ICASSP'96*, Atlanta, Georgia, USA, vol. 2, pp. 1228-1231.

[23] T. Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Tran. Speech and Audio Processing*, 9(4):441-457, 2001.

[24] HTK speech recognition toolkit. [Online] Available: http://htk.eng.cam.ac.uk/.

[25] L. Lu, H. Jiang, H.J. Zhang, "A robust audio classification and segmentation method," *Proc. ACM Int. Conf. Multimedia*, Canada, 2001, pp. 203-211.

[26] T.-S. Chua, etc., "TRECVID 2005 by NUS PRIS," in *Proc. TREC Video Retrieval Evaluation*, Gaithersburg, MD, USA, 2005.

[27] S.-F. Chang, W. Chen, and H. Sundaram, "Semantic visual templates - Linking features to semantics," *Proc. ICIP'98*, Chicago, USA, pp.531-535.

[28] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, J. S. Jin, "A unified framework for semantic shot classification in sports video," *IEEE Tran. Multimedia*, 7(6):1066-1083, 2005.

[29] M.R. Naphade and T.S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE Tran. Multimedia*, 3(1):141-151, 2001.

[30] A. Amir, etc., "IBM research TRECVID-2005 video retrieval system," *Proc. TREC Video Retrieval Evaluation*, Gaithersburg, MD, USA, 2005.

[31] H. Yang, L. Chaisorn, Y. Zhao, S.-Y. Yeo, and T.-S. Chua, "VideoQA: question answering on news video," *Proc. ACM Int. Conf. Multimedia'03*, Berkeley, CA, USA, pp. 632-641.

[32] C.-S. Xu, J. Wang, Y. Li, K. Wan, and L.-Y. Duan, "Live sports event detection based on broadcast video and web-casting text," *Proc. ACM Int. Conf. Multimedia'06*, CA, USA.

[33] N. Babaguchi, K. Kawai, and T. Kitabashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Tran. Multimedia*, 4(1):68-75, 2002.

[34] Reuters-21578 Text Categorization Test Collection.
[Online] Available: http://www.daviddlewis.com/resources/testcollections/reuters21578/

[35] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, 54(1):1-47, 2002.

[36] K. Lang, "Newsweeder: learning to filter netnews," *Proc. ICML'95*, USA, pp.331-339.

[37] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," *Proc. ECML'98*, Germany, pp.137-142.

[38] A. Hampapur, K. Hyun, and R. Bolle, "Comparison of sequence matching techniques for video copy detection," *Proc. SPIE'02*, pp.194-201, vol.4676.

[39] J. Yuan, L.-Y. Duan, Q. Tian, and C.-S. Xu, "Fast and robust short video clip search using an index structure," *Proc. of ACM MIR Workshop'04*, New York, USA, pp. 61-68.

[40] K. Kashino, etc., "A quick search method for audio and video signals based on histogram pruning," *IEEE Tran. Multimedia*, 5(3):348-357, 2003.

[41] LIBSVM. [Online] Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[42] C. Colombo, A. D. Bimbo, and P. Pala, "Retrieval of commercials by semantic content: The semiotic perspective," *Multimedia Tools and Applications*, 13(1):93-118, 2001.

[43] Jinhui Yuan, etc., "Tsinghua Univeristy at TRECVID 2005," *Proc. TREC Video Retrieval Evaluation*, Gaithersburg, MD, USA, 2005.