Digesting Commercial Clips from TV Streams

Ling-Yu Duan and Yan-Tao Zheng
Institute for Infocomm Research

Jinqiao Wang and Hanqing Lu
Chinese Academy of Sciences

Jesse S. Jin
University of Newcastle

A commercial system that performs syntactic and semantic analysis during a TV advertising break could facilitate innovative new applications, such as an intelligent set-top box that enhances the ability of viewers to monitor and manage commercials from TV streams.

Although the costs of creating, producing, and airing a TV commercial are staggering, television is one of the most cost-efficient media. Television's most outstanding attribute is its ability to reach a vast number of consumers at the same time. Its other advantages are impact, credibility, selectivity, and flexibility.[1] But advertisers face a serious problem. The recent development of digital video recording and playback systems has provided a means for the viewer to skip the advertisements, either manually or by automatic means. According to a Forrester Research's survey (see http://www.news.com/2100-1024-5200073.html) a majority of national advertisers plan to cut spending on TV commercials by 20 percent in the next five years due to these ad-skipping devices.

Our proposed scheme addresses two challenging tasks: commercial boundary detection (ComBD) and commercial classification (ComCL) in terms of advertised products and services. The first involves video parsing; the latter, semantic video indexing. In TV streams, a commercial block consists of a series of individual spots. We can view each spot as a semantic scene. Commercial video parsing detects such scene transitions within a block.

Semantic commercial video indexing can be likened to classified newspaper ads, where consumers can easily find useful information. Semantic commercial video indexing is designed to provide consumers with useful information through video content analysis techniques.

Various video clip matching methods can identify commercials (ComID) using any one of several existing methods, but one important issue is to identify different ad versions for a product or service effectively. In our system, we apply a visual concept—that is, an image frame marked with product information (FMPI)—which incorporates the commercial production knowledge to represent an ad.

Research shows that most people don't mind TV advertising in general, although they dislike certain types of commercials. With the advance of digital TV set-top boxes—in terms of powerful processors, large hard disks, and Internet access—consumers need a TV commercial management system that detects commercial segments, determines individual commercial boundaries, identifies and tracks new commercials, and summarizes commercials by removing repeated instances. Given a decent interface, this system could change a TV viewer's passive relationship to advertising. A user could apply positive actions (for example, search, browse, summarize) to the commercial video archive, which could indirectly improve the reach of TV commercials.

Besides business issues, the industrial applicability of digesting ads depends upon whether people are willing to browse video commercials. This willingness is related to the varying degrees of relevant and valuable information conveyed to the user. Unlike infomercials, the vast majority of brief spots—ranging in length from a few seconds to one minute—brand a product in the marketplace. We consider the digested ads to be the video-based alert for emerging products and services. When people browse ads and dig up an ad's offer of interest, they might use a search engine to collect additional information.

Commercial boundary detection

Commercial videos are characterized by dramatic changes in lighting, chromatic composition, and other factors, including shot

1070-986X/08/$25.00 © 2008 IEEE Published by the IEEE Computer Society

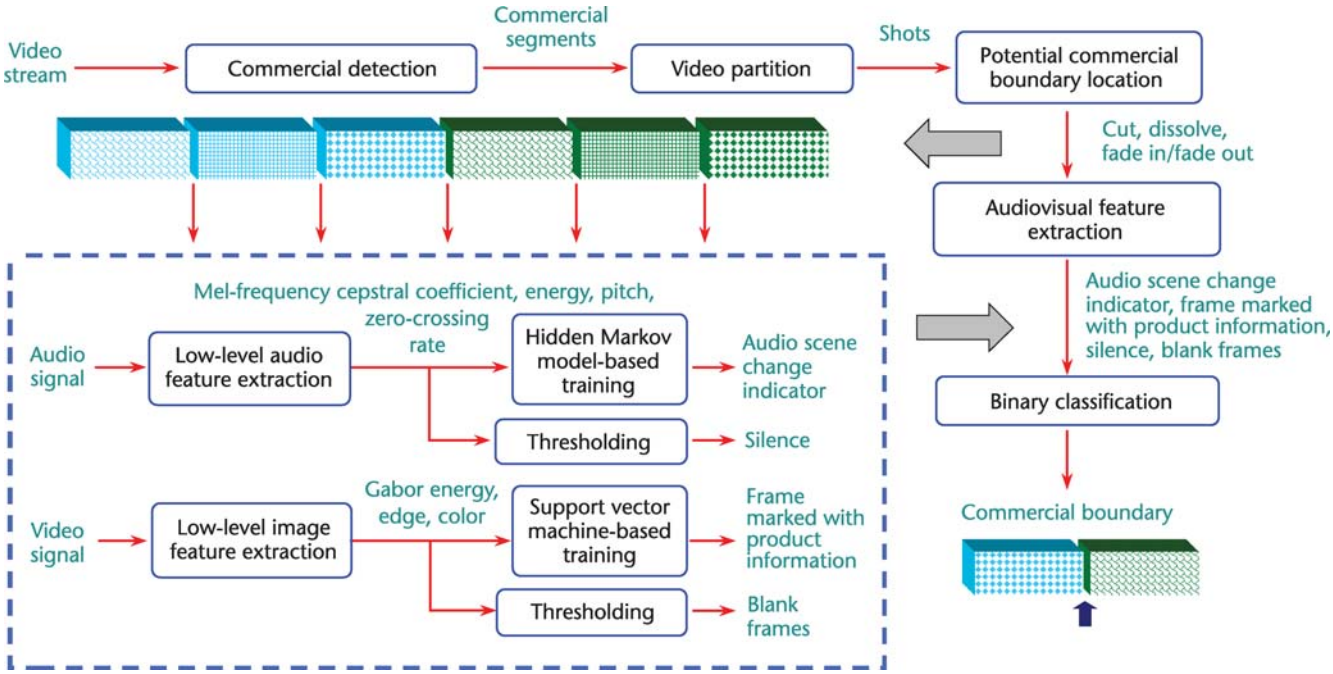

Figure 1. Detecting individual commercials' boundaries.

length, motion, sound, and, of course, creative stories. These factors make existing scene transition detection methods less effective for individual ComBD, as shots lack uniformity. Hence, our approach reduces the problem of commercials' boundary detection to that of a binary classification of true versus false scene changes at candidate positions consisting of video shot change points. It's reasonably assumed that a TV commercial scene transition always comes with a shot change (that is, cuts, fade-ins and fade-outs, and dissolves).

Solution and framework

Figure 1 illustrates the framework. Our system extracts multimodal features within a symmetric window at each candidate point. While different or multiscale window sizes can be applied to different feature types, we apply supervised learning to fuse the multimodal features. Two techniques—audio scene change indicators (ASCI) and FMPIs—can help characterize the computable video contents of interest to signify an individual commercial's boundaries. Because it isn't feasible to decipher a commercial video's temporal arrangement through a predefined set of shot classes, midlevel features condense high-dimensional, low-level features by using adequate classifiers to generate as many useful concepts as possible supported by commercial video production rules or knowledge.

While general commercial detection is a preliminary stage in our approach, other approaches have been proposed elsewhere.[2-5] For example, one study reported a 92 percent accuracy rate on a heterogeneous dataset.[2] Our implementation relies on the detection and tracking of TV logos because TV logos often don't appear during commercials. We achieved satisfactory results of $F1 = 97.76$ to 99.80 percent on opaque, semitransparent, and animated TV logos from eight TV channels, including NBC, CNN, and MSNBC. Note, $F1$ is an even combination of precision and recall. It is defined as $2 \cdot \text{precision} \cdot \text{recall}/(\text{precision} + \text{recall})$. More details can be found elsewhere.[6] As our focus is on the boundaries of individual spots within a commercial break, the use of logo-based commercial detection does not affect this scheme's applicability.

Frames marked with product information

We present a visual concept FMPI to determine those candidate regions containing a commercial boundary.

Using an FMPI to locate the most probable boundary candidates. We use FMPIs to describe those images containing visual information explicitly illustrating an advertised product or service. These frames express visual information in three ways: text, computer graphics, and frames from live footage of real

Figure 2. Image frames marked with product information.

things and people. Figure 2 shows FMPI frame examples. The textual section may consist of brand name, store name, address, telephone number, cost, and so on. Alongside the textual section, a drawing or a photo of a product might be placed with computer graphics. Live footage of real things or people is usually combined with graphics to avoid impersonality.

Production rules reveal the spatial relationship between the presence of FMPIs and individual commercials' boundaries. For convenience, we define the video shot containing at least one FMPI as an FMPI shot. Sometimes we have trouble determining the precise offer in a commercial, but an FMPI shot is a useful prop. Some commercials might contain irregularly interposed FMPI shots, as branding can be reinforced by endless repetition. Occasionally, a commercial might show an FMPI shot at the beginning. We therefore can consider an FMPI shot to be an indicator that can help determine a much smaller set of commercial boundary candidates from large amounts of video shot changes.

Constructing an FMPI recognizer. We rely on the combination of texture, edge, and color features to represent an FMPI. As the frame's layout is a significant factor in distinguishing an FMPI, it's beneficial to incorporate the spatial information. One common approach is to divide an image into subregions and impose positional constraints on the image comparison (a process called *image partitioning*). Dominant colors help construct an approximation of color distributions. We can easily identify these distributions from color histograms. Because Gabor filters exhibit optimal location properties in the spatial domain as well as in the frequency domain, we use them to capture rich texture in FMPIs.[7] Edge is a useful complement to texture when an FMPI produces stand-alone edges as a contour of an object, as texture relies on a collection of similar edges. Combined features yield better results than using a single feature.

Figure 3 illustrates our feature-extraction procedure. We determine dominant colors by selecting maximum bin values and edge densities using Canny edge detection. We apply both dominant colors to the subimages and the whole image. For local texture features, we apply Gabor filters (one with center frequency and four with equidistant orientations) to each subimage. By combining local features and global ones, our implementation constructs 141 feature dimensions. More details can be found elsewhere.[8]

To train the FMPI recognizer, we use supporting vector machines, which work well for data with a large number of features and contain fewer parameters. Our implementation resorts to the C-support vector classification.[9] To determine an FMPI shot, we can apply FMPI recognition to keyframes only.

Audio scene change indicator

We model audio scene changes to facilitate identifying commercial boundaries.

Using ASCI to characterize audio changes occurring at commercial boundaries. Different TV commercials often exhibit dissimilar

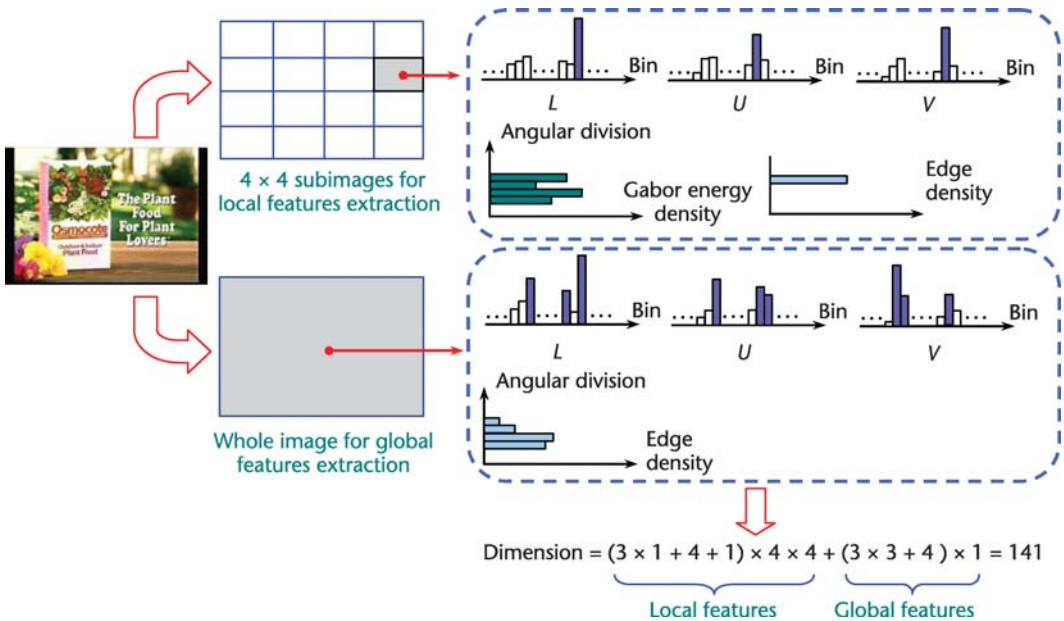

Figure 3. An example of low-level visual feature extraction for training the frame marked with a product information recognizer.

audio characteristics. but a proper modeling of audio scene changes (ASCs) can facilitate the identification of commercial boundaries.

Given an audio segment (for example, 4 seconds) at a candidate boundary, ASCI provide a probabilistic representation of ASCs. As Figure 4 shows, we use a hidden Markov model (HMM) to train two models for two dynamic ASC and non-ASC patterns. Our method classifies any unknown segments using the model with the highest posterior probability.

Like FMPI, ASCI is an indicator but cannot secure true boundaries due to dynamic audio characteristics inherent to commercial videos. Our solution is to fuse multimodal features—for example, ASCI + FMPI.

Using HMM to train recognizers. The HMM we use to train ASC and non-ASC recognizers is a Gaussian mixture (left to right). We use a diagonal covariance matrix to estimate the Gaussian mixture distribution. Suppose we have two HMM models for representing ASC and non-ASC, the forward-backward algorithm generates two likelihood values of an observation. We use the HTK toolkit (see http://htk.eng.cam.ac.uk/) for this process.

Currently, our ASCI considers 43 dimensional audio features comprising

- Mel-frequency cepstral coefficients (MFCCs) and their first and second derivates (36 features);

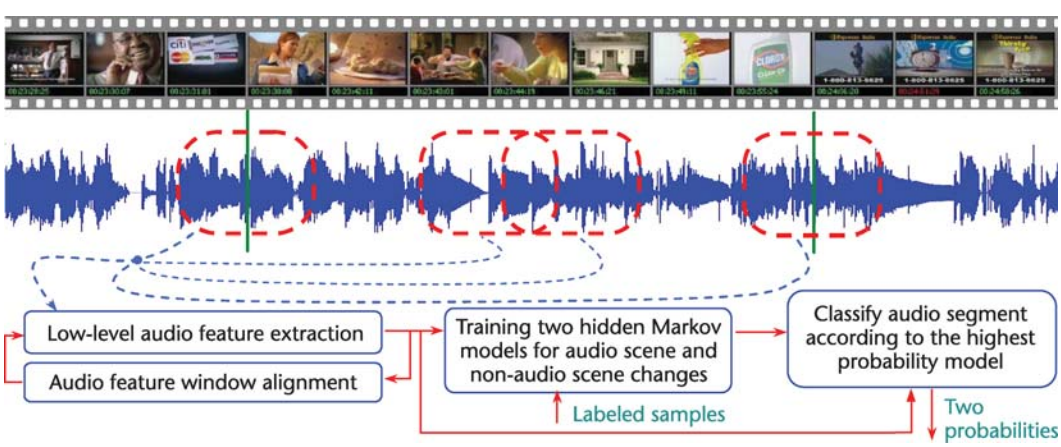

Figure 4. Training an audio scene change indicator.

- mean and variance of short-time energy log measure (two features);
- mean and variance of short-time zero-crossing rate (two features);
- short-time fundamental frequency, or pitch, (one feature);
- mean of the spectrum flux (one feature); and
- harmonic degree (one feature).

Readers are referred elsewhere for more feature details.[10] We segment an audio signal into a series of successive 20 milliseconds analysis frames by shifting the sliding window of 20 ms with an interval of 10 ms. Our approach computes features for each analysis frame. Within each analysis frame we compute short-time energy, zero-crossing rate, spectrum flux, and harmonic peaks once every 50 samples at an input sampling rate of 22,050 samples per second where the size of sliding window is set to 100 samples. We calculate means and variances of short-time energy and zero-crossing rates for seven values from seven overlapping frames, and calculate the mean of spectrum flux for six values from seven neighbor frames. The harmonic degree is the ratio of the number of frames having harmonic peaks to the frame number seven. We directly compute pitch and MFCCs from each frame.

Aligning the audio feature window. Referring to Figure 4, we must address the alignment problem for two reasons. First, TV commercial boundaries have a maximum offset of ±0.25 to ±1.0 seconds between an ASC and its associated video scene change. Second, due to video production, a mixed soundtrack isn't necessarily synchronized to a video track. Thus, a symmetric window exactly at shot transitions cannot extract the most effective matching features from the nearby ASC.

An alignment procedure seeks to locate the most likely ASC point within the neighborhood of a shot change. Let W_i and W_j be two audio analysis windows, and their difference denoted by $d(W_i,W_j)$. By using the Kullback-Leibler (K-L) distance metric, we can write the difference as

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

where $p_i(x)$ and $p_j(x)$ denote the probability distribution functions estimated by the features extracted from W_i and W_j. We first consider ASC computing at one scale of an analysis window (that is, all the audio analysis windows are of a fixed size). Let W_i with $i = 1, 2,\ldots, N$ 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 as different change peaks occur for different window sizes. Since we don't know a priori what sound we are analyzing, we use multiscale computing. We first make use of multiple window sizes $\{Win_{scale}\}_{scale=1,\ldots,S}$ to yield a cluster of difference value series, denoted by $\{Distance_{scale}\}_{scale=1,\ldots,S} = \{D_{i,scale} | i = 1,\ldots, N_{scale}\}_{scale=1,\ldots,S}$. We then normalize each series of $Distance_{scale}$ to [0, 1] through dividing difference values $D_{i,scale}$ by the maximum of each series $Max(Distance_{scale})$. We then determine the most likely ASC point ω by locating the highest accumulated values.

We calculate the probability $p(\omega_\lambda)$ of an ASC point, where ω_λ is the candidate window position, as

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

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

where M denotes the total number of candidate window positions and 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 [−500 ms, 500 ms]. That is, $WS = 1{,}000$. We extract and arrange audio features within the adjusted 4-second feature windows and employ 11 scales (that is, $S = 11$) 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. We use a single Gaussian probability distribution function and apply a 20 ms sliding window with an interval of 10 ms.

Silence and black frames

We can separate spots with a short break of several black frames or moments of silence or reduced audio in some TV channels.[4] We detect silence by examining the audio energy level. We measure short-time energy functions every 10 ms and smooth them using an eight-frame finite impulse response filter. The smoothing imposes a minimum length constraint on the silence. We apply a threshold and categorize the segment that has energy below the threshold as silence. We detect a black frame by measuring the mean and the variance of intensity values within a frame, using a similar threshold method. A sequence of consecutive black frames (eight, for example) rises above the threshold and indicates the presence of black frames to the system.

Feature fusion

Our approach fuses the features of FMPI, ASCI, silence, and black frames—extracted from a temporal window at a candidate boundary—with a binary classifier as indicated in Figure 1. Our implementation relies on an SVM classifier to accomplish this fusion. To evaluate the effectiveness of FMPI and ASCI, we empirically conducted the fusion by combining different features.

ASCI yields two probability values: p(ASC) and p(non-ASC). Silence and black frames also yield two values p(Silence) and p(Black Frames) to indicate their presence within a 4-second temporal window (with 2 seconds each for left and right shots). For FMPI, $2 \cdot n$ neighbor video shots at a candidate boundary (Left n shots, right n shots) produce $2 \cdot n$ values $\{p_i(\text{FMPI})\}_{i = 1,\ldots, 2n}$ to indicate the presence of FMPI shots. The complete feature is $2 \cdot n + 4$ dimensional, where we empirically use $n = 2$.

The fusion procedure purely relies on SVMs and doesn't involve any feature selection, weighting, or rules. The simplicity of this system derives from concepts that abstract commercial production knowledge, and we are considering other linear classifiers and linear fusion schemes as well.

ComCL by-products and services

Compared with news or sports, commercials are essentially creative in terms of copyrighting and production techniques. We cannot use intermediate visual features or specialized concept detectors to model intrinsic semantics of commercials using audiovisual features, which means we must resort to extrinsic knowledge to narrow the semantic gap.

Solution and framework

Textual resources are becoming a useful channel for event detection[11,12] and high-level retrieval.[13,14] The textual sources can be an acoustic speech recognition (ASR) or optical character recognition (OCR) transcript,[13] closed caption,[12] and Web-casting text.[11] The use of textual information has two obvious advantages: clear linkages with semantics, and many available text-based external knowledge databases—for example, WordNet (see http://wordnet.princeton.edu/), dictionaries, encyclopedia, and topic-wise document corpora like Reuters-21578 (see http://www.daviddlewis.com/resources/testcollections/reuters21578/).

Hence, we resort to textual resources for addressing commercial classification with respect to products or services. By using text, our approach transforms the problem of semantic video classification to that of automated text categorization (see Figure 5, on the next page).[15] We assume that ASR/OCR can deliver useful textual hints about advertised products and services. First, we parse the deficient transcripts of ASR/OCR to extract keywords, by which search is carried out to retrieve semantically informative articles from the Internet. The commercial category information is enriched by the document representation of retrieved articles. Second, we use topic-wise documents from public corpora or from other external sources such as the Internet. Finally, we train text categorizers to determine the commercial category. Readers are referred elsewhere for examples.[8]

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 list L_i of nouns and noun phrases from S_i with the natural language processor. It selects keywords $K_i(kw_{i1},\ldots, kw_{i\ell})$ applying the following steps:

1. Check S_i against a predefined dictionary of brand names;
2. If the brand name occurs in S_i, select it as the only keyword kw_i and search it on Wikipedia (see http://en.wikipedia.org/wiki);

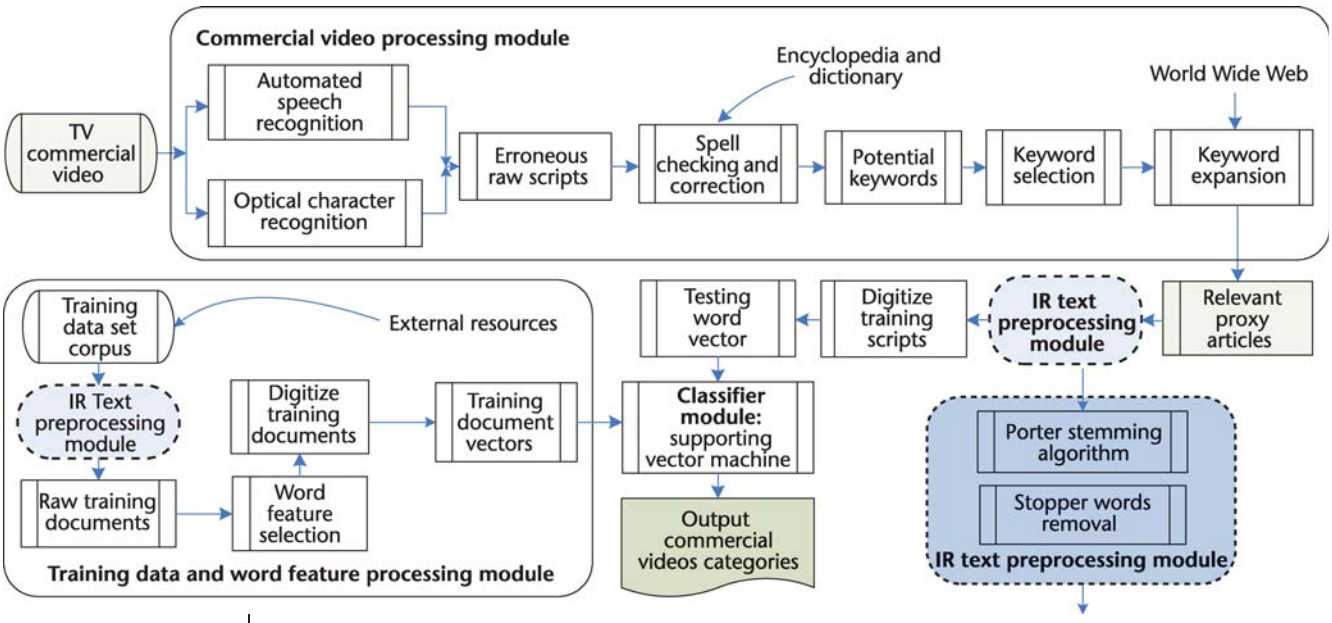

Figure 5. A TV commercial classification approach.

3. Otherwise, heuristically select from L_i the n nouns and noun phrases with the largest font size from OCR and the last m from ASR as keywords to search using a Web search engine.

We use Google as a search engine because its superior performance assures the searched articles' relevancy. Among returned articles, our approach selects the one with the highest relevancy rating as d_i, which we denote as the proxy article of $TVCom_i$. Exploiting d_i reduces the problem of TV commercial video classification to that of text categorization.[15] That is, we 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 predefined commercial category set 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.

Function modules

The commercial classification framework is composed of four major modules.

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). The PSA removes the common morphological and inflexional endings from words in English so that different word forms map to the same token. The SWRA eliminates words of little or no semantic significance—such as ''the,'' ''you,'' and ''can.'' Both testing and training documents go through this module before any other process.

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 first extracts the raw semantic information via ASR and OCR. The accuracy of OCR depends on the image character resolution. Large-size text contains more significant information than small. As shown in Figure 6, it's easy for OCR to recognize the large-size text ''Free DSL Modem, Free Activation,'' which contains more category related information than does the small and difficult-to-recognize text ''after rebates with 12 months commitment.'' Hence, our approach selects the n nouns and noun phrases with the largest font size from OCR as keywords. Subsequently, we apply spell checking and correction to the transcripts. Then, we correct the misspelled vocabulary terms and remove the terms not found in dictionaries. We use an English-language dictionary and encyclopedia for spell checking, as a dictionary might not include nonvocabulary terms, such as brand names. With the corrected transcript S_i, we get the proxy article d_i, from which we generate the testing vector.

Training data and word feature module. This module generates the training dataset and word feature space for text categorization. For the training dataset, we construct topic-wise document corpora from available public IR corpora or related articles manually collected from the Web. Currently, we use the categorized Reuters-21578 and 20 Newsgroup (see http://people.csail.mit.edu/jrennie/20Newsgroups/) corpora. In this way, the training corpus can possess large amounts of training documents and cover wide topic areas. The topics of these corpora might not exactly match the categories of TV commercials. Our solution is to choose topics related to the commercial category and combine them to 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 to form the training set for the finance category.

Next, we employ a document frequency technique to select word features on the 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, we normalize each document vector to eliminate the influence of different document lengths.

Classifier module. The classifier module performs text categorization of proxy articles and determines the categories of respective TV commercials. Sebastiani reviews various text-categorization techniques and reports that SVMs deliver consistently outstanding performance.[15] We use SVMs as the classifier in our implementation. Joachims presented the promising characteristics of SVMs to demonstrate their suitability for text categorization: They can handle high-dimensional input spaces—for example, 10,000 dimensions—and sparse document vectors.[16]

ComID

The boundaries of individual TV commercials reduce ComID to video clip matching. To address color distortion and different versions from postediting effects, we use a group-of-frames (GoF) compact signature to characterize dynamic spatiotemporal patterns. Our signature consists of ordinal and color features. Ordinal features contain spatial information, which is inexpensive to acquire. Color features involve color-range information by accumulating color histograms. Although a color histogram itself is vulnerable to color distortion, our experiments have shown the combination of ordinal and color features improves ad identification.

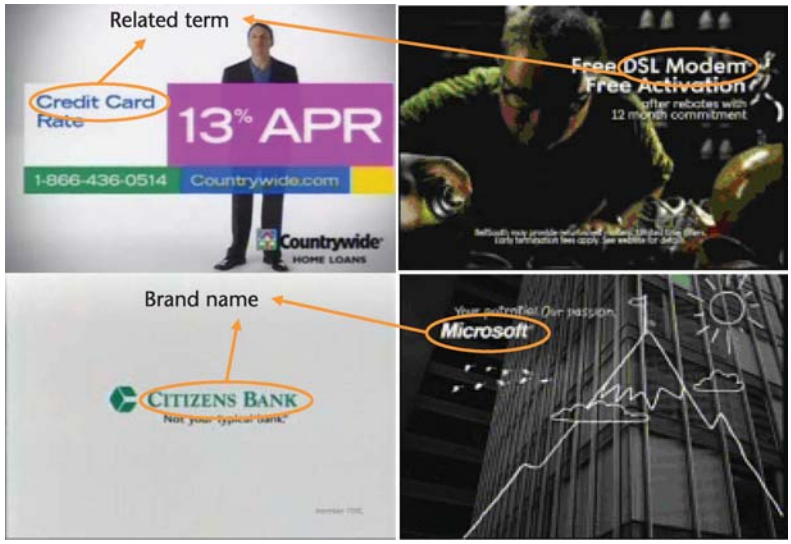

Figure 6. Examples of key frames containing significant semantic information.

To extract ordinal features, we reduce each frame to 2 × 2 pixels. For each *Y*, *Cb*, or *Cr* channel, we calculate the average pixel values within each subimage; we then use the ordinal measure process.[17] Given a GoF, for each channel $c = Y, Cb, Cr$, the ordinal pattern distribution (OPD) histogram

$$H_c^{opd}$$

is formed as

$$H_c^{opd} = (h_1, h_2, \ldots, h_l, \ldots, h_N)$$
$$0 \le h_i \le 1 \quad \text{and} \quad \sum_i h_i = 1$$

where h_i is the normalized bin value indicating the occurrences of an ordinal pattern i and $N = 4! = 24$ is the OPD histogram dimension, that is, the number of possible patterns. The total dimension of ordinal features is thus $3 \times 24 = 72$.

The advantages of an OPD histogram representation are twofold. First, it's robust against frame-size changes and color shifting. Second, its contour can characterize a video clip in a global

manner, while it's insensitive to video frame-rate changes and various local frame changes, unlike keyframe-based representations.

For color features, we employ the cumulative color distribution (CCD) histogram H_c^{ccd} defined as

$$H_c^{ccd} = \frac{1}{M} \sum_{i=b_K}^{b_{K+M-1}} H_i(j) \quad j = 1, \ldots, B$$

where H_i denotes the color histogram of an individual frame i, M is the total number of frames within a video segment, and B is the color bin number. In our experiments, we use $B = 24$ (a uniform quantization). Hence, the total dimension of color features is $3 \times 24 = 72$, by considering three color channels.

Given query clip Q and candidate clip S from commercial databases, the similarity is experimentally defined as the reciprocal of linear combination of the average distance of OPD and the minimum distance of cumulative color distribution among three channels:

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

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

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

where Euclidean distance $D(\cdot, \cdot)$ is used, w denotes the weight. In experiments, we use $w = 0.5$.

Another useful clue for ComID is FMPI. As discussed previously, FMPI shots highlight the commercial's offer. To promote a product or service, advertisers might use different storytelling videos to generate different ad versions where our proposed GoF signature at the clip level cannot fulfill ComID. Extensive experimental observations—for example, Cannes Lions Live ad corpus (see http://www.canneslions.com/winners_site/film/)—have indicated FMPI shots are often kept uniform among different ad versions to communicate persistent messages. Hence, we might constrain the GoF signature computation to those FMPI shots within a commercial. We would apply the same similarity measure.

ComBD experiments

We present the empirical results of detecting commercial boundaries.

TV commercial video database

We have built a TV commercial video database of 499 individual commercials covering 390 distinct commercials—as some commercials contained more than one instance These commercials cover three content concepts: ideas (for example, vehicle safety), products (for example, vehicles and food items), and services (for example, banking and insurance). We collected these commercial clips from the Text Retrieval Conference (TREC) video-retrieval evaluation 2005 corpus.

FMPI classification results

Our FMPI recognizer has achieved a promising accuracy up to $F1 = 89.6$ percent (recall/precision = 88.25/91.00 percent) over 4,632 images comprising 1,046 FMPI frames and 2,987 non-FMPI frames. We determine this accuracy by averaging the results of 10 random half-and-half training and testing partitions. We use LIBSVM (see http://www.csie.ntu.edu.tw/~cjlin/libsvm/) to accomplish C-support vector classification learning. We also use radial basis function $\exp(-\gamma\|x_i - x_j\|^2)$, $\gamma > 0$ and we must tune four parameters: γ, penalty C, class weight w_i, and tolerance e. We weight w_i for SVMs to deal with unbalanced data, which sets the cost C of class i to $w_i \times C$. e sets the tolerance of termination criterion. We set class weights as $w_i = 5$ for the FMPI class and $w_0 = 1$ for the non-FMPI class. We set e to 0.0001 and tune γ between 0.1 and 10 and C to between 0.1 and 1. We set an optimal pair of $(\gamma, C) = (0.6, 0.7)$.

As Figure 7 shows, a set of recall precision curves are yielded by using different visual features and different parameter pairs (γ, C). Compared with color features, texture features play a more important role. Combining color and texture features significantly improves the performance, which is further enhanced by fusing color, texture, and edge (which is a useful complement of texture).

ASCI classification results

We compare performance between a K-L- and HMM-based methods and between before and after alignment. Table 1 lists the results. The dataset comprises 2,394 non-ASC samples and 1,932 ASC samples. We applied a half-and-half training and testing partition and use the HMM structure of eight hidden states and 12 mixture components. Using an alignment

process increases the $F1$ of ASC and the overall accuracy by 3.9 to 4.6 percent. Against a K-L distance metric, HMM can improve the $F1$ of ASC and the overall accuracy by 2.9 to 4.2 percent. The alignment plays an important role. In addition, the overall accuracy of ASC and non-ASC is important as those two probabilities jointly contribute to a boundary classifier. We achieved a promising overall accuracy of 87.9 percent with HMM along with an alignment process.

ComBD classification results

Our experimental dataset produces 498 true boundaries and 2,050 false ones. For unbalanced data, we set class weights as $w_1 = 5$ for a true boundary class and $w_0 = 1$ for a false boundary class.

Figure 8 (next page) illustrates the simulation results of ComBD. We achieved a promising accuracy of $F1 = 89.22$ percent (recall/precision = 91.00 percent/87.50 percent) via half-and-half training and testing with FMPI and ASCI only. This performance provides a basis for a reliable ComBD system as FMPI and ASCI are independent of postediting effects. We obtain a further $F1$ improvement from 89.22 percent to 93.70 percent by fusing FMPI, ASCI, silence, and black frames; whereas using black frames only yields a poor result of $F1 = 81.0$ percent (recall/precision = 87.00/75.80 percent). The inferior result of ASCI + silence + black frames clearly shows the improvement by introducing FMPI.

Performance can vary with different streams. However, we employed a heterogeneous dataset for a fair evaluation. The use of only black frames (as suggested elsewhere[4]) would produce an even worse result—less than 81.0 percent—if they weren't used as a delimiting flag, easily omitted by TV stations, to separate spots.

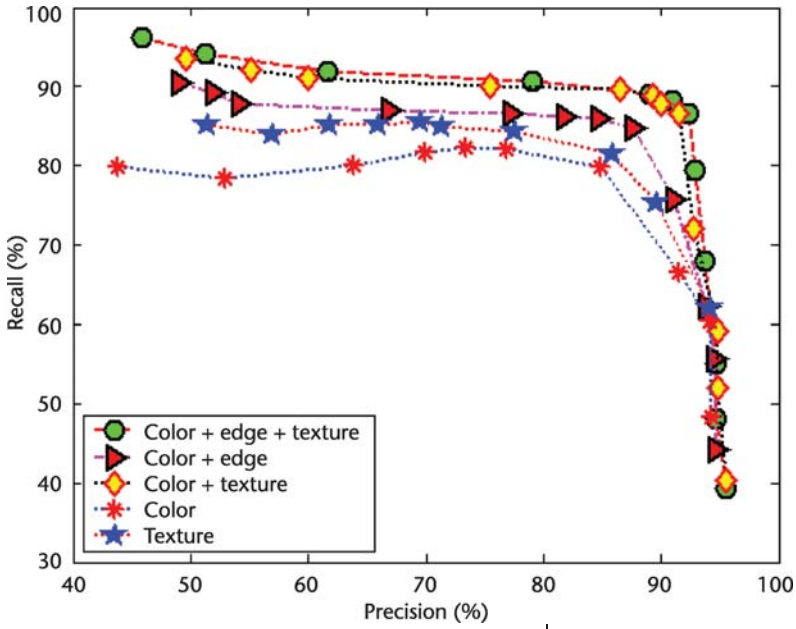

Figure 7. The recognition yield results for a frame marked with product information. Recognition using different features and parameters.

ComCL experiments

We present the empirical results of classifying commercials.

Commercial data observations and parameters

From the commercial video database, we extracted 191 distinct English ones. The 191 TV commercials are distributed in eight categories based on their advertised products or services. Our experiments involve four categories: automobile, healthcare, IT, and finance. Although they don't exclusively cover all commercials, they include 141, 74 percent of total commercials. This is a large enough number to show the effectiveness of our approach.

It's straightforward to determine appropriate topics in the available corpus to match the automobile, healthcare, and finance categories and form the training documents. We did not include the food category because we couldn't find suitable topics in Reuters-21578 and 20

Table 1. Experimental result on an audio scene change (ASC) recognizer.

Method	Alignment	Precision (%)	Recall (%)	$F1$ (%)	Accuracy of ASC and non-ASC (%)
Kullback-Leibler	No	72.8	76.6	74.6	79.8
Kullback-Leibler	Yes	76.7	81.8	79.2	84.0
Hidden Markov model	No	76.1	80.5	78.2	83.6
Hidden Markov model	Yes	79.5	84.9	82.1	87.9

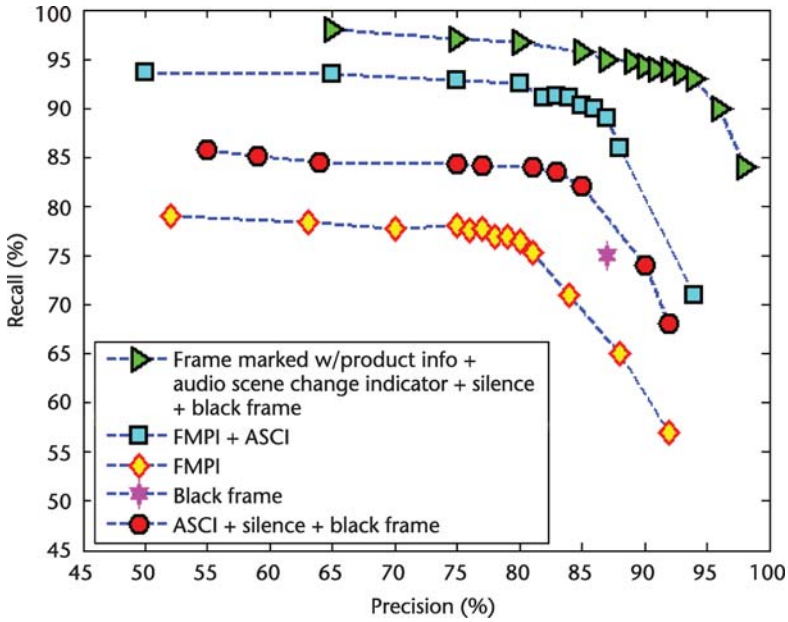

Figure 8. ComBD results yielded by using different features and supporting vector machine parameters.

Newsgroup. Thus we chose IT as the fourth category. An alternative is to collect relevant documents manually, a process we'll include in our future work.

For each category, we collected 1,000 training documents from Reuters-21578 and 20 Newsgroup. Altogether the training documents amount to 4,000. At the feature selection phase, we set the document frequency threshold to two. We examined 9,107 word features to determine their integrity and qualification as training data. With a three-fold cross-validation we reached an accuracy of up to 96.9 percent, using a radial basis function kernel and SVMs parameters C and γ of 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 two. The recognition of brand names plays an important role as brand names are the best keyword candidates. ASR can recognize brand names in 8 percent and OCR can recognize brand names in 56 percent of total commercials.

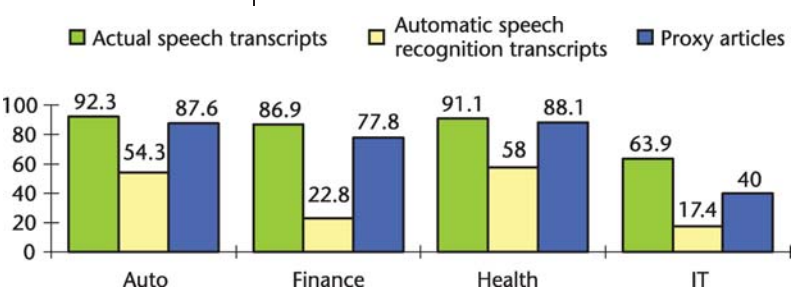

Figure 9. **F1** *values of three classifications based on different sources.*

Results evaluation

Figure 9 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, but do not necessarily provoke serious degradation on the final performance. Compared with ASR transcripts, proxy articles have improved the performance drastically; the overall classification accuracy is increased from 43.3 percent to 80.9 percent. With manually recorded speech transcripts, the overall classification accuracy reaches up to 85.8 percent. With the exception of IT, all other categories achieve satisfactory results. The reason for lower accuracy in the IT category lies in the mismatch of topic definition between the training documents and the testing commercials. In the training data, the IT category mainly covers computer hardware and software. However, in the testing data, it includes other IT products, such as printers and photocopy machines. In addition, ASR transcripts are applied to text categorization.

In general, accurate brand names help deliver correct classification. As the classification accuracy is 80.9 percent, we can roughly infer that the classification of some 16.9 percent (80.9 − 56 − 8 = 16.9 percent) of total commercials still can benefit from Google and text categorization when the brand names are not extracted.

ComID experiments

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 percent accuracy for matching among commercials. Moreover, we conduct the sliding-window-based matching plus the active search technique over video streams.[18] As shown in Figure 10a, our signature obtains comparable results with features having a $3 \times 720 = 2{,}160$ dimension.[17] Our feature is $6 \times 24 = 144$ dimensional, 15 times smaller than that of

Hampapur, Hyun, and Bolle.[17] As shown in Figure 10b, compared with using ordinal or color features only, using combined features delivers better results.

Conclusion

Ad agencies generate original ideas to represent products and services. This creative design or production could challenge our scheme's generalization of TV environments. But reducing ad boundary detection to binary classification, transforming video classification into text categorization, and using external knowledge to expand sparse textual semantics from ASR and OCR makes sense, in general. In addition, FMPI represents a utility widely used in ad production and the computable FMPI concept is useful in syntactic and semantic ad analysis, while ASCI provides a generic approach to address the alignment between shot transitions and potential audio scene changes nearby.

A few open issues remain. We need to explore the impact of production formats (for example, demonstration, product alone, spokesperson, and so on) on ComBD and the role of visual coherence of shots in ComBD. In the future, we will carry out ad classification on an extensive dataset and more categories. We will seek out a systematic approach to accurate keyword selection in ComCL. Finally, we'll introduce a collection of computable, visual concepts on scenes and objects that can classify an ad's lack of textual semantics. **MM**

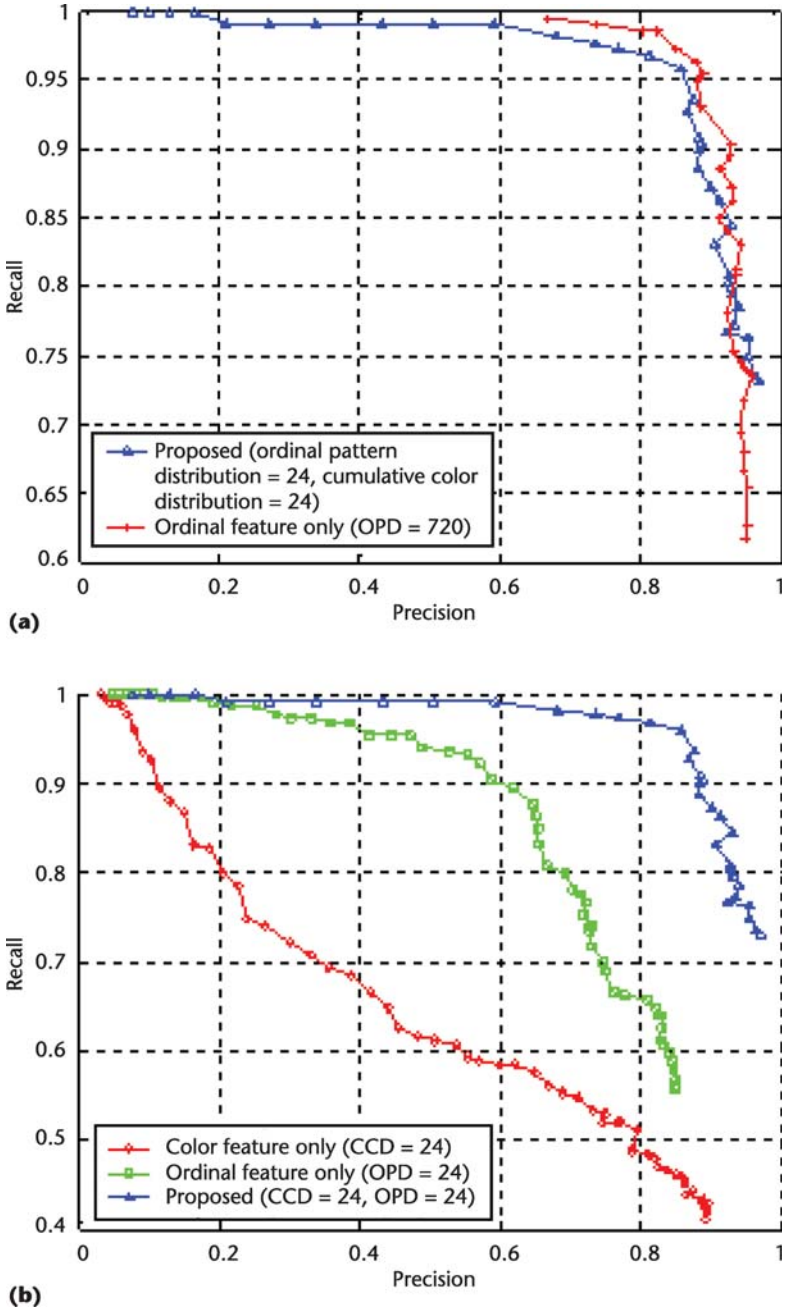

Figure 10. Performance of our signature versus ordinal features.

Acknowledgment

The National Natural Science Foundation of China (Grant No. 60475010) partially supported this work.

References

1. J.V. Vilanilam and A.K. Varghese. *Advertising Basics! A Resource Guide for Beginners*, Response Books, 2004.
2. M. Mizutani, S. Ebadollahi, and S.-F. Chang, ''Commercial detection in Heterogeneous Video Streams Using Fused Multimodal and Temporal Features,'' *Proc. IEEE Int'L Conf. Acoustic, Speech, and Signal Processing*, IEEE Press, 2005, pp. 157-160.
3. L. Agnihotri et al., ''Evolvable Visual Commercial Detector,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* vol. 2, IEEE Press, 2003, pp. 79-84.
4. R. Lienhart, C. Kuhmunch, and W. Effelsberg, ''On the Detection and Recognition of Television Commercials,'' *Proc. IEEE Int'l Conf. Multimedia Computing and Systems*, IEEE Press, 1997, pp. 509-516.
5. X.-S. Hua, L. Lu, and H.-J. Zhang, ''Robust Learning-Based TV Commercial Detection,'' *Proc. IEEE Int'l Conf. Multimedia and Expo*, IEEE Press, 2005, pp. 149-152.
6. J. Wang et al., ''A Robust Method for TV Logo Tracking in Video Streams,'' *Proc. IEEE Int'l Conf. Multimedia and Expo*, IEEE Press, 2006, pp. 1041-1044.

Related Work

In the context of video indexing and retrieval, TV commercial videos have been widely investigated.[1–6] To distinguish our work, we review related work in TV commercial video analysis and discuss the weaknesses of previous approaches in addressing commercial boundary detection (ComBD) and commercial classification (ComCL).

Most previous work on TV commercial video analysis focuses on automatically locating a commercial within a video stream to develop commercial skip applications.[1–4] These approaches exploit audiovisual features including blank frames, scene breaks, action,[3] and so on, to characterize commercial video segments. Subsequently, some approaches employ heuristic rules[3] or machine learning algorithms[1,4] to generate a commercial discriminator. Shots or image sequences are a commonly used level of granularity,[1,4] because some useful features—for example, scene change rate[1] or shot frequency[4]—are based on shots directly, and some statistically meaningful features—for example, blank frame rate and audio class histogram,[1] average and variance of edge change ratio, frame difference,[4] and so on—have to undergo the accumulation over a temporal window. Generally speaking, such features-based commercial detectors only allow roughly locating commercial breaks. To determine boundaries precisely, say one single shot, some recent work has introduced heuristic rules[4] and generative models[1] to implicitly or explicitly incorporate more constraints from commercial duration and temporal transition characteristics. Moreover, grouping semantically related shots (for example, program shots versus commercial shots) has been proposed as a postprocessing to refine the boundaries.[3,4] Despite successes in prototyping and performance tuning of a commercial detector, little existing work addresses structural and semantic analysis within a commercial segment.

We know of no previous work that exploits a commercial's promotional content. Colombo, Bimbo, and Pala's semiotic categorization of commercials is the work most focused on semantic content analysis of commercial videos.[6] They use heuristic rules in the commercial production to associate a set of perceptual features with four major commercial types, namely practical, playful, utopic, and critical. This work is different from our proposed approach to commercial classification. Regardless of the video story's meaning,[6] their systems are designed to extract high-level semantics associated with the cinematic elements and narrative forms synthesized using them[7] by emphasizing production knowledge. We can view this as a problem of computational media aesthetics.[7] In contrast, our work seeks to understand the advertising messages communicated to viewers.

ComBD is a significant stage for ComCL and commercial identification. Previous work suggested the use of blank or monochrome frames and quiet frames to segment each spot with a commercial break.[3] These methods assumed that two consecutive commercials are separated by a short break of several monochrome frames or audio depression occurrences. Derived from postediting effects, TV stations can easily omit these. Even if monochrome frames are used consistently, other editing effects (for example, fade-in and fade-out) and black/white frames occurring within an individual commercial would decrease the ability to discriminate between commercials. Determining audio segments of silence is technically sound. Yet silence is not a consistent and reliable indicator either. Although sound does help a TV viewer realize the transition from one commercial to the other, hidden Markov model-based approaches working on low-level audio features—even those coupled with visual features—cannot suffice for precisely partitioning different spots.[8] This is due to the complex audio content and the diverse temporal changes within a commercial itself. A desirable approach would seek intermediate multimodal features inherent in commercial video content. Those effects-related features should be dealt with as a complement but not as the central strategy.

References

1. M. Mizutani, S. Ebadollahi, and S.-F. Chang, ''Commercial detection in Heterogeneous Video Streams Using Fused Multimodal and Temporal Features,'' *Proc. IEEE Int'l Conf. Acoustic, Speech, and Signal Processing*, IEEE Press, 2005, pp. 157-160.
2. L. Agnihotri et al., ''Evolvable Visual Commercial Detector,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, vol. 2, IEEE Press, 2003, pp. 79-84.
3. R. Lienhart, C. Kuhmunch, and W. Effelsberg, ''On the Detection and Recognition of Television Commercials,'' *Proc. IEEE Int'l Conf. Multimedia Computing and Systems*, IEEE Press, 1997, pp. 509-516.
4. X.-S. Hua, L. Lu, and H.-J. Zhang, ''Robust Learning-Based TV Commercial Detection,'' *Proc. IEEE Int'l Conf. Multimedia and Expo*, IEEE Press, 2005, pp. 149-152.
5. J. M. Gauch and A. Shivadas, ''Identification of New Commercials Using Repeated Video Sequence Detection,'' *Proc. IEEE Int'l Conf. Image Processing*, vol. 3, IEEE Press, 2005, pp. 1252-1255.
6. C. Colombo, A. D. Bimbo, and P. Pala, ''Retrieval of Commercials by Semantic Content: The Semiotic Perspective,'' *Multimedia Tools and Applications*, vol. 13, Kluwer Academic, 2001, pp. 93-118.
7. C. Dorai and S. Venkatesh, ''Computational Media Aesthetics: Finding Meaning Beautiful,'' *IEEE Multimedia*, vol. 8, no. 4, 2001, pp. 10-12.
8. J. Huang, Z. Liu, and Y. Wang, ''Joint Scene Classification and Segmentation Based on a Hidden Markov Model,'' *IEEE Trans. Multimedia*, vol. 7, no. 3, 2005, pp. 538-550.

7. B.S. Manjunath and W.Y. Ma, ''Texture Features for Browsing and Retrieval of Image Cata,'' *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, 1996, pp. 837-842.
8. L.-.Y. Duan et al., ''Segmentation, Categorization, and Identification of Commercials from TV Streams Using Multimodal Analysis,'' *Proc. ACM Int'l Conf. Multimedia*, ACM Press, 2006, pp. 201-210.
9. V. Vapnik. *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.
10. T. Zhang and C.-C. Jay Kuo, ''Audio Content Analysis for Online Audiovisual Data Segmentation and Classification,'' *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 4, 2001, pp. 441-457.
11. C.-S. Xu et al., ''Live Sports Event Detection Based on Broadcast Video and Web-Casting Text,'' *Proc. ACM Int'l Conf. Multimedia*, ACM Press, 2006, pp. 221-230.
12. N. Babaguchi, K. Kawai, and T. Kitahashi, ''Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration,'' *IEEE Trans. Multimedia*, vol. 4, no. 1, 2002, pp. 68-75.
13. T.-S. Chua et al., ''TRECVID 2005 by NUS PRIS,'' *Proc. TREC Video Retrieval Evaluation* (Trecvid), Nat'l Inst. Standards and Technology, 2005; http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/nus.pdf.
14. A. Amir et al., ''IBM Research TRECVID-2005 Video Retrieval System,'' *Proc. Trecvid*, Nat'l Inst. Standards and Technology, 2005; http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/ibm.pdf.
15. F. Sebastiani, ''Machine Learning in Automated Text Categorization,'' *ACM Computing Surveys*, vol. 54, no. 1, 2002, pp. 1-47.
16. T. Joachims, ''Text Categorization with Support Vector Machines: Learning with Many Relevant Features,'' *Proc. Euro. Conf. Machine Learning*, Springer, 1998, pp. 137-142.
17. A. Hampapur, K. Hyun, and R. Bolle, ''Comparison of Sequence Matching Techniques for Video Copy Detection,'' *Proc. SPIE Storage and Retrieval for Media Database*, vol. 4676, 2002, pp. 194-201.
18. K. Kashino, T. Kurozumi, and H. Murase, ''A Quick Search Method for Audio and Video Signals Based on Histogram Pruning,'' *IEEE Trans. Multimedia*, vol. 5, no. 3, 2003, pp. 348-357.

Ling-Yu Duan is a research scientist at the Institute for Infocomm Research, Singapore. His research interests include video and motion computing, multimedia information retrieval, statistical learning, and pattern recognition. Duan has a PhD in information technology from the School of Design, Communication, and Information Technology at the University of Newcastle, Australia. Contact him at lingyu@i2r.a-star.edu.sg.

Jinqiao Wang is a PhD candidate at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include multimedia information retrieval, image content analysis and classification, machine learning, and computer vision. Wang has an ME from Tianjin University in mechatronics engineering. Contact him at jqwang@nlpr.ia.ac.cn.

Yan-Tao Zheng is a PhD candidate at the National University of Singapore. His research interests include multimedia information retrieval, image content analysis and classification, machine learning, and computer vision. Zheng has a BE in Computer Engineering from Nanyang Technological University, Singapore. Contact him at stuytz@i2r.a-star.edu.sg.

Hanqing Lu is a deputy director at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include image and video analysis, medical image processing, and object recognition. Lu has a PhD in computer science from Huazhong University of Sciences and Technology. Contact him at luhq@nlpr.ia.ac.cn.

Jesse S. Jin is the Chair Professor of IT at the School of Design, Communication and IT, University of Newcastle. His research interests include multimedia, medical imaging, computer vision, and pattern recognition. Jin has a PhD in computer science from the University of Otago, New Zealand. Contact him at Jesse.Jin@newcastle.edu.au.