LETTER

Commercial Shot Classification Based on Multiple Features Combination

Nan LIU[†a)], **Yao ZHAO**[†], **Zhenfeng ZHU**[†], *Nonmembers*, *and* **Rongrong NI**[†], *Member*

SUMMARY This paper presents a commercial shot classification scheme combining well-designed visual and textual features to automatically detect TV commercials. To identify the inherent difference between commercials and general programs, a special mid-level textual descriptor is proposed, aiming to capture the spatio-temporal properties of the video texts typical of commercials. In addition, we introduce an ensemble-learning based combination method, named Co-AdaBoost, to interactively exploit the intrinsic relations between the visual and textual features employed.

key words: *video categorization, video analysis, commercial detection, Co-AdaBoost, mid-level descriptor*

1. Introduction

TV commercials have become an inescapable part of modern day life and plenty of commercials are shown in broadcast videos which can be conveniently recorded by various multimedia devices for time-shifted easy access and consumption if necessary. It becomes clear that an automatic commercial detection scheme is increasingly in demand by both TV viewers and media professionals for their respective requirements, such as commercial filtering and commercial cataloguing/evaluation.

As the key starting point to an effective commercial detection system, commercial shot classification has commanded a lot of attention in recent years. Some previous works focused on the utilization of broadcast editing rules, such as the occurrence of black/silent frames [1] and the absence of subtitles [2] in advertisement time, to distinguish commercials from general programs. This approach, however, is heavily dependent on the specified rules and would fail in some countries where few of these rules are used. Aiming at alleviating this problem, others resorted to exploring various characteristics of commercials versus general programs from training samples and building a classifier to perform classification. For instance, a variety of context-based audio-visual features were proposed by Hua et al. [3] and Zhang et al. [4], taking account of temporal information. Moreover, Mizutani et al. [5] fused audio/visual/temporal local features of commercials in the context of their global temporal characteristics to detect commercial segments. Nevertheless, most of the prior studies have neglected the usage of textual information, which is one of the most important commercial characteristics. Although a simple case of application, employing local information of text occurrence, was found in [5], an in-depth research on textual descriptor was rarely explored.

The main goal of our research is to present a novel multiple features combination strategy (see Fig. 1.) to robustly discriminate commercial shots from those of general programs. The two main contributions are: Firstly, we propose a novel mid-level textual descriptor by exploiting the spatio-temporal properties of the video texts. Secondly, an ensemble-learning based combination method, called Co-AdaBoost, is introduced to improve the generalization ability between commercials and general programs in the multiple feature space.

2. Text Pattern Variation Indicator

Commercial is one of the most important media forms to convey the commodity, service provision or brand information to consumers. Aiming at generating sustained appeal of the 'products' promoted in the advertisements, a large number of text blocks, such as brand names and catch-phrases, are presented in the salient areas for a rather limited time to highlight their names or functions. But these texts are extremely unwonted in the majority of general programs, except some subtitles that appear in the bottom of the frames (as shown in Fig. 2). Even if there are a certain number of texts in news programs, their duration is much longer than that in commercials due to the fact that viewers need sufficient time to catch their meanings along with the contextual contents of the news. Hence, the occurrence frequency of text blocks can be reasonably taken to form an effective characteristic to discriminate commercials from general programs. In addition, the variation pattern of text blocks in commercials is usually more complex than that in general programs, in terms of the occurrence location, size and orientation (see Fig. 3).

To extract the occurrence frequency and variation pattern of the video texts in commercials, we propose a novel textual descriptor, named Text Pattern Variation Indicator (**TPVI**). As shown in Fig. 1, we employ a robust Co-Training based video text detection technique [9], previously developed by the authors for the purpose of complex commercial text detection, to locate binary areas with text presence in a key frame of each shot (from Fig. 4, it is found that the appearance of text blocks in commercials is more complicated than that in general programs). The key frame

Manuscript received January 29, 2010.
Manuscript revised May 25, 2010.

[†]The authors are with Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China.
a) E-mail: 05112073@bjtu.edu.cn

DOI: 10.1587/transinf.E93.D.2651

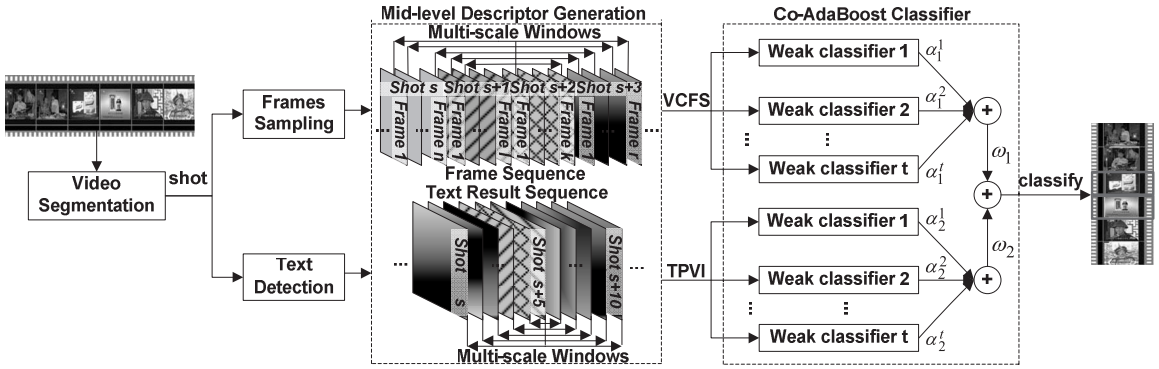
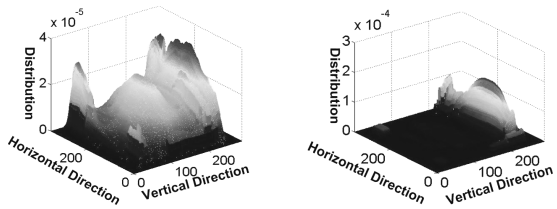


Fig. 1 Proposed multiple features combination scheme for commercial shot classification based on two kinds of visual and textual mid-level descriptors.



(a) Commercials (b) General programs

Fig. 2 Illustration of statistical distribution of text area position for commercials and general programs.

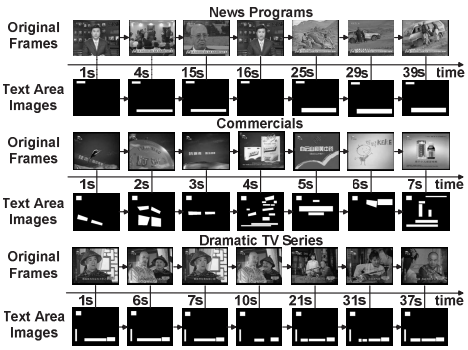


Fig. 3 Illustration of different variation patterns of text blocks in commercials, news programs and dramatic TV series.

Fig. 4 Examples of text area images of commercial key frames.

is simply chosen as the middle frame of each shot. Based on the result of statistical analysis of the collection of binary images as shown in Fig. 2 and Fig. 3, five significant features are introduced, employing N_t shot-level multi-scale sliding windows; the elaboration of each feature is given as follows:

(1) Text Block Frequency (TBF): we utilize the TBF, which consists of the temporal density (*td*) and the variance in unit time (*tv*) of the quantity of text blocks appeared in a sliding window, to reflect the text occurrence frequency characteristic. It is defined as:

$$td = \sum_{t=-W_t}^{W_t} Q(t) \Big/ \sum_{t=-W_t}^{W_t} Shot\_len(t)$$
$$tv = \sum_{t=-W_t}^{W_t} \left[Q(t) - \frac{1}{2W_t} \sum_{t=-W_t}^{W_t} Q(t) \right] \Big/ \sum_{t=-W_t}^{W_t} Shot\_len(t) \quad (1)$$

where W_t is half the size of the sliding window and $Q(t)$ is the total number of the text blocks occurred in each key frame within the sliding window.

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

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

where $I(x, y, t)$ is the binary text area image of each shot, which is the location distribution of output image of the text detector. Moreover, N, M, x_c and y_c are the size and center of $I(x, y, t)$, respectively. Then, the ratio $P(t)$ of the text areas in a key frame is used to substitute for $Q(t)$ in Eq. (1) to form the ratio-based representation, i.e. RTA.

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

(4) Text Orientation Histogram (TOH): The moment of inertia [6] is firstly utilized to calculate the orientation angle α of each text block. Then we employ a histogram with 3 bins, which represent horizontal ($0° \le \alpha < 10°$), vertical ($80° < \alpha \le 90°$) and slantwise ($10° \le \alpha \le 80°$) directions, respectively, to delineate the orientation distribution of the text blocks appeared in each key frame within a sliding window. Therefore, the TOH is characterized as the temporal density and the variance in unit time for each bin over the multiple histograms.

(5) Randomness of Text Occurrence (RTO): As clearly

shown in Fig. 3, the occurrence patterns of text blocks in commercials are revealed to be more random compared with those in general programs. Thus the randomness of text occurrence can be described as:

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

$$\text{where } \mathrm{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x \le 0 \end{cases} \text{ and } \sum_{u=1}^{W_t} e^{-\alpha|u|} = 1 \quad (3)$$

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

3. Visual Change on Frame Sequence

As we know, the spatial and temporal variations in visual contents of commercials are more drastic than those in general programs owing to the considerably limited duration of commercials. For the purpose of reinforcing the distinction of commercial characteristics, we present a mid-level descriptor Visual Change on Frame Sequence (**VCFS**) to delineate the local and global visual variations for each shot and its contexts.

To construct VCFS, a series of key frames are equally sampled from each shot with a 30-frame interval. Note that the key frame here is different from the middle frame used in TPVI. Aiming at the construction of salient intermediate descriptor, various global and local properties are exploited with N_v frame-level multi-scale sliding windows. The HSV histogram is adopted to form a global representation vector $V_g(t)$ for each key frame. With respect to the local information, each frame is partitioned into $h \times v$ blocks and then a set of local properties, including HSV histogram, edge change ratio and gray-scale frame difference, are extracted from each block to construct a vector $V_l(t)$. Next, the first- and second-order statistical moments of the variation on $V(t) = \{V_g(t), V_l(t)\}$ across multiple frames $V(t + u)$ contained in a sliding window with a certain scale are given as:

$$\begin{aligned} C_1^v(t) &= \frac{1}{2W_v} \sum_{u=-W_v}^{W_v} V(t) - V(t+u) \\ C_2^v(t) &= \frac{1}{2W_v} \sum_{u=-W_v}^{W_v} \left[V(t) - V(t+u) - C_1^v(t) \right]^2 \end{aligned} \quad (4)$$

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

$$C^v = \left[\frac{1}{R} \sum_{t=1}^{R} C_1^v(t), \quad \frac{1}{R} \sum_{t=1}^{R} C_2^v(t) \right] \quad (5)$$

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

4. Co-AdaBoost

The main idea of Co-AdaBoost is to utilize an ensemble of multiple weak learners, which are sequentially selected from two different descriptors based on a set of updated weights over the training set, to build a stronger classifier. The pseudo-code is shown in Fig. 5. Just like the traditional AdaBoost [7], all weights are initially set to be uniform, but on each round the weights of the incorrectly (correctly) classified examples are increased (decreased). But the key difference lies in that the penalty degree of the weight is simultaneously controlled by h_j^t ($j = 1, 2$), inspired in part by the interactive cross-modal learning fashion [8]. Therefore, we can select more informative samples, on which the agreement cannot be reached for h_j^t ($j = 1, 2$), from the training set and place a bigger weight on them so that the multiple weak learners can be forced to focus on these examples in the next round to achieve better generalization ability. Specifically, the weak learner, adopted in Co-AdaBoost, classifies all examples less than a threshold as belonging to one class and greater than a threshold as another class. Moreover, the *WeakLearn* (see Fig. 5) denotes the selection strategy for the optimal threshold which can minimize the error rate ε^t based on D_j^{t-1}.

Input : a set of training examples $S_j = \{(x_{ji}, y_i)\}$ where $j = 1,2$ $i = 1,2,\cdots,m$, $y_i \in \{-1,1\}$ and the maximum iteration round T.

1. Initialize the weights of training samples for each descriptor: $D_{ji}^0 = 1/m$.

2. Do for each round of iteration $t = 1,2,\cdots,T$:

2.1. **Do** for each descriptor $j = 1,2$:

-Utilize the *WeakLearn* to train the weak learner h_j^t

- Calculate the error rate of h_j^t: $\varepsilon_j^t = \frac{1}{2} \sum_{i=1}^{m} D_{ji}^{t-1} \mid h_j^t(x_{ji}) - y_i \mid$

- Set the coefficient of h_j^t: $\alpha_j^t = \frac{1}{2} \ln[(1 - \varepsilon_j^t)/\varepsilon_j^t]$.

2.2. **Do** for each descriptor $j = 1,2$:

- Update the weights according to the following rules:

if $h_1^t(x_{1i}) = h_2^t(x_{2i})$, $D_{ji}^t = D_{ji}^{t-1} \min_j \{\exp[-\alpha_j^t y_i h_j^t(x_{ji})]\}$,

if $h_1^t(x_{1i}) \neq h_2^t(x_{2i})$, $D_{ji}^t = D_{ji}^{t-1} \exp[-\alpha_j^t y_i h_j^t(x_{ji})]$.

- Normalize the weights: $D_{ji}^t = D_{ji}^t / Z_j^t$ where Z_j^t is a normalization factor.

3. Construct the ensemble of h_j^t:

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

Output : $H = \sum_{j=1}^{2} \omega_j H_j^T$.

Fig. 5 Co-AdaBoost training process for multiple features combination.

5. Experimental Results

A series of experiments are conducted based on a collection of 13.8 hours videos comprising the data derived from TRECVID05 and videos captured from several Chinese TV Channels. We select 8.6 hours long videos containing 8723 shots for training and the remaining 5.2 hours (4731 shots) for testing. And the general programs include news and dramatic TV series. The evaluation measures including precision, recall and accuracy [3], [4] are utilized to evaluate the classification performance.

5.1 Performance of Mid-Level Descriptor

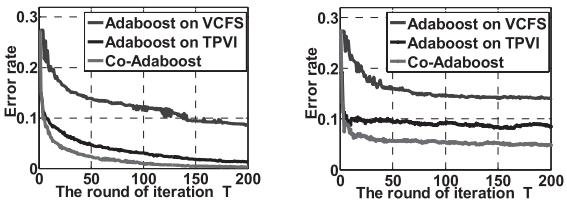
To investigate the effect of automatic shot boundary and text detection results on the performance of the proposed descriptors, LIBSVM with the kernel of Radial Basis Function (RBF) is employed as the benchmark classifier in our experiments and the parameters are obtained via cross-validation. In Table 1, *Descriptor+AT_shot+AT_text* denotes the descriptor with automatic shot boundary and text detection, whereas *Descriptor+ML_shot+ML_text* denotes the descriptor with manually labeled shot boundary and text area. As clearly shown in Table 1, both visual and textual descriptors based on automatic detection results achieve promising performance. Particularly, for the *TPVI+AT_shot+AT_text*, 7 % improvement is achieved with comparison to the *VCFS+AT_shot*, demonstrating the effectiveness of exploiting the essential characteristics of the text blocks appeared in commercials. Meanwhile, the robustness of the proposed descriptor is also satisfactory with the decrease of 4.79 % and 7.26 %, respectively, against the descriptors based on manual labeled results.

5.2 Combination of Two Mid-Level Descriptors

The classification error rates of *Co-AdaBoost* in each round are described in Fig. 6 with the training and testing data. Specifically, the shot boundary and text area image for VCFS and TPVI are extracted automatically. As shown in Fig. 6, the variation of error rate becomes stable at round $T = 200$. Thus considering the tradeoff between the classification accuracy and computational complexity, we set the iteration round $T = 200$ in our experiments. Moreover, it is clear that the result of *Co-AdaBoost* strategy convincingly outperforms independent *AdaBoost* on each individual descriptor. To evaluate the combination effect of our proposed strategy, we also combine two individual *SVM* classifiers by summing the weighted prediction probabilities in comparison with *Co-AdaBoost*. As we can see from Table 2, the performance of *Co-AdaBoost* is more promising than the SVM combination method with a nearly 2 % improvement. Note that the weights in *combination of SVM* are 0.45 (VCFS) and 0.55 (TPVI), and in *Co-AdaBoost* are 0.35 (VCFS) and 0.65 (TPVI), respectively, which achieve the best performance in our experiment.

Table 1 Performance comparison for VCFS and TPVI with automatic and manually labeled shot boundary and text detection results.

Mid-level descriptor	Precision	Recall	Accuracy
VCFS + ML_shot	82.46%	87.25%	89.97%
VCFS + AT_shot	74.21%	75.13%	85.18%
TPVI + ML_shot + ML _text	98.86%	99.33%	99.42%
TPVI + AT_shot + AT_text	94.04%	77.96%	92.16%



(a) Error rate of training samples (b) Error rate of testing samples

Fig. 6 Error rate comparison of *Co-AdaBoost* and the independent *AdaBoost* on individual descriptor.

Table 2 Performance comparison for different combination strategies.

Combination Strategy	Precision	Recall	Accuracy
Combination of SVM	93.03%	85.38%	93.32%
Co-AdaBoost	**93.05%**	**90.62%**	**95.31%**

6. Conclusions

In this paper we have presented an effective commercial shot classification scheme for the purpose of automatic detection of TV commercials. In addition to using traditional visual features, a novel textual mid-level descriptor, TVPI, was proposed, capitalizing on the spatio-temporal properties of the video texts. Moreover, we introduced an ensemble-learning based combination method, named Co-AdaBoost, to improve the generalization ability by interactively exploring the intrinsic relations across multiple features. The experiments showed the effectiveness of the proposed scheme.

Acknowledgments

This work was supported in part by NSFC (No. 60776794), PCSIRT (No. IRT0707), Sino-Singapore JRP (No. 2010DFA11010), Fundamental Research Funds for the Central Universities (No. 2009JBZ006-3) and Open Foundation of NLPR.

References

[1] D.A. Sadlier, S. Marlow, N. O'Connor, and N. Murphy, "Automatic TV advertisement detection from MPEG bitstream," Pattern Recognit., vol.35, no.12, pp.2719–2726, Dec. 2002.
[2] Y.P. Huang, L.W. Hsu, and F.E. Sandnes, "An intelligent subtitle detection modal for locating television commercials," IEEE Trans. Syst. Man Cybern. B, Cybern., vol.37, no.12, pp.485–492, April 2007.
[3] X.S. Hua, L. Lu, and H.J. Zhang, "Robust learning-based TV commercial detection," Proc. IEEE ICME2005, pp.149–152, Amsterdam, Netherlands, July 2005.

[4] L. Zhang, Z.F. Zhu, and Y. Zhao, "Robust commercial detection system," Proc. IEEE ICME2007, pp.587–590, Beijing, China, July 2007.
[5] M. Mizutani, S. Ebadollahi, and S.F. Chang, "Commercial detection in heterogeneous video streams using fused multi-modal and temporal features," Proc. IEEE ICASSP2005, vol.2, pp.157–160, Philadelphia, USA, March 2005.
[6] H.P. Li and D. Doermann, "A video text detection system based on automated training," Proc. 15th IEEE ICPR, vol.2, pp.223–226, Barcelona, Spain, May 2000.
[7] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," Proc. 11th annual conf. on Computational learning theory, pp.92–100, Madison, USA, 1998.
[8] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Computer and System Sciences, vol.55, no.1, pp.119–139, Aug. 1997.
[9] N. Liu, Y. Zhao, and Z.F. Zhu, "Commercial video text detection scheme based on co-training strategy," Technical Report, BJTU, March 2008.