TV COMMERCIAL CLASSIFICATION BY USING MULTI-MODAL TEXTUAL INFORMATION

Yantao Zheng[1]*, Lingyu Duan*[2]*, Qi Tian*[2]*, Jesse S. Jin*[3]
yantaozheng@comp.nus.edu.sg, {lingyu, tian}@i2r.a-star.edu.sg, Jesse.Jin@newcastle.edu.au

[1]NUS Graduate School, National University of Singapore, Singapore 117597
[2]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[3]The School of Design, Communication and Information Technology, University of Newcastle, NSW 2308, Australia

ABSTRACT

In this paper, we propose an approach for TV commercial video classification by the categories of advertised products or services (e.g. automobiles, healthcare products, etc). Since automatic speech recognition (ASR) and optical character recognition (OCR) can deliver meaningful textual information related to products or services, TV commercial video classification is formulated as the problem of text categorization. However, there exist two challenges. Firstly, the background music of TV commercials makes ASR techniques yield erroneous and deficient output transcripts. Secondly, even if ASR and OCR could work perfectly, the limited textual information from TV commercials do not suffice to train a generic and non-overfitting text categorizer. For the first issue, our approach resorts to the external resources to expand deficient ASR and OCR transcripts. The output transcripts of ASR and OCR are parsed to yield a few keywords, on which a Web searching is executed to retrieve relevant and semantically informative articles from World Wide Web (WWW). The retrieved articles are then utilized to construct textual feature vectors and perform text categorization on behalf of commercials. For the second issue, a topic-wise document corpus is constructed from the public corpora like Reuters-21578 or from the articles manually collected from WWW for the training of text categorizers. Experimental results have shown that the proposed approach alleviates the negative effects from weak ASR/OCR performance and yield a promising classification accuracy of 80.9%.

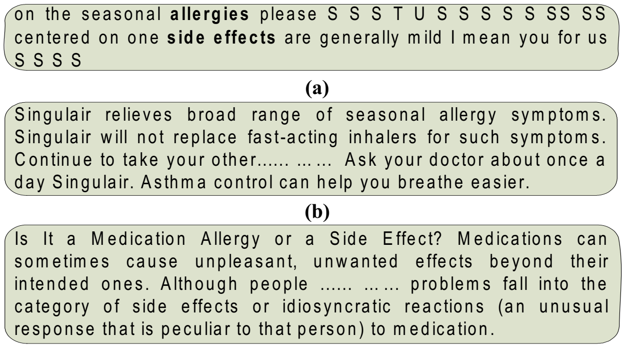
1. INTRODUCTION

Automatic TV commercial video classification by advertised products or services undoubtedly contributes to the management and monitoring of TV commercials. It may facilitate TV commercial industry and commercial design education by providing a well-indexed commercial database. Few research works have explored the semantic classification of TV commercial videos. Previous research works [1] [2] [3] [4] are basically focused on commercial skipping type applications by locating and removing commercial segments. Various low-level audiovisual features have been proposed to distinguish commercial segments from program segments. In this paper, our focus is switched to the semantic video analysis within an individual TV commercial.

Speech and key frame images are two important resources for structural and semantic video content analysis. The textual information can be extracted from these resources by applying techniques such as automatic speech recognition (ASR) and optical character recognition (OCR). In [5], the ASR-based transcripts were utilized to perform news video question-answer analysis. In [6], OCR was applied to detect text in key frames for news stories categorization.

Moreover, contrast to news and sports videos, TV commercials do not have any available closed-captioned text or web broadcast text. Therefore, we are motivated to resolve the problem of commercial video classification by analyzing the textual outputs of ASR and OCR. Speech content consists of affluent semantic information about advertised products or services. OCR may provide valuable hints such as brand name, slogan, etc. Accordingly, the commercial video classification problem is reduced to the text categorization on the basis of the transcripts generated by ASR and OCR.

However, the ASR accuracy is closely related to the signal-to-noise ratio (SNR) of audio. Unlike news videos, TV commercial videos often involve complex background music to attract TV audiences' attention and dispose them favorably towards the products or services. Unfortunately, the ASR performance would be corrupted by the "noise" introduced by music. Fig. 1(a) shows an example of ASR transcript of the TV commercial video of *Singulair*, which is a medicine curing asthma and allergy. Fig. 1(b) presents a part of the manually recorded speech transcript of this commercial. The transcript comparison between Fig. 1(a) and Fig. 1(b) indicates that the background music impedes ASR techniques from delivering a semantically meaningful and coherent message describing the advertised commodity. Fortunately, the output of ASR and OCR usually contains words more or less related to the advertised commodity's category, like <allergy> and <side effect> highlighted in Fig. 1(a) and circled words in Fig. 2. Fig. 2 shows four key frame images of four commercials, which contain either the text related to advertised commodity's category like <Credit Card> for the finance category, or even the commodity's brand name like <Microsoft> for IT.



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

(a)

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.

(b)

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.

(c)

Fig. 1. *Singulair* TV commercial (a) ASR transcripts. (b) manually recorded speech transcripts. (c) article searched from Web

1-4244-0367-7/06/$20.00 ©2006 IEEE

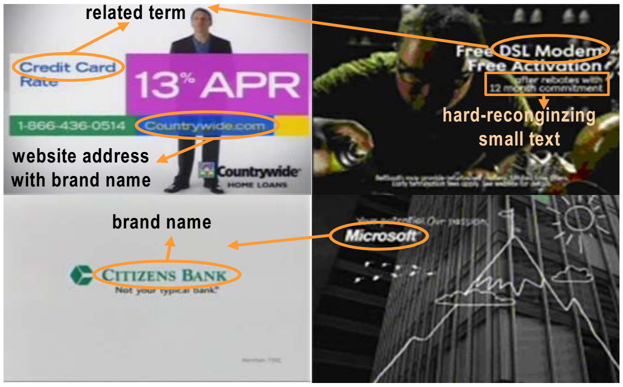


Fig. 2. Frame images with rich semantic information

Since our ultimate purpose is to classify a TV commercial video into its advertised commodity's category, e.g. *Singular* commercial to healthcare, it is preferred but not necessary to exploit the actual speech transcript in the 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. 1(c) is obtained by Google search with keywords <allergy> and <side effect>. Obviously, this article can be classified to the healthcare category. Therefore, we propose an approach to make use of informative articles from external resources to circumvent the weak speech recognition performance in commercial videos and to finally improve the commercial classification result.

2. PROBLEM FORMULATION

The proposed approach firstly preprocesses the output transcripts of ASR and OCR on TV commercial video *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 a natural language processor [7]. A set of keywords $K_i\,(kw_{i1},\ldots,\ kw_{ix})$ are selected by applying the steps below:

1. Check S_i against a predefined dictionary of brand names.
2. 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).
3. Otherwise, from L_i, the *n* nouns and noun phrases with largest font size from OCR and the last *m* from ASR are heuristically selected as keywords and searched via a web search engine.

Google search engine is used in our implementation, as its superior performance assures the searched article's relevancy. Among returned articles, the one with highest relevancy rating is selected as d_i, which we denote as the proxy article of TV commercial *i*. By exploiting d_i, TV commercial classification is reduced to the problem of text categorization [8]. That is to approximate a classifier function $\phi : D \times C \rightarrow \{T, F\}$ to assign a Boolean value to each pair $(d_i, c_j) \in D \times C$, where D is the domain of proxy article d_i and C is the set of predefined commercial category c_j. A value T assigned to (d_i, c_j) indicates the proxy article d_i under c_j, while a value F assigned to (d_i, c_j) means d_i not under c_j.

3. TV COMMERCIAL CLASSIFICATION FRAMEWORK

Overall the proposed approach consists of four modules: Information Retrieval (IR) Text Preprocessing Module, Commercial Video Module, Training Data & Word Feature Module and Classifier Module, as shown in Fig. 3.

3.1 IR Text Preprocessing Module

This module functions as a vocabulary term normalization process that is widely used in the setting-up of IR systems. It applies two major steps: the Porter stemmer algorithm and the stop word removal algorithm [9]. The Porter stemming algorithm is a process of removing the common morphological and inflexional endings from words in English. Stop word removal is to eliminate words of little or no semantic significance, such as "the", "you", "can" etc. As shown in Fig. 3, both testing and training documents go through this module before any other process runs on them.

3.2 Commercial Video Module

This module aims to expand the deficient and less-informative transcripts from ASR and OCR with relevant and informative articles searched from WWW like Google and encyclopedia webs.

For each incoming TV commercial video *i*, the module firstly extracts the raw semantic information via ASR and OCR on the key frame images. Key frames are uniformly extracted every four seconds. 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 ones. As shown in the right upper image in Fig. 2, it is easy for OCR to recognize the text of large size "Free DSL Modem, Free Activation", which contains more category related semantic information than the small and hard-recognizing text "after rebates with 12 months commitment". Therefore, OCR's failure in recognizing small texts does not necessarily degrade the final performance much. It is also 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 of ASR and OCR. 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 normal English dictionary may not include non-vocabulary terms like brand names. Based on corrected transcript S_i, the proxy article d_i is obtained from the steps stated in Section 2. With word features, the testing document vector will be generated from d_i.

3.3 Training Data & Word Feature Module

This module is to generate the training dataset and feature space for text categorization. Firstly, a topic-wise document corpus is constructed from pubic IR corpora or relevant articles manually collected from WWW as the training dataset of text categorizer. In this way, the training corpus can possess large amount of training documents and wide coverage of topics. Such training corpus can avoid the potential over-fitting problem, which may be caused if the textual information of a limited set of TV commercials is taken as training data. In our experiments, the categorized Reuters-21578 [10] and 20 Newsgroup [11] corpora are combined to construct the training dataset. The defined categories of these corpora may not exactly match ones of TV commercials. Our solution

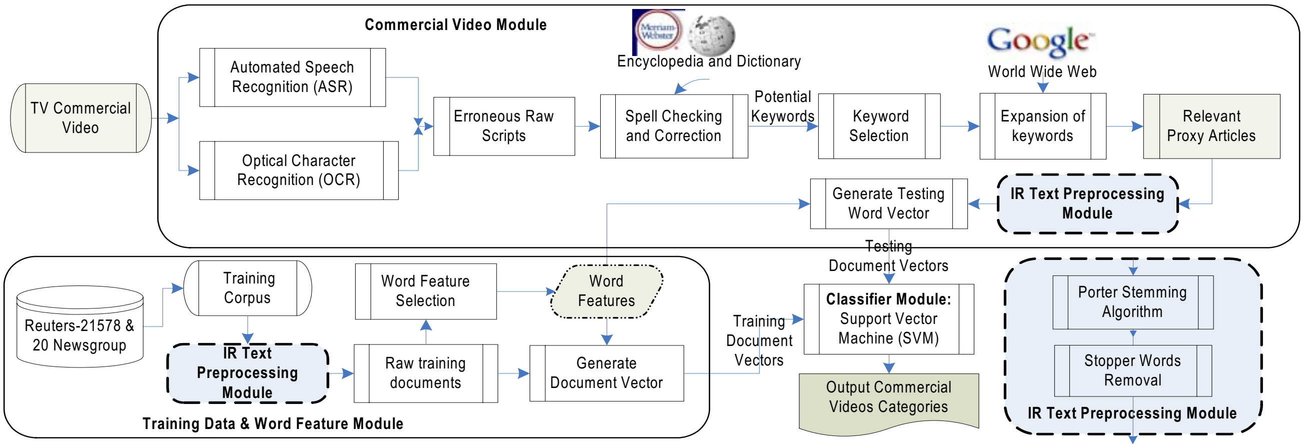

Commercial Video Module
TV Commercial Video
Automated Speech Recognition (ASR)
Optical Character Recognition (OCR)
Erroneous Raw Scripts
Encyclopedia and Dictionary
Spell Checking and Correction
Potential Keywords
Keyword Selection
Google
World Wide Web
Expansion of keywords
Relevant Proxy Articles
Generate Testing Word Vector
IR Text Preprocessing Module
Testing Document Vectors
Reuters-21578 & 20 Newsgroup
Training Corpus
IR Text Preprocessing Module
Word Feature Selection
Word Features
Raw training documents
Generate Document Vector
Training Data & Word Feature Module
Training Document Vectors
Classifier Module: Support Vector Machine (SVM)
Output Commercial Videos Categories
Porter Stemming Algorithm
Stopper Words Removal
IR Text Preprocessing Module

Fig. 3. System architecture of TV commercial classification framework

is to select categories of these corpora that are related to commercials' one and combine them to jointly construct the training dataset for that commercial category. For example, the documents from the categories of "earn", "money" and "trade" in Reuters corpus are merged together to construct the training dataset for the finance category.

Next the document frequency technique [12] is applied on training dataset to perform word feature selection. Document frequency is one of the simplest feature selection and high feature dimensionality reduction methods. Its promising performance and approximately linear computational complexity reported in [12] promote its usage in our approach. The document frequency fw_j measures the number of documents, in which a word w_j occurs. If fw_j exceeds a predetermined threshold, w_j will be selected as a feature. The document vectors are accordingly generated in the constructed feature space. For each document, the number of occurrences of word w_j is taken as the value of feature w_j. Finally, each document vector is normalized to unit length, so as to eliminate the influence of different document lengths.

3.4. Classifier Module

The Classifier Module aims to perform text categorization of proxy articles, and furthermore, determine the categories of respective TV commercials. In [8] and [13], various text categorization techniques have been reviewed and SVM is reported to deliver consistently outstanding performance. Thereby, SVM is used as classifier in our implementation. [14] presented the following promising characteristics of SVM to theoretically demonstrate its suitability for text categorization task.

- Capability to handle high dimensional input space. Text classification usually involves a feature space with extremely high (around 10,000) dimensions.
- Capability to tackle sparse document vectors. Document vectors usually contain only a few non-zero entries, due to the short length of documents and large feature space

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents both our findings from the observation of TV commercial data and experimental results. The proposed approach is evaluated by comparing its classification results with ones of manually recorded actual speech transcripts and ASR transcripts.

4.1. TV Commercial Data Observations and Parameter Setting

We cut 499 English TV commercials and extracted 191 distinct ones from TRECVID05 [15] database. Based on their advertised products or services, the 191 distinct TV commercials are distributed in eight categories, as illustrated in Fig. 4. Our experiments involve four categories: Automobile, Finance, Healthcare and IT. Though they do not exclusively cover all TV commercials, yet they count up to 141 and 74% of total commercials. Therefore, they should be able to demonstrate the effectiveness of the proposed approach. For each category, 1,000 training documents are selected from corpora Reuter and 20 Newsgroup. Altogether the training documents amount to 4,000. In word feature selection phase, the document frequency threshold is set to 2, and 9107 word features are selected. Prior to training SVM, these 4,000 documents were evaluated by a three-fold cross validation to examine their integrity and qualification as training data. The cross validation accuracy reached up to 96.9%, where Radial basis function (RBF) kernel was used and SVM parameter *cost* and *gamma* were determined to be 8,000 and 0.0005.

In keyword selection phase, the statistics show that in 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 parameter n and m to be 2. The recognition of brand names from ASR and OCR plays an important role, as brand names are the best keyword candidates. Fig. 5 presents the number of commercials, in which OCR and ASR recognized their brand names successfully. It shows OCR can recognize brand names in a considerable amount of commercials, especially in automobile ones. Overall, OCR can recognize brand names of 56% of all commercials. OCR recognizes brand names from two major

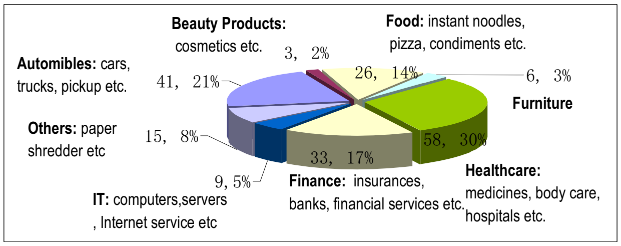


Fig. 4. TV commercial category distribution

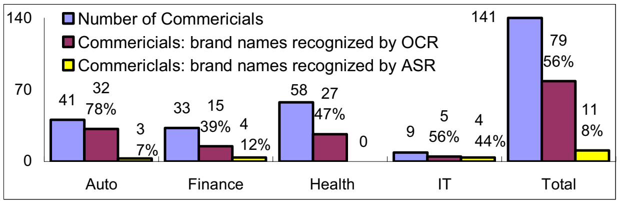


Fig. 5 Commercials with brand names recognized by ASR & OCR

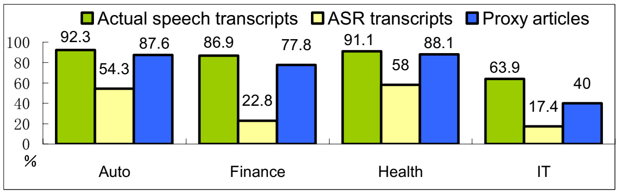


Fig. 6. F1 values of three classifications

Table 1. Results of classifications with manually recorded speech transcripts, ASR transcripts and proxy articles

.	**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

sources: the trade name text and the website address on frame image, as shown in upper left image of Fig. 2.

4.2. Experimental Result Evaluations and Discussion

The classification based on manually recorded speech transcripts of commercials is firstly performed. As Table 1(a) shows, except IT, all other categories achieve satisfactory classification result and the overall classification accuracy reaches 85.8%. The reason of low accuracy in IT category lies in the mismatch of category definition between the training data and testing commercials. In the training data, IT category mainly covers computer hardware and software. However, in testing commercials, it includes other IT products, like printers and photocopy machines. ASR transcripts are also applied to perform text categorization. As Table 1(b) shows, the ASR transcripts deliver bad results in all categories. Table 1(c) shows the classification results with proxy articles. Compared with ASR transcripts, the classification results have been improved drastically and the overall classification accuracy increases from 43.3% to 80.9%. Fig. 6 displays the F1 values of classifications based on all three types of inputs. For most categories, the proxy articles deliver slightly lower accuracies than the 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. CONCLUSION

In this paper, we propose a novel approach to classify TV commercial videos by advertised commodities' categories. The proposed approach aims to improve the classification accuracy by utilizing the external resources (World Wide Web) to circumvent speech recognition problem in commercial videos. The experiments conducted over 141 distinct TV commercials deliver 80.9% overall classification accuracy. Hence, it can be concluded that the proposed approach is able to perform satisfactory TV commercial classification.

There exist several open issues with the proposed system. Firstly, the current experiments involve four categories of TV commercials only. The training dataset should be extended to include more categories. Secondly, the keyword selection rules are heuristic. It could fail, if the selected keywords from ASR and OCR transcripts were not able to identify the commercial's category. One improving solution can be incorporating a keyword significance checking to filter out the "trivial" keywords.

6. REFERENCES

[1] L. Agnihotri, etc. "Evolvable visual commercial detector," *Proc. CVPR'2003*, Wisconsin, USA, Jun. 2003.

[2] M. Mizutani, S. Ebadollahi and S.F. Chang, "Commercial detection in heterogeneous video streams using fused multi-modal and temporal features," *ICASSP 2005*, Philadelphia, USA, 2005

[3] R. Lienhart, C. Kuhmunch, and W. Effelsberg, "On the detection and recognition of television commercials," *ICMCS'1997*, Ottawa, pp. 509-516.

[4] A. G. Hauptmann and M. J. Witbrock, "Story segmentation and detection of commercials in broadcast news video," in *Proc. Conf. Advances in Digital Libraries*, Santa Barbara, 199

[5] H. Yang, L. Chaisorn, Y. Zhao, S.Y. Neo, and T.S Chua, "VideoQA: question answering on news video", *ACM Multimedia 2003*, Berkeley, CA, USA, pp. 632-641, 2003

[6] W. Qi, L. Gu, H. Jiang, X.R. Chen, H. Zhang, "Integrating Visual, Audio and Text Analysis for News Video", *ICIP 2000*, Vancouver, BC, Canada, 10-13 Sept. 2000

[7] C. Toklu, S.P. Liou, and M. Das, "VideoAbstract: A New Hybrid Approach to Video Summary Generation", *ICME 2000.* New York, USA, vol.3, pp.1333-6, 2000

[8] F. Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, pp 34. 2002

[9] S. Jones, Karen, and P. Willet, *Readings in information retrieval*, San Francisco: Morgan Kaufmann, 1997

[10] Reuters. Reuters-21578 text categorization test collection, 1997.http://www.daviddlewis.com/resources/testcollections/reuters21578

[11] K. Lang, "Newsweeder: Learning to filter netnews", *ICML 95*, Tarragona (Catalonia, Spain), pp. 331–339, 1995.

[12] Y. Yang, and J.O. Pedersen, "A comparative study on feature selection in text categorization", *ICML 97*, pp. 412-420, 1997

[13] Y. Yang and X. Liu, "A re-examination of text categorization methods", *Proceedings of SIGIR-99*, Berkeley, CA, USA, 1999 [14] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", *ECML'98*, Chemnitz, Germany, 1998

[15] TRECVID. TREC Video Retrieval Evaluation 05, National Institute of Standards and Technology, U.S.A. 2005