LINKING VIDEO ADS WITH PRODUCT OR SERVICE INFORMATION BY WEB SEARCH

Jinqiao Wang[1], Ling-yu Duan[2], Bo Wang[1], Shi Chen[1], Yi Ouyang[1], Jing Liu[1], Hanqing Lu[1], Wen Gao[2]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{jqwang, bwang, schen, youyang, jliu, luhq@nlpr.ia.ac.cn }
[2]Institute of Digital Media, School of EE & CS, Peking University, Beijing 100871, China
{lingyu, wgao@pku.edu.cn }

ABSTRACT

With the proliferation of online media services, video ads are pervasive across various platforms involving internet services and interactive TV services. Existing research efforts such as Google AdSense and MSRA VideoSense/ImageSense have been devoted to the less intrusive insertion of relevant textual or video ads in streams or web pages through text/image/video content analysis whereas the inherent semantics of video ads is much less exploited. In this paper, we propose to link video ads with relevant product/service information across E-commerce websites or portals towards ad recommendation in a cross-media manner. Firstly, we carry out semantic analysis within ad videos in which *Frames Marked with Product Images* (FMPI) are extracted. Secondly, we link ad videos with relevant ads on the Web by utilizing FMPI to search visually similar *Product Images* (e.g. appearance or logo) and to collect their accompanying text (brand name, category, description, or other tags) over popular E-commerce websites or portals such as EBay, Amazon, Taobao, etc. We search visually similar product images with Local Sensitive Hashing (LSH) in a Naïve Bayes Near Neighbor classifier. Finally, we may recommend more relevant products/services for ad videos through ranking those matched product images and categorizing useful tags of top ranked ads from the Web. Preliminary experiments have been carried out to demonstrate the idea of linking ad videos with product/service information from the Web.

Index Terms— video ad, video retrieval, ad recommendation

1. INTRODUCTION

Undoubtedly video ads have become a popular advertising form to promote goods, services, and ideas via online media services. TV ad is generally considered as the most effective mass-market advertising format that could be reflected by high prices charged for airtime during popular events. Recently, with the proliferation of user generated content (UGC) websites, Internet is dramatically changing the way of entertainment. More and more people prefer to watch programs and network videos over PC instead of a TV set, where viewer relevant or non-relevant video ads are often inserted. Ideally ads function as a sort of information medium, which could provide valuable information about products or services that they do not know but might want. In addition to entertainment, Internet is changing the way of shopping. Nowadays, online shopping is becoming popular and people can enjoy efficient shopping at their comfort home. Through linking video ads with more or less related products promoted on websites, online ad recommendation in a cross-media manner is becoming increasingly critical to assist the decision-making of online consumers.

Most relevant works in ad videos focus on detection [4][2], retrieval [1][3], insertion [5][8] and categorization [6][7]. Ad detection aims to locate and skip ads in video streams for effective browsing of TV programs (e.g. in DVR). Ad retrieval is to identify and locate a particular ad from video streams or databases. For ad insertion, Mei et al. [5] tried to associate relevant video ads with the content of playing videos and to seamlessly insert the ads at proper points in video streams. Liu et al. [9] proposed to insert ads at the low attention regions in videos. Ad categorization is meant to classify video ads into one of predefined classes. Colombo et al. [6] proposed an approach to semiotic analysis of ad videos. They utilized heuristic rules to associate a set of perceptual features with four major commercial production types, i.e., practical, playful, utopic and critical. Duan et al. [2] proposed a multimodal approach to classify ad videos by advertised products or services.

Enormous money is spent on video ad production to capture the attention of audience. Many ads are produced so elaborately that a video ad can be considered as a miniature movie (say 30 seconds). Such creative arts design makes ad video analysis fairly challenging, which involves ad detection, retrieval, insertion and categorization. For example, in [2], we have revealed that deficient texts extracted by ASR/OCR in ad videos are weak in extracting useful semantics for effective ad recommendation.

In this paper, we propose a scheme to link ad videos with pervasive product/service information across Internet towards ad recommendation in a cross-media manner. Firstly, we analyze ad videos to capture the subset of informative images about advertised products/services, i.e. FMPI images [2]. Then we try to collect relevant image and textual information from external resources (i.e., the Web) through matching the crawled product images embedded in E-commerce websites or portals against recovered FMPI images from video ads. We parse the tags of top matched images and make more meaningful recommendation of product/service information. To accomplish effective and efficient matching, we extract SURF (speeded up robust features) points [10] from FMPI images, and employ a Naïve Bayes Near Neighbor classifier to retrieve visually similar product images extensively crawled from websites (in advance). To accelerate the matching process, Local Sensitive Hashing (LSH) [12] is employed. Finally we may make online recommendation in other ads formats by linking video ads with rich product information from the Web.

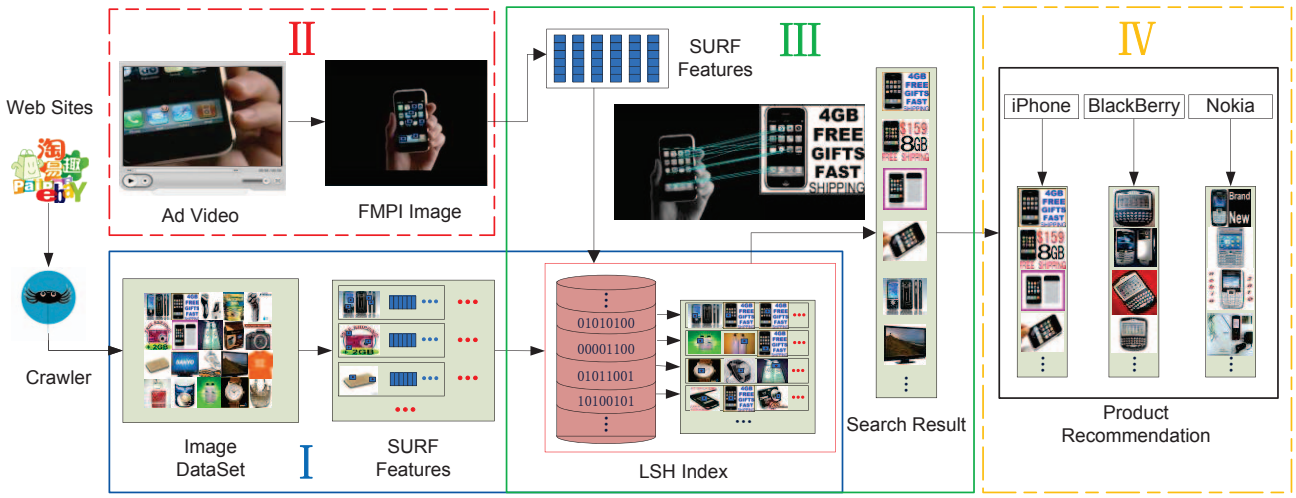


Fig.1. Overall framework of linking video ads with product/service information from the Web

2. FRAMEWORK

Fig.1 illustrates the overall framework of linking video ads with popular E-commerce websites/portals, which includes four major modules: 1) indexing product images crawled from the Web, 2) ad video analysis, 3) searching visually similar product images, and 4) product/service recommendation enriched by the Web.

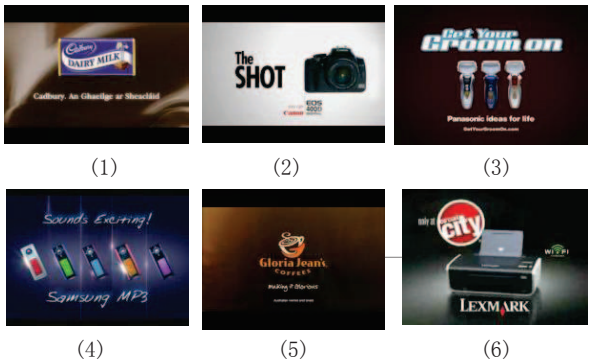
To make use of external resources from the Web, we build up a large dataset of product images (e.g., appearance, brand logos) by collecting images extensively from popular shopping websites, such as eBay, Amazon, etc. When crawling web pages, in addition to capture product images, we store useful tags including category, brand name, price, and shop place, etc. On the image dataset, we extract SURF features for all stored images and create an index of SURF features by using LSH. Ad video analysis is to extract FMPI images from each ad video. Searching visually similar product images works on selecting the stable SURF features of FMPI images to search goods images via LSH-based matching so as to establish the linking between ad videos and related products.

On the basis of web product search, we reorganize and rank the raw search results of product images and recommend more relevant products/services with the surrounding tag information, as indicated in Fig.1. Further classification can be done by our multimodal approach as reported in [2].

3. VIDEO ADS ANALYSIS

On a large dataset, frame-by-frame matching is infeasible for searching product images from each video ad. How to succinctly represent video ads is critical for efficient retrieval. Thus we resort to commercial production rules; that is, we detect the presence of FMPI images to identify a video ad [2]. As illustrated in Fig.2, an FMPI image can be dealt as a kind of document image involving graphics (e.g., corporate symbols, logos), images (e.g., products, setting and props) and texts (e.g., brand names, headlines or captions and contact information). FMPI images are used to highlight the advertised products, service or ideas.

FMPI images exhibit a uniform and clear pattern. Moreover, FMPI images are distinct among different ads. An important fact is that if the ad videos from different companies involve the same or quite similar FMPI images, copyright issues would be raised. We utilize FMPI images to identify and match ad videos. In practice, our FMPI recognizer works on key frames only.



(1) (2) (3)

(4) (5) (6)

Fig.2. Examples of FMPI images

Our FMPI recognizer is trained by SVM. An FMPI image is represented by the properties of color, texture, and edge features. As described in [2], a 141-d visual feature vector comprising 128-d local features and 13-d global features is constructed. C-Support Vector Classification (C-SVC) is employed, and the radial basis function (RBF) kernel is used. We use the probability output of SVM classifier to determine FMPI images and the key frame with the highest probability within an ad video is selected as FMPI.

It is worthy to note that an FMPI image may produce diverse views of a product image, which may deviate from the (frontal) images crawled from Web. The availability of only one view in an FMPI image could be limited for retrieving product images. Hence, we warp an FMPI image with an affine model (six parameters) to generate more views with different scales and rotation angles for robust features against noise, illumination and viewpoints. In our experiments, we set scale range $s \in [0.1, 2]$, rotation angle θ and $\phi \in [-\pi/4, \pi/4]$.

As local image features, SURF has been successfully applied to object recognition and classification [10]. We extract SURF

features for the views transformed by an affine model and select the most stable *n* local features by the frequency of SURF points. The resulting top *n* (say 100) SURF points are used to search visually similar product images on the Web.

4. SEARCH PRODUCT IMAGES FROM WEB

On the basis of stable SURF points, we search relevant product images collected from the Web. Instead of "bag-of-words" model, a Naïve Bayes Near Neighbor classifier [11] is employed to search visually similar images. To improve the retrieval efficiency, LSH[12] is employed to make approximate similarity queries that only examine a small fraction of the dataset. Given a video ad *v* with an FMPI image Q, Our goal is to find the optimal product image class C which can match the video ad *v:*

$$C^* = \arg\max_{C_p} P(C|Q) = \arg\max_{C_p} \sum_{i=1}^{n} \sum_{j \in I_k} f(p_i, p_j)$$

$$= \arg\max_{C_p} \sum_{j \in I_k} \sum_{i=1}^{n} f(p_i, p_j)$$

where $p_1, p_2, ..., p_n$ denotes the set of stable SURF features in FMPI image Q of video ad *v*, p_i the SURF point at *i* positions l_i in Q. $f(p_i, p_j)$ measures the similarity between feature point p_i in Q and point p_j in a product image I_k from the dataset,

$$f(p_i, p_j) = \begin{cases} 1, d(p_i, p_j) \le r \\ 0, else \end{cases}$$

where *r* is the indexing radius. To further incorporate rough spatial position of feature points [11], we augment SURF descriptor with the location information in the distance measure, and the L2 distance is denoted by:

$$d(p_i, p_j) = \| p_i - p_j \|^2 + \alpha \| l_i - l_j \|^2$$

where *l* is the location of a SURF point in an image, α is the weight of location information.

For traditional approaches in content-based image search, a key challenge lies in how to efficiently match feature sets, which involves two factors: (1) the high demand of memory space; and (2) the high complexity in computing the similarity between sets.

Differently we employ LSH technique to accomplish image matching. LSH performs probabilistic dimension reduction in high-dimensional space. The basic idea is to hash the input data using hash functions so that those similar points have much higher probability of being mapped into the same buckets (note that the number of buckets is much smaller than the universe of possible input points). Therefore, one can quickly determine near neighbors by hashing the query point and retrieving the elements in the buckets containing that point. In our scheme, ideally well matched product images can be determined by selecting the images having more feature points in the same index buckets as the query FMPI image. The stable SURF points function as the query points while those feature points in image dataset are retrieved from the corresponding buckets decided by LSH mapping. For a query FMPI image, the searching time of product images (100 points) with LSH in the naïve Bayes Near Neighbor classifier only takes 40ms, on the PC with CPU 2.0 and 1G RAM.

5. PRODUCT OR SERVICE RECOMMENDATION

Through visual search with FMPI image from video ad, we obtain the related product images of the web site. Now we have to reorganize and rank the raw search results of product images with the help of text information. By taking advantage of product tags from the web, such as name, brand and price, we try to reorganize the product search results and recommend more relevant products/services toward providing a rich recommendation to users based on the tag similarity. For example, when we recommend products for a digital camera video ad, the final recommendation are illustrated in Fig.3. Besides product image, our scheme can equip viewers with other useful product information such as name and price. User also can click the image hyperlink to access the origin web page to find more details.

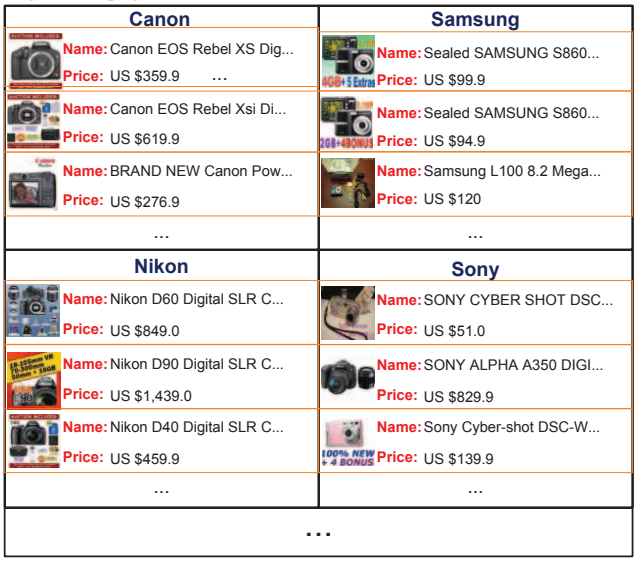


Fig.3. An example of digital camera recommendation.

6. EXPERIMENTS

6.1. Experimental Setting

The experimental data involves two parts: video ads and product information database containing images and tags. We have selected ten popular classes of video ads as listed in Table 1. Some 20~40 ad videos are included in each class. These video are mainly download from [14]. Our product information database includes about 20000 information items which can be divided into 18 categories by the types of products. Each piece of information item contains brand name, image, price, sales info, and other descriptions. Some 20000 product images are included. The resources of our product database mainly come from eBay [13].

6.1. Experimental results

In our experiments, the recognition accuracy of FMPI image up to F1 =90.2% is obtained with color, edge and texture features, Comparatively, texture features play a more important role. The combination of color and texture features results in a significant improvement of performance.

Now let us evaluate the performance of product/service recommendation by calculating the mean average precision (MAP) of different classes of video ads. For each product class, we first compute the average precision (AP) obtained by using the FMPI image from each ad video as the query and we average these AP to

obtain MAP of the class. AP is a common metric in information retrieval that measures precision at all depths of a search process and averages all measurements up to a given depth. Given a query and *k* relevant items excluding the query, let $rank_i$ be the rank of the *i*th retrieved relevant item, then average precision is defined as following:

$$\text{average precision} = \frac{1}{k}\sum_{i=1}^{k}\frac{i}{rank_i} \quad (6.1)$$

Table.1. The MAP scores at a depth of 100 of 10 types products.

Ad Video Class	MAP@100 Ours	MAP@100 [15]
Cell Phone	**0.28958**	0.11485
Chocolate	**0.26982**	0.20736
Coffee	0.12592	**0.17791**
Digital Camera	**0.13111**	0.08054
Electric Shaver	**0.24105**	0.20577
Mp3&iPod	**0.19760**	0.09101
Printer	0.06376	**0.10321**
Shampoo	**0.13805**	0.06126
Television	**0.09047**	0
Wristwatch	0.05784	**0.09830**

As listed in Tab.1, we have compared our approach with a recently reported image search approach [15] using 144-d color correlogram and 24-d Polynomial Wavelet Tree (PWT) in terms of recommending product info for 10 classes of video ads. We can see that our approach has achieved promising results for 8 classes of video ads like Cell Phone, Chocolate, Electric Shaver, etc.

In our current implementation, there are several factors affecting the precision of recommendation as below: (1) *The quality of an FMPI image*. The quality, resolution and content of a query ad video determine the goodness of FMPI. An input FMPI containing blur product image can lead to much worse results. This is the reason why search result of *Printer* is worse. (2) *Size and texture of product images in the dataset*. Large and texture-rich images usually have more feature points so they are easier to be retrieved. Experiments show that most of noisy images are related to one of above two factors. (3) *The appearance of products*. A class of products exhibiting several chief brands or more uniform shape such as cell phone tends to yield better results. For example, the lower result of *coffee* is due to its variant appearances from different packages, while the approach [15] with global features can obtains better performance.

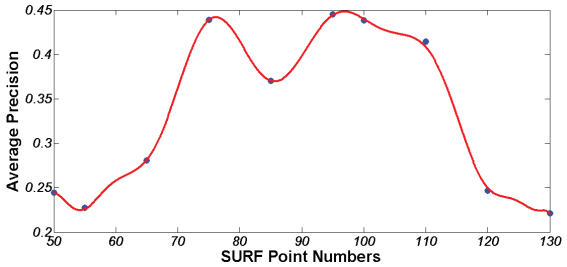


Fig.4 Recommendation results with different SURF point numbers.

Moreover, we investigate the effect of the number of SURF points on the recommendation. As shown in Fig.4, we can see that AP is improved with the increase of SURF points, while AP descends when the number of SURF points is greater than 100.

7. CONCLUSION

An online product/service recommendation system relating with video ads can combine various techniques in video analysis, image retrieval and multimodal fusion. We have proposed to link video ads with relevant product/service information across E-commerce websites or portals towards ad recommendation in a cross-media manner. Experiments have demonstrated the idea of our approach. In future we will improve the performance of ad recommendation by fusing visual and text information. Moreover, we will remove those duplicated (or near duplicated) product images to improve the efficiency of retrieval. In particular, such kind of linking can be combined with our proposed multimodal approach in [2] to furnish more comprehensive video ad categorization by products/services.

8. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China No. 60675003, 60833006, 60723005, and partially supported by the research fund from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and also partially supported by the research award of Microsoft Research Asia Internet Services Theme.

9. REFERENCES

[1] Junsong Yuan, Ling-Yu Duan, Qi Tian, and Changsheng Xu, "Fast and robust short video clip search using an index structure," in Proc. *ACM MIR'04*, 2004, pp. 61–68.

[2] Ling-Yu Duan, JinqiaoWang, Yantao Zheng, Jesse S. Jin, Hanqing Lu, and Changsheng Xu, "Segmentation, categorization, and identification of commercials from tv streams using multimodal analysis," in Proc. *ACM MM'06*, 2006, pp. 202–210.

[3] A. Shivadas and J.M. Gauch, "Real-time commercial recognition using color moments and hashing," in Proc. *ACM MIR'06*, Oct 2006.

[4] X.-S. Hua, L. Lu, and H.-J. Zhang, "Robust Learning-based TV Commercial Detection". *Proc.ICME'05*, pp.149-152.

[5] Tao Mei, Xian-Sheng Hua, Linjun Yang, and Shipeng Li, "VideoSense- Towards Effective Online Video Advertising," *ACM MM'07*, Augsburg, Germany, Sept. 2007.

[6] C. Colombo, A.D. Bimbo, and P. Pala. "Retrieval of commercials by semantic content: The semiotic perspective." *Multi-media Tools and Applications*, Vol. 13, pp.93-118, 2001.

[7] Jinqiao Wang, Lingyu Duan, Lei Xu, Hanqing Lu and Jesse S. Jin, "TV Ad Video Categorization with Probabilistic Latent Concept Learning," *ACM MIR '07*, Sep. 22, 2007.

[8] Tao Mei, Xian-Sheng Hua, Shipeng Li."Contextual In-Image Advertising" *ACM Multimedia*, pp. 439-448, 2008.

[9] Huiying Liu, Shuqiang Jiang, Qingming Huang, Changsheng Xu. "A generic virtual content insertion system based on visual attention analysis," *ACM Multimedia 2008*, pp.379-388.

[10] H. Bay, T. Tuytelaars, and L.V. Gool. "SURF Speeded Up Robust Features," *Proc. ECCV'06*, 2006.

[11] O. Boiman, E. Shechtman, M. Irani, "In Defense of Nearest-Neighbor Based Image Classification." *Computer Vision and Pattern Recognition (CVPR)*, Ankorage, June 2008.

[12] A. Andoni, P. Indyk, E2LSH0.1 *User manual*, June 2000.

[13] eBay. http://www.ebay.com

[14] YouTube. http://www.youtube.com

[15] Haoying Ding, Jing Liu and Hanqing Lu. "Hierarchical clustering-based navigation of image search results." *ACM MM'08*. 2008.