ONLINE VIDEO ADVERTISING BASED ON USER'S ATTENTION RELAVANCY COMPUTING

Jinqiao Wang, Yikai Fang, Hanqing Lu

Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
{jqwang, ykfang, luhq}@nlpr.ia.ac.cn

ABSTRACT

Information overload has become an important problem in the internet, and that all kinds of existing ads flood into people's eyes causes scarcity of user's attention. To provide relevant information under user's control, we propose an online video advertising framework based on user's attention relevancy computing. Users receive relevant video ads in exchange of their attention consumption. Multimodal concept detectors are trained to annotate the video databases, and a multimodal video ads categorization and related concept-to-ad relevancy and ad-to-concept relevancy ranking algorithm are proposed to compute user's attention relevancy. Experiments and a subjective evaluation show the feasibility and effectiveness of the proposed approach.

Index Terms— Video ad, concept retrieval, ad insert

1. INTRODUCTION

With the development of information technology and storage ability, news, blogs, podcasts, photos, videos and cool space pages etc are increasing more and more. Online ads especially the video ads along with the web pages are increasing rapidly. But all the existing types of ads are incrusive and irrelevant to the users, which makes user boring and offensive. Also the ads make it increasingly difficult to juggle all the news sources and keep on top of things. Recently, content-targeted or contextual advertising systems, such as Google's AdSense program [1], Yahoo's Contextual match product and ValueClick's Ad Network [2] are becoming an increasingly important part of the revenue source for today's web. Typical content-targeted advertising systems [7] analyze a web page, such as a blog, a news page, or another source of information, to find representative keywords on that page. These keywords are then sent to an advertising system, which matches the keywords against a database of ads. Advertising appropriate to the keyword is displayed to the user.

Existing Online video ads are mainly concentrated on ad insertion on the start, middle and end of the videos [3], spatially replacing a specific region with product advertisement in sports videos, and personalized ad insertion in an interactive TV environment. There are several problems in these systems. The first is that most video ads inserted in video streams are intrusive and boring without considering user's attention. The second is that traditional video ads are inserted in the middle of the program or in highlights of the sports video undermines the continuity of the video program. The third is textual keywords [3] used in online video advertising is not enough for measuring the relevancy of rich content videos.

We propose to online video advertising based on user's attention relevancy computing. There are two characteristic: one-to-one relevancy service and under user's control. As long as a consumer sees relevant content, he/she is going to stick around and that creates more opportunities to sell. Literally, the longer a user stays on a site reading news or watching videos etc, the higher the chance that person will click on the ads. For online video broadcasting, since some concepts or objects ask for consumer's attention in exchange for the opportunity to show him/her advertising. Control is not just to protect consumer's information but also to put the user in control of her information. The user chooses what services he/she wants to receive, in exchange for their attention information.

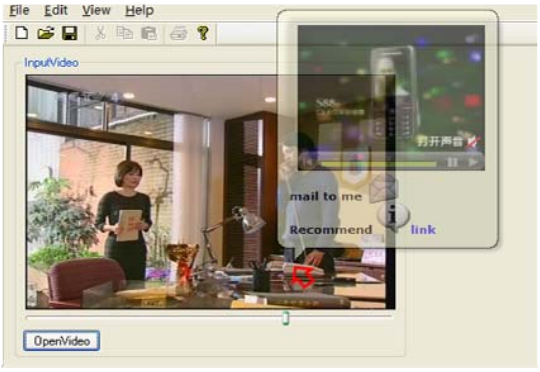

Fig.1. An example of online video advertising with user's attention relevancy computing. The cell phone video ad is triggered when the user's mouse (red arrow) stops at the phone.

Fig.1 gives an example of our online video advertising model, which proposes concepts, objects, scenes, entity and logo etc based video ad placing. If the user's mouse stops on the interested objects, concepts etc in the video, the video ads are triggered to in exchange for user's attention by a recommend engineer with a semi-transparent window overlapping in the video player window. If the user moves his mouse to the ad, then the semi-transparent ad windows changes into an opaque one. Also the user can visit the product site or recommend the product to his friend.

2. ONLINE VIDEO ADVERTISING FRAMEWORK

The online video advertising framework includes semantic concepts (the objects, texts, scenes or highlights) annotation, video ad categorization and video ad recommendation through user's attention relevancy computing. As shown in Fig.2, the videos are annotated by multimodal concepts detectors before broadcasting in the internet. Video ads are classified based on product and service

with visual and textual features. Finally, textual and visual attention relevancy ranking including concept-to-ad relevancy and ad-to- concept relevancy is computed and combined by the recommendation engineer.

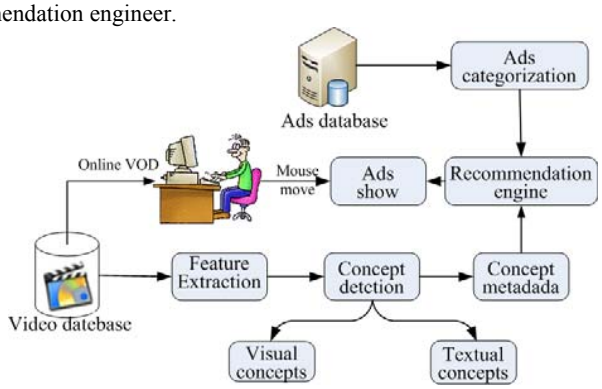

Fig.2. Online video advertising framework.

3. SEMANTIC CONCEPT DETECTION WITH MULTIMODAL FEATURES

The type of online video advertising based semantic concepts can be classified into region based, frame based or shot/scene based ads. We mainly focus on the user's attention relevancy computing based on the three types. Like the rules in concept selection in Kodak's consumer video benchmark database [6], we choose semantic concepts as detectability, observability and relevance to user's attention. Our lexicon includes semantic concepts related objects, scene, occasion, people, and camera motion. In our system, the manually constructed visual semantic space comprises 13 concepts, which are summarized in three categories as:

a) People (people, face, hair).
b) Occasion (indoor, outdoor, road, office, sky, water, building).
c) Object (car, phone, credit card).

Our concept detection approach is based on local features. Local features have been shown to be powerful for their invariance to occlusions and viewpoints [9]. We extract local visual features in each frame for region based and frame based advertising, while in key frames of a video for scene or shot based advertising. The local regions are extracted by random sampling, Harris, and SURF [10], respectively. The Harris detector locates corner-like regions, and the SURF detector extracts blob-like regions. We employ the 36-dimensional PCASIFT descriptor [9] to compute a gradient orientation histogram for each local region. The region descriptors are quantized according to the nearest neighbor rule. Subsequently, a visual vocabulary is constructed by applying K-means clustering to all the local descriptors extracted from training images, and those means remain as visual terms. Small clusters are pruned out as noises. Finally, each ad video can be represented as a collection of visual words.

With the bag-of-words model, we employ SVMs to complete a series of binary supervised learning of concept classifiers expect human face. The face detector in [11] is applied to detect faces. For other twelve concepts, we train twelve SVMs-based binary classifiers. With all the concept detectors to annotate the videos before broadcasting, we can provide user attention relevant precision video ads based on user's interest and profile.

4. MULTIMODAL VIDEO ADS CATEGORIZATION

Before recommending a video ad to user based on the relevancy computing, ad categorization based on production and service is the first step. We introduce PLSA models to automatically discover latent visual and textual concepts for representing ad videos [4]. Co-occurrence of local visual features and expanded textual features is modeled to represent ad categories in latent semantic space. Different from discovering aspects in documents or images, our PLSA models work on the bag-of-words representation of videos. That is, the PLSA models have to work on local visual features and textual features derived from a set of key frames. We use the joint probabilistic distribution of latent concepts to represent ad categories.

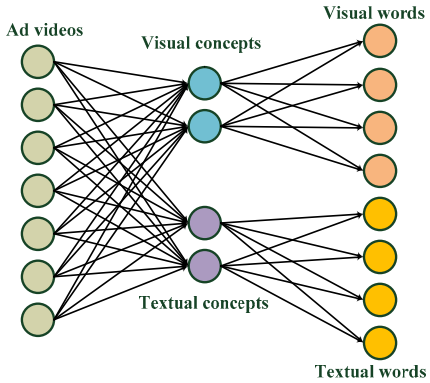

Fig.3. Multimodal ad categorization by PLSA model

PLSA modeling of visual and textual concepts is illustrated in Fig. 3. Given a collection of ad videos $C = c_1, c_2, ..., c_n$, we want to represent an ad as a bag of visual words from the visual vocabulary $W_v = \omega_1^v, \omega_2^v, ..., \omega_m^v$ and a bag of textual words from textual vocabulary $W_v = \omega_1^t, \omega_2^t, ..., \omega_s^t$. The collection of ads is thus represented by a $m \times n$ visual co-occurrence matrix M^v and a $s \times n$ textual co-occurrence matrix M^t. By the PLSA model, we can derive unobservable latent visual concepts (aspects or topics) $z_e^v (z_e^v \in (Z^v = z_1^v, z_2^v, ..., z_p^v))$ associated with the occurrence of a visual word ω_i^v in an ad c_j, and unobservable latent textual concepts $z_e^t (z_e^t \in (Z^t = z_1^t, z_2^t, ..., z_q^t)$. The numbers of latent concepts p and q are predefined; however, the learning of concepts runs in a data-driven manner. According to Bayesian rules, the probability of an observation (ω_i, c_j) in adopting the latent concept $z_k (z_k^v \; or \; z_k^t)$ is modeled as:

$$P(\omega_i, c_j) = \sum_{z_k \in Z} P(\omega_i \mid z_k) P(z_k \mid c_j) \qquad (1)$$

Where $P(z_k \mid c_j)$ is the probability of latent concept z_k occurring in ad c_j, $P(\omega_i \mid z_k)$ is the probability of word $\omega_i (\omega_i^v \; or \; \omega_i^t)$ occurring in latent concept z_k. We can identify the visual/textual words belonging to a latent concept by ranking words by $P(\omega_i \mid z_k)$. In the context of ad categorization, the probabilistic distribution of latent concepts in an ad video is then encoded by

parameters $P(z_k \mid c_j)$, which are finally applied to classify ads in a semi-supervised manner.

The PLSA model expresses each ad video as a convex combination of the aspect-specific distributions of p latent concepts. The unobservable category-related latent concepts $z_e^v (z_e^v \in Z^v = z_1^v, z_2^v, ..., z_p^v)$ are thus determined by the PLSA model. Referring to Eq.1, the conditional probabilities $\{P(z_1^v \mid c_j), P(z_2^v \mid c_j), ..., P(z_p^v \mid c_j)\}$ can be used to represent ad videos in the latent semantic space characterized by the visual concepts $\{z_1^v, z_2^v, ..., z_p^v\}$. According to Bayesian rules, a visual word is represented in the latent semantic space by a parametric formula $P(z_k \mid \omega_i) = P(z_k)P(\omega_i \mid z_k) / P(\omega_i)$. Visual words associated with similar semantics are often generated by one latent concept, while a visual word with multiple semantic meanings can receive higher generative probabilities from more than one related concepts. Finally, an ad c_j is represented as:

$$\varphi(c_j) = \left(P\left(z_1^v \mid c_j\right), P\left(z_2^v \mid c_j\right), ..., P\left(z_p^v \mid c_j\right)\right) \quad (2)$$

Likewise, PLSA is utilized to model the latent semantic space for ads documents. The latent concepts $z_e^t (z_e^t \in Z^t = z_1^t, z_2^t, ..., z_q^t)$ are consequently formed to represent the ad video c_j as:

$$\phi(c_j) = \left(P\left(z_1^t \mid c_j\right), P\left(z_2^t \mid c_j\right), ..., P\left(z_q^t \mid c_j\right)\right) \quad (3)$$

Based on the visual and textual concepts, we represent an ad by $\left(\varphi\left(c_j\right), \phi\left(c_j\right)\right)$ according to Eqs. (2) & (3). We resort to two SVMs to visual and textual features, respectively, and then linearly combine the soft outputs of two classifiers.

5. MULTIMODAL VIDEO ADS RANKING USER'S ATTENTION RELAVANCY COMPUTING

A video ad that is triggered by users is relevant to user's attention. How to rank the ad videos as the user's favorite concepts or objects is critical for online video advertising. Vector model is used in [8] to calculate the textual relevancy between ads. In this paper, in addition to textual ad ranking, visual ad ranking is also chosen in the relevancy computing. To embody the relevancy to user's attention, we combine concept-to-ad relevancy and ad-to- concept relevancy to rank the ad videos.

For a concept D_x and a video ad A_y, the attention relevancy is defined as:

$$R(D_x, A_y) = \omega_1 R_{ca}(D_x, A_y) + \omega_2 R_{ac}(D_x, A_y) \quad (5)$$

Where ω_1 and ω_2 are the weights. R_{ca} denotes the relevancy from the concept D_x to the video ad A_y, R_{ac} denotes the relevancy from the video ad A_y to the concept D_x. For each user attention region, object or scene, we expand the shot to five shots and detect the latent visual and textual concepts used in video ads in shot level, then R_{ac} is calculated by the cosine distance of the concept vectors:

$$R_{ac} = \frac{P_v(A_y) \cdot P_v(D_x)}{\| P_v(A_y) \| \cdot \| P_v(D_x) \|} + \frac{P_t(A_y) \cdot P_t(D_x)}{\| P_t(A_y) \| \cdot \| P_t(D_x) \|}$$

Where $P_v(\cdot)$ and $P_t(\cdot)$ are the concepts vectors respectively.

Similar to the calculation of R_{ac}, each concept detector is applied to the key frames of all video ads. R_{ca} is the average probabilistic output. Through the combination of the visual, textual relevancy to user's attention, and the concept-to-ad, ad-to-concept relevancy computing, the final video ad ranking can better meet user's requirement.

5. EXPERIMENTS

Our experimental online videos use five long videos including two movies, one home video and two sitcoms. Video ads data are extensively collected from TRECVID'05 & '06 news corpus and several Chinese TV channels. We collect in total 406 distinct ones including 191 distinct English ones used in [5]. All the ad videos are in MPEG-1 format (29.97 fps, 352 × 240). By their advertised products/services, such 406 ads are distributed in 8 classes, i.e., Automobile, Finance, Health care, IT, Food, Beauty Products, Furniture, others. For evaluation purposes, we further form 4 subclasses: car, credit card, body care and phone, which belong to Automobile, Finance, Health care, and IT, respectively, each subclass having 50 ads. The selection of experimental categories considers three factors: percentage distribution, closeness to our daily life, and of course algorithm evaluation.

In our experiments, we will give the semantic concepts annotation results, ad video categorization results with visual and textual features, and a user study is performed to illustrate the performance of the online video advertising approach.

5.1. Results of Semantic Concepts Annotation

To explore the semantic concepts in movies, news and sitcoms, we explicitly model 13 semantic concepts including indoor, outdoor, road, office, sky, water, building, car, phone, credit card, people, face, and hair. The learning of these scenes and objects completely resorts to external image resources comprising public databases such as TU Darmstadt, UIUC car, VOC 2006, Caltech and MIT-CSAIL, and some images collected from Google search. The training data size of each concept ranges from 400 to 1000 images. Concept detection is applied to key frames of a video. Given an image, each concept classifier has a probability output to determine the presence of a concept. Note that not each key frame can be classified into one of 13 concepts, and a key frame can be classified into multiple concepts.

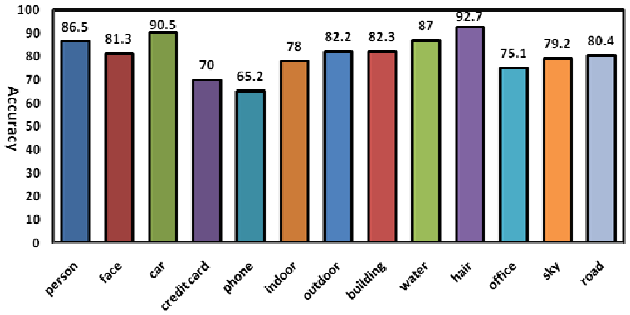

Fig.4. Detection accuracy of 13 visual concepts.

In our experiment, some 600 local regions are extracted from a key frame. PCASIFT descriptors are utilized to represent the regions. The codebooks are learned by applying k-means to the descriptors of local regions over training images. Small clusters are

removed. A separate codebook comprising 240 to 350 visual words is formed for each visual concept. Fig.4 shows the average detection results over our video dataset.

5.2. Results of Video ads Categorization by Products or Service

	Car	Body care	Phone	Credit card
Car	26	0	2	2
Body care	0	27	1	2
Phone	1	1	26	2
Credit card	2	5	2	21

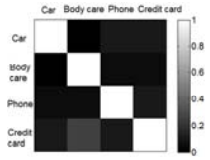

Fig.5. Confusion matrixes of ad categorization by fusing multi-modal features with two SVMs.

We employ the linear combination strategy for visual and textual SVM outputs. Classification performances are evaluated as listed in Fig. 5. The weights for linear combination are set equal in the experiments. The linear fusion has achieved the best performance of 86.7% (car), 93.3% (body care), 90% (phone), and 76.7% (credit card). Fusing multi-modal concepts has greatly improved the results. Promising performance reveals our approach's applicability.

5.2. Subjective Evaluation on Online Video Advertising

To evaluate the performance of Video advertising, we conducted a subjective user study to evaluate our work. Five evaluators were invited to participate in the user study, and each individual was assigned with five videos. When viewing each of the online video advertising result, the evaluators were asked to give a score from 1 to 5 (higher score indicating better performance) to show their performance level based on the following aspects:

- Intrusiveness. Is the video ads are intrusive for users?
- Relevancy. How about the relevancy between the semantic concepts and video ads?
- Continuity. Do the video ads interrupt your enjoying movie?
- Acceptability. Whether can you accept the video ads?

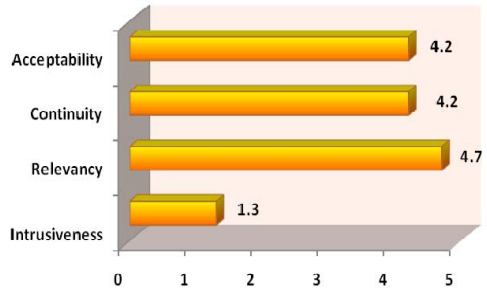

Fig.6. Average subjective evaluation results of the online video advertising in five videos.

The average results for the five videos are listed in Fig. 6. We can see that the intrusiveness level is lower for all the users, and most users are acceptable for this kind of online advertising system. Without the video ad inserted in the middle of the movies, users are also satisfied with the continuity of the movie content. The average subjective evaluation results are higher than [3] for we use a more flexible way to use user's attention to in exchange for the ad service. With the introduction of the visual relevancy and through the combination of the concept-to-ad relevancy ranking and ad-to-concept relevancy ranking, the users are satisfied with the ad relevancy to their attention.

6. CONCLUDIONS

In this paper, an online video advertising approach is proposed based on user's attention relevancy computing. The video data are annotated by multimodal concept detector, and the video ads are categorized by PLSA base visual and textual features. A multi-modal user's relevancy ranking algorithm is applied to combine concept-to-ad relevancy and ad-to-concept relevancy. Finally, a user study shows the performance of the proposed approach. Through the exchange of user's attention to the relevant video ads, the online video advertising is effective.

6. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No. 60675003), and the 863program No. 2006AA01Z315.

11. REFERENCES

[1] Google AdSense. www.google.com/adsense.

[2] ValueClick. http://www.valueclick.com/.

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

[4] 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.

[5] Lingyu Duan, Jinqiao Wang,Yantao Zheng, and Jesse S. Jin, Hanqing Lu and Changsheng Xu, "Segmentation, Categorization, and Identification of Commercial Clips from TV Streams Using Multimodal Analysis," *ACM MM'06*, 2006.

[6] A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak Consumer Video Benchmark Data Set: Concept Definition and Annotation." *ACM MIR'07*, Augsburg, Germany, September 2007.

[7] S. Mccoy, A. Everard, P. Polak, and D. F. Galletta, "The effects of online advertising," *Communications of The ACM*, 50(3):84-88, 2007.

[8] W.-T. Yih, J. Goodman, and V. R. Carvalho, "Finding advertising keywords on web pages", *WWW'06*, 2006.

[9] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors." *Computer Vision and Pattern Recognition*, 2004.

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

[11] Chang Huang, Haizhou Ai, Yuan Li, and Shihong Lao, "Vector Boosting for Rotation Invariant Multi-View Face Detection." *Proc.ICCV'05*, pp.446-453, Beijing, China, 2005.