Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 7, July 2018, pp. 1411-1420
DOI: 10.23940/ijpe.18.07.p4.14111420

Discriminative Image Representation based on Multi-Cues for Computational Advertising

Zhize Wu[a], Shouhong Wan[b], and Ming Tan[a,*]

[a]*Department of Computer Science and Technology, Hefei University, Hefei, 230601, China*
[b]*School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230027, China*

Abstract

Image representation is a key step in image advertising recommendations. Traditional image representation methods, based on the local description, generate a histogram of visual words to represent images. However, it is very difficult to establish a discriminative and descriptive codebook with the local description only. Therefore, we propose a novel image representation method by integrating visual saliency, color feature and local description. Moreover, the proposed multi-cues image representation has been applied to a new image advertising scenario, i.e., delivering image advertisements in a list of images, such as the results of an image search. To evaluate our proposal, we have crawled a dataset, named Pop2016, which consists of image lists and advertising images with 31 pop labels. The performance of the advertising recommendations is measured in terms of the precision@n and the mean average precision. Experimental results show that the proposed algorithm outperforms several traditional methods.

Keywords: image advertising; computational advertising; image representation; multi-cue; visual word; visual saliency

(Submitted on April 13, 2018; Revised on May 25, 2018; Accepted on June 16, 2018)

1. Introduction

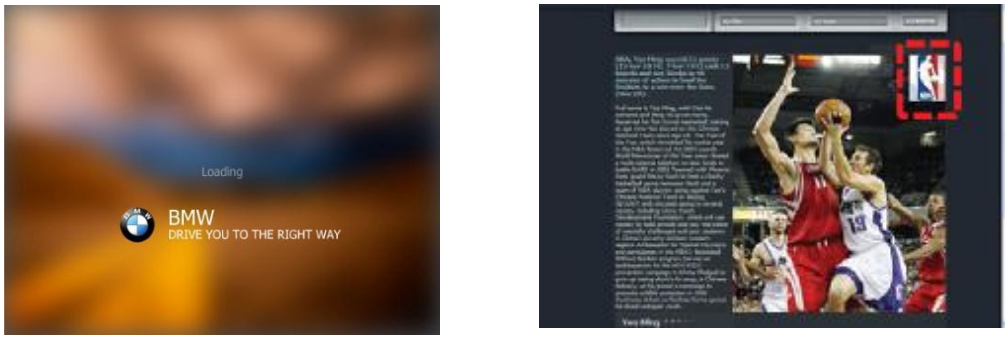
Advertisements are extremely concerned by all walks of life and are the major revenue stream for many internet enterprises. Differing from traditional advertisements, internet advertisements are a combination of many newly-developing technologies, i.e., large scale search and text analysis, information retrieval, statistical modeling, machine learning, classification, optimization, and microeconomics [1]. With the development of these skills in internet advertisements, computational advertising is an emerging new scientific domain at the intersection.

There are two main types of advertising deliveries: Guarantee delivery (GD) and Non-Guarantee delivery (NGD). GD delivers a fixed advertisement at an advertising space purchased from the website by the advertiser. Thus, an identical advertisement is presented for all users, even though it is likely not one's cup of tea. On the other hand, NGD delivers different contents according to the interests of the users. Compared to GD, NGD is more flexible and efficient. Thus, designing a reasonable recommendation algorithm for delivering suitable advertisements to a given user is a significant study. Most of the NGD algorithms still focus on delivering text advertisements based on natural language processing. However, users are more likely interested in visualized forms, e.g., images and videos. Therefore, research on how to deliver image advertisements is renewed and promising.

Existing works for image advertising in the field of NGD mainly include two types of methods. The first type is to present an advertisement during the image loading [12], as shown in Figure 1(a). The advertising time of this kind of method, however, is short, thus influencing the advertising effect. The second type is to deliver a related image advertisement based on the text labels of the images or the contents of the websites [3], as shown in Figure 1(b). However, most of the images on the web do not have a text label. Although the content of the website can also be used to extract the main features, the advertisement recommended based on the selected features may be unrelated to the actual meaning of the

* Corresponding author.
E-mail address: hftm163@163.com

image, because the meanings of the extracted features can vary. It is difficult to select related features without using the knowledge of the image.



(a) (b)

Figure 1. Forms for delivering image advertisements

To the best of our knowledge, thus far, only the studies in [16] and [18] consider the content of the image to recommend advertisements. In [16], the image features combined with the text features are used to deliver text advertisements, rather than to deliver image advertisements directly. In [18], the content of an image is transformed to a four-layer structure, including the theme, the semantic area, the visual word, and the pixels, to deliver a related image advertisement. However, the locations of the advertisements are neglected, because this might affect the users' experience. Furthermore, for both [16] and [18], the knowledge about the images is only extracted from the local shape features of the images without using other features, such as visual saliency and color information.

In this paper, we first propose to extract the visual salience and color information, in addition to the local descriptions, from the images to represent these images. Then, we apply this image representation method for a new image advertising scenario, i.e., delivering image advertisements in a list of images (e.g., the results of image searches or photo albums), instead of in a single image. For such a scenario, the most relevant image advertisement is first determined to be delivered based on the similarity between the image advertisement and the image list. The similarity is calculated based on the proposed image representation method. Then, a suitable placement for each image advertisement is calculated according to the similarities between the advertisement and the image list to prevent the destruction of the website's structure. Thus, we can recommend the advertisements that users will more likely be interested in and improve their quality of experience.

The remainder of the paper is structured as follows. Details of the proposed application scenario are described in Section 2. Section 3 presents the framework of the proposed image advertising algorithm. The experimental results are shown in Section 4. Section 5 concludes this work.

2. Proposed Application Scenario

We design a new image advertising scenario for a list of images (e.g., image searching results, web albums, results of the search when shopping online), as shown in Figure 2. Figure 2(a) shows the image results of searching "beach" on Google. We can notice that the image list is the main body of the webpage, and these images are with the same label. In actuality, there are many similar web pages when users search for their interesting keywords or manage their photo albums.

(a) (b)

Figure 2. (a) Results of image searching in Google; (b) Example of advertisement recommendation

At present, the great majority of computational advertising studies mainly concentrate on the general web pages, but neglect the image list web pages. Actually, the latter webpages can truly reflect the interests of a user. For example, as shown in Figure 2(b), when a user searches "beach", we can infer that the user may be interested in beach tourism. By inserting a relevant advertisement in the image list, the user is more likely to click on it. Unlike the traditional method of finding interests based on the images uploaded by a given user, this proposal is more adapted to the user's changing interests.

We can find that delivering advertisements in the results of image searches or web albums is a pretty promising study. For this application scenario, we propose a novel image advertising algorithm based on the multi-cues representation, which is a combination of different features including shape, color, and visual attention. Through this image representation method, the most similar advertisement can be found and placed in a suitable place without influencing the user's experience.

3. Proposed Application Scenario

3.1. Algorithmic Sketch

The aim is to recommend the best advertisement and put the advertisement in a suitable location. Because the images with the same label may have various contents because of the ambiguity of the meaning, as shown in Figure 3, the similarity between each of the two images should be calculated based on the content of images directly, instead of only based on the text label. We propose to calculate the similarity based on a multi-cues image representation. Namely, an image is represented based on the combination of visual saliency [19], color information, and local descriptions, rather than the local descriptions only.

Figure 3. Advertisements list with the same label "apple"

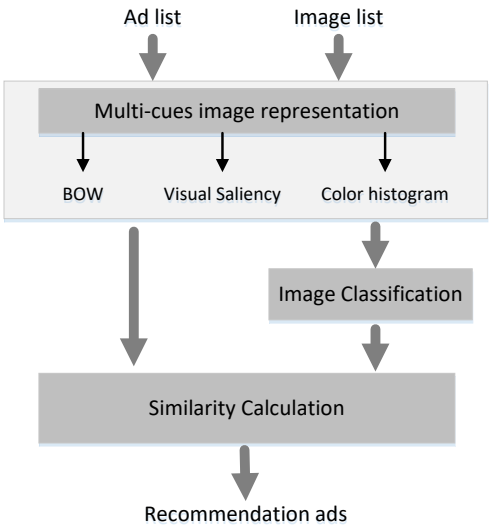


Figure 4. Sketch of image advertising recommendation

The basic idea of the proposed algorithm is as follows and illustrated in Figure 4. First, each image in the advertisement set and the image list is represented based on the combination of shape, color, and visual attention. Then, divide the image list into several clusters based on similarity. Finally, replace images from the cluster with several of the most relevant advertisements. Both the relevance of the advertisements and the location of the replacements are calculated based on the similarity between the advertisements and the images.

3.2. Algorithm Description

Formally, given a set of labels $C=\{c_1,c_2,\cdots,c_K\}$, an image list $I\_List_k=\{I_1,I_2,\cdots,I_M\}$ with M images, and an ad list $Ad\_List_k=\{ad_1,ad_2,\cdots,ad_N\}$ with N advertisements, the essence of this problem is sorting Ad_List_k based on the relevance between ad_n and I_List_k for each class c_k. As shown in Figure 4, the key steps of our image advertising algorithm include image representation, classification, similarity calculation, and delivering placement.

Image Representation. The most popular image representation model, the Bag-of-Words (BOW) model, first extracts various descriptors of an image. Then, it clusters these descriptors into visual words. Finally, an image is represented by coding and pooling its descriptors into a histogram based on the codebook. The process is shown in Figure 5.

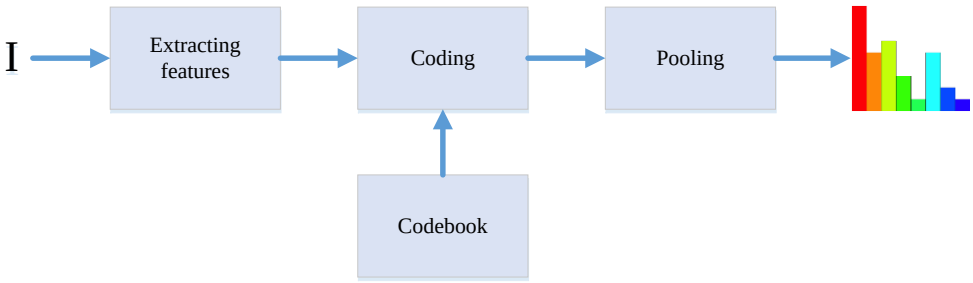


Figure 5. BOW-based image representation scheme. It contains three steps: extracting features, coding, and pooling

However, the BOW model treats an image as an orderless collection of visual words for which many details may be lost. Thus, we propose an improved BOW model, in which the visual saliency information is considered. Based on the visual saliency information, the foreground of an image, containing the most essential object, can be extracted. And the background is ignored because it is helpless for image understanding.

Motivated by the spatially local coding method [11], we represent an image as the combination of the extracted features and the salient values (i.e., the visual saliency information) of these features. The spatial information of these features can be partially considered through this representation, because whether a feature belongs to the foreground can be judged using its salient value. As a result, we do not need to compute the histogram for each region. Therefore, the dimension can be reduced markedly. This image representation scheme based on the visual saliency and BOW is shown in Figure 6. Specifically, first, the features should be extracted and the salient maps should be computed for both training images and testing images. Then, depending on the extracted descriptors and the salient maps, a new feature vector v_i consisting of an appearance feature vector a_i and its visual saliency information $s(x_i,y_i)$ is constructed as Equation (1):

$$v_i=\left[a_i,\lambda s(x_i,y_i)\right] \tag{1}$$

where $s(x_i,y_i)$ is the salient value of the position (x_i,y_i) on the saliency map computed according to the recent studies [6, 11]; and λ is a visual saliency weighting factor controlling the importance of the visual saliency in a feature vector. Thus, image *I* can be represented by Equation (2):

$$I\mapsto\left[v_1,v_2,\cdots,v_n\right] \tag{2}$$

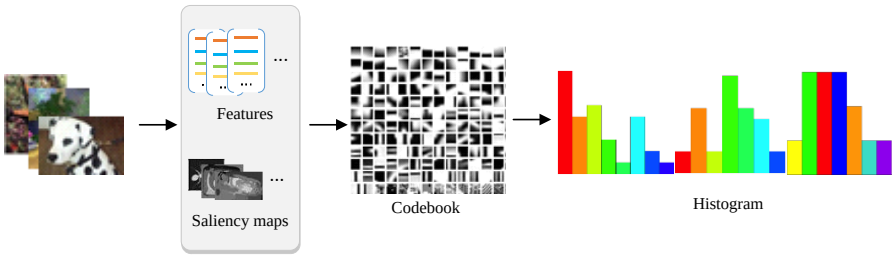


Figure 6. Image representation based on visual saliency and BOW

After that, we cluster all feature vectors $\{v_1, v_2, \cdots\}$ in the training image set to build a codebook $z = [z_1, z_2, \cdots z_k]$. Each visual word z_k in the codebook has already contained visual saliency information. We then represent all training images as histograms by coding and pooling.

Furthermore, the color features (RGB color histogram, shown as Equation (3) are also considered in the coding scheme to improve the descriptive and discriminative ability of the image representation. Thus, each image is represented based on multi-cues, including shape, salience, and color feature.

$$I \mapsto \{\mathrm{r}_1, \mathrm{r}_2, \cdots, \mathrm{r}_P, g_1, g_2, \cdots, g_P, b_1, b_2, \cdots, b_P\} \tag{3}$$

Image Classification. As mentioned before, various contents may exist in the image list with the same label because of the ambiguity of the meaning of the label. As a result, the various contents might influence the performance of advertising recommendations. However, it is impractical to filter these images by traditional image classifiers for large-scale image processing, because a classifier is required for each potential label. Fortunately, existing studies [2, 7, 17, 18] show that an image similar to the labeled image has a similar label in most cases, as shown in Figure 7. Therefore, we propose to perform unsupervised classification for the images with the same label, i.e., dividing the images into several clusters directly according to the image contents.

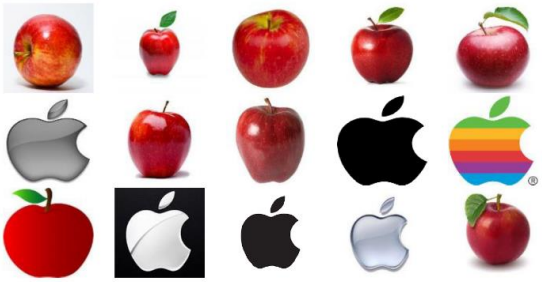


Figure 7. Image list with the same label

The classification process is given as follows. Each image *I* is represented as a feature vector which is a combination of the BOW saliency histogram and RGB histogram, shown in Equation (4).

$$I \mapsto \{\mathrm{z}_1, \mathrm{z}_2, \cdots, \mathrm{z}_K, \mathrm{r}_1, \mathrm{r}_2, \cdots, \mathrm{r}_P, g_1, g_2, \cdots, g_P, b_1, b_2, \cdots, b_P\} \tag{4}$$

Then, *k*-means is adopted to cluster these images. As the BOW histogram includes more local features than the RGB histogram, we set $k > p$. Because the number of images in each cluster is different (as shown in Figure 6) and the cluster with the most number of images is the major part, we choose the images in this major cluster to conduct advertising recommendations.

Similarity Calculation. The placement of an image advertisement should be close to the relevant images when the recommendation of the advertisement is based on the image content because it is very important for maintaining the quality of the user's experience. However, the images of the major cluster may be disconnected in location. For example, in Figure 8, the major cluster contains three groups of images, i.e., G_1, G_2, G_3. Obviously, users may be more likely to pay attention to G_1. Thus, we first find the largest connected region by calculating each connected region $\{\mathrm{G}_1, \mathrm{G}_2, \cdots, \mathrm{G}_N\}$ based on the locations of the images. Then, the relevance between an advertisement and each image in the selected region is calculated.

Specifically, the advertisements and the images of the major cluster are first represented by the feature vector given in Equation (4). Then, the relevance between each advertisement and each image in the major cluster is computed using Equation (5). Note that the distance function here is L_2 norm (Euclidean distance). Based on $Sim(ad_i, \mathrm{G})$, we can sort these advertisements and determine which advertisement to deliver.

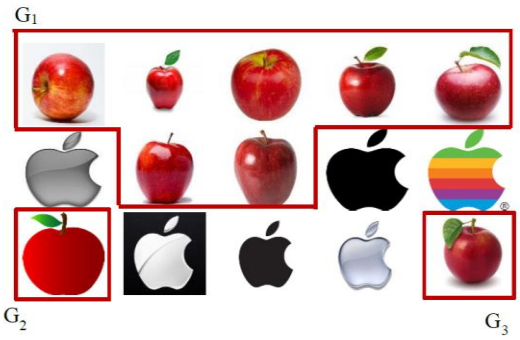


Figure 8. Image group in the image list with the same label

$$Sim(ad_i, G_j) = \Big/ \sum_{j=1}^{N} norm\_l2(\mathrm{ad}_i, \mathrm{I}_j) \tag{5}$$

Delivering Placement. Because the advertisements to be delivered are selected based on their similarity with the contents of the images, the advertisements should be placed around these images to maintain the quality of the user's experience. Specifically, the lowest relevant image from the images around the largest connected region with the advertisement image is calculated according to Equation (5). Then, replace this image with an advertisement.

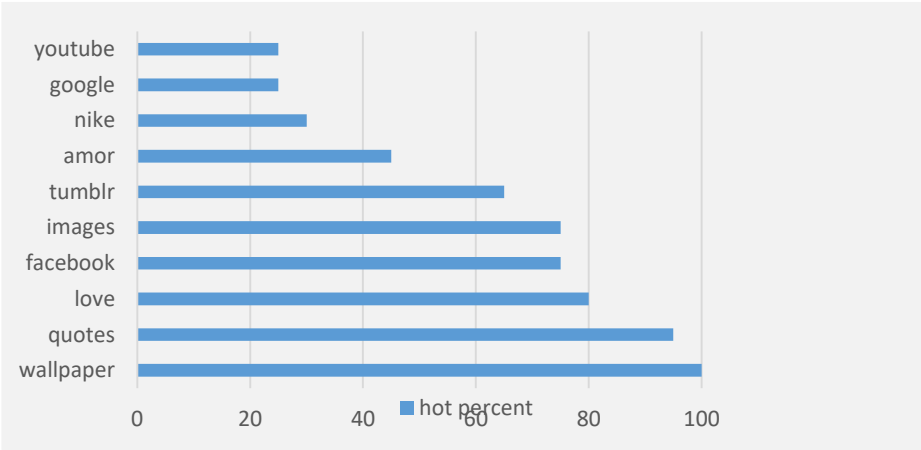


Figure 9. Top 10 queries in 2016

4. Experiments

4.1. Dataset

Currently, there is no existing dataset to verify the proposal. Thus, we crawl the dataset from the internet. Because our focus is computational advertising, we should consider the pop labels as much as possible so that the advertisements are delivered to more users. The hot keywords of image queries can be obtained from Google Trends. Figure 9 shows the top 10 queries in 2016. We select 31 keywords as the labels, as shown in Table 1. Then, we search in Google Images with each label and take the top 20 images to form the corresponding image list. Meanwhile, we type the label in Google Search to select the top 10 Google advertisements that consist of images for the advertisement list. For each class, we manually sort the advertisement images for evaluating our algorithm later on. We term this dataset Pop2016.

Table 1. Label set of Pop2016 (arranged in alphabetical order)

2016	apple	art	baby	beach	car
cat	Christmas	dog	face	facebook	flower
football	galaxy	game	girls	hairstyle	Halloween
London	love	map	minecraft	Nike	quotes
school	sea	tattoos	thanksgiving	tumblr	wallpaper
worldcup					

4.2. Experimental Setup

When the dataset is ready, we extract dense SIFT for all images in Pop2016. Dense SIFT [15] is a dense vision of SIFT [8], and the comparative evaluation of Fei-Fei and Perona [6] has shown that dense features work better for classification. Moreover, the dense SIFT extracts SIFT descriptors at every position of the image while the normal SIFT descriptors are only extracted at some regions detected by DoG, so dense SIFT is more suitable than SIFT or other descriptors that are used for analyzing the different salient value of visual words in the different location. Here, we extract dense SIFT features with a step width of 8 pixels, and the descriptor is extracted at 8*8 pixels.

We randomly choose 1,000,000 descriptors from the training images and cluster them to build a codebook. Instead of using *k*-means to cluster visual words, we use FLANN (Fast Library for Approximate Nearest Neighbors) [11] to cluster descriptors approximately during each iteration of *k*-means, just as in [10]. We conduct experiments at 90% *k*-means accuracy. For the color feature, we use the RGB histogram with 24 dimensions. Thus, each image is represented by the proposed multi-cues representation. We show the 20 images of an image list in the form of a 4*5 grid. The main body of each image list is calculated using *k*-means. Finally, we sort the advertisements based on their relevancies with the main body.

The performance of the proposed algorithm is measured in terms of precision@n (P@n) and mean average precision (mAP). The forms of these indicators are given in Equation (6) and Equation (7).

$$P@n = \frac{\sum_{i=1}^{n} \pi_i}{n} \tag{6}$$

$$\mathrm{m}AP = \frac{\sum_{i=1}^{n} \phi_i}{n} \tag{7}$$

Here, π_i is the recall of the *i*-th recommended advertisement. When $n = 5$, taking the Top 5 among the recommendations, if *i*-th recommended advertisement is one of the Top 5, π_i=1, else π_i=0. ϕ_i indicates the precision of the *i*-th recommended advertisement and can be calculated using Equation (8), where i_m is the recommendation manual index and i_c is the computational index.

$$\phi_i = \min(\mathrm{i}_m, \mathrm{i}_c) \Big/ \max(\mathrm{i}_m, \mathrm{i}_c) \tag{8}$$

4.3. Experiment Results and Discussions

Comparison Experiments. In the first experiment, we compare our method with the basic BOW model [5] to verify that adding visual saliency and color information can improve recommendation performance. We use hard vector quantization coding and average pooling in both methods. Table 2 shows the results.

It can be seen that for the same codebook size, much better results are obtained using our method. When the codebook size increases, both P@n and mAP also increase.

We also compare our method with the Spatial Local Coding (SLC) method [10]. The parameter λ is set to 1.5 as in [10]. The soft assignment, i.e., choosing the nearest 10 neighbor visual words to assign a descriptor, is adopted for both methods in the coding step. Table 3 shows our method achieve a better performance over SLC. Moreover, comparing the results of the proposed algorithm in Table 2 and Table 3, we can see that coding using soft assignment results in higher P@n and mAP.

Parameter Analysis. The parameter λ controls the weighting of visual saliency. When $\lambda = 0$, the model is equal to the basic BOW model which does not contain visual saliency information. When λ is high, the cluster features are very similar in visual saliency. When λ is low, we cluster features with few visual saliency information. The effect of the parameter λ is shown in Figure 10. Figure 10 shows that for the cases where $\lambda = 0$ or the value is high, bad performance will be obtained.

Table 2. Comparison results with basic BOW model

Methods	Codebook size	P@n	mAP
Bag-of-Words	200	55.46%	52.82%
	300	57.79%	53.65%
	500	58.22%	56.27%
	1000	64.74%	62.05%
Proposed algorithm	200	72.13%	70.25%
	300	72.43%	70.20%
	500	76.84%	73.37%
	1000	80.02%	77.86%

Table 3. Comparison results with spatial local coding

Method	Codebook size	P@n	mAP
SLC + soft assignment	1000	79.36%	80.23%
	1500	81.28%	80.20%
	2000	80.00%	78.76%
	3000	81.57%	79.00%
Proposed algorithm + soft assignment	1000	82.80%	79.53%
	1500	83.05%	82.64%
	2000	86.22%	84.21%
	3000	84.37%	82.80%

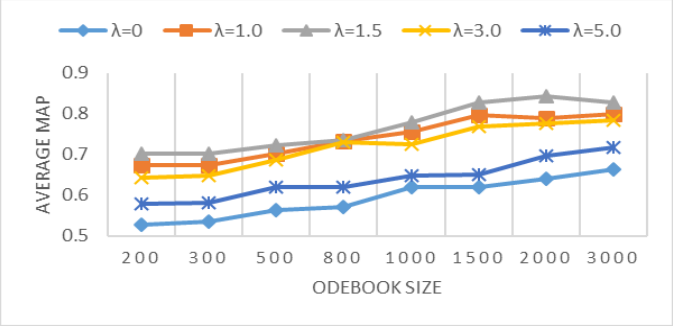


Figure 10. Average mAP with different parameter values

Instance Analysis. Figure 11 shows the image list with "apple" label in the form of a 4*5 grid. For this instance, we first get the top four results by calculating the advertisement relevance based on the proposed representation. Then, we find the display positions for the delivered advertisement based on the relevance. The result is presented in Figure 12. It shows that the recommendation is considerably suitable and the structure of the original page is not disrupted.

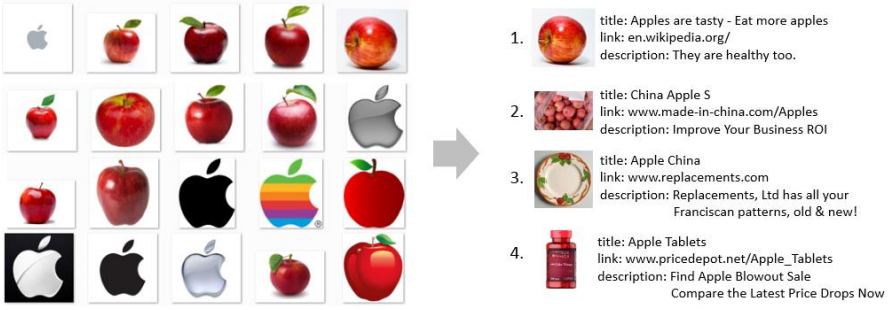


Figure 11. Top 4 recommendations for "apple" label

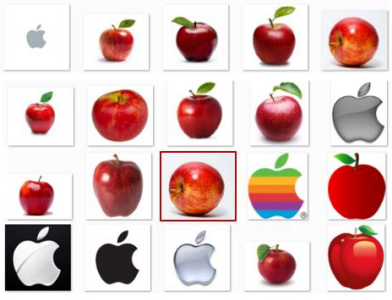


Figure 12. Advertisement delivering position for "apple"

5. Conclusions

In this paper, we design a new image advertising scenario, i.e., delivering image advertisements in a list of images, such as the results of an image search. We suggest delivering the most relevant image advertisements and putting them in suitable places with the purpose of not influencing the users' experience. The similarity between the image advertisement and the image list is calculated based on the proposed multi-cues image representation method. The placement of each image advertisement is also calculated according to the similarities. The experimental results show that the proposed method outperforms the compared traditional methods.

Acknowledgements

This work was supported by the grant of the National Natural Science Foundation of China (No. 61672204), the grant of the Major Science and Technology Project of Anhui Province (No. 17030901026), and the grant of the Key Constructive Discipline Project of Hefei University (No. 2016xk05).

References

1. "Introduction to Computational Advertising," Available at https://web.stanford.edu/class/msande239
2. Ben-Haim, Nadav, B. Babenko, and S. Belongie. "Improving Web-based Image Search via Content based Clustering." Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on IEEE, 2006:106-106.
3. Y. Chen, O. Jin, G. Xue, G. R. Xue, J. Chen, and Q. Yang, (2010). Visual Contextual Advertising: Bringing Textual Advertisements to Images. Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, Usa, July. DBLP.
4. M. M. Cheng, J. Warrell, W. Y. Lin, S. Zheng, V. Vineet, and N. Crook, (2013). Efficient Salient Region Detection with Soft Image Abstraction. IEEE International Conference on Computer Vision (pp.1529-1536). IEEE Computer Society.
5. Csurka, Gabriella, *et al*. "Visual Categorization with Bags of Keypoints."Workshop on the 8th European Conference on Computer Vision, Prague, Czech, (2004) May 11-14
6. S. Goferman, L. Zelnik-Manor, "Context-aware Saliency Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 10, (2012) p. 1915-1926
7. Fergus, Robert, F. F. Li, P. Perona, and A. Zisserman. "Learning Object Categories from Google's Image Search." Proceedings of the Tenth IEEE International Conference on Computer Vision, Beijing, China, (2005) Oct.17-20
8. J. Huang, X. Yang, X. Fang, W. Lin, "Integrating Visual Saliency and Consistency for Re-ranking Image Search Results," IEEE Transactions on Multimedia, vol. 13, no. 4, (2011) p. 653-661
9. F. F. Li, and P. Perona. "A Bayesian Hierarchical Model for Learning Natural Scene Categories." Proceeding of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, (2005) Jun. 20-26
10. D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints."International journal of computer vision, vol. 60, no. 2 (2004) p. 91-110
11. McCann, Sancho, and D. G. Lowe. "Spatially Local Coding for Object Recognition." Proceeding of the 11th Asian Conference on Computer Vision, Daejeon, Japan, (2012) Nov. 05-09
12. M. Tao, L. Li, X. S. Hua, and S. P. Li. "ImageSense: Towards Contextual Image Advertising." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 8, no. 1, (2012), p. 6
13. Muja, Marius, and D. G. Lowe. "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration." Proceeding of the 4th VISAPP, Lisbon, (2009) Feb. 5-8
14. Vedaldi, Andrea, and B. Fulkerson. "VLFeat: An Open and Portable Library of Computer Vision Algorithms." Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, (2010) Oct. 25-29
15. X. J. Wang, M. Yu, L. Zhang, R. C., and W. Y. Ma. "Argo: Intelligent Advertising by Mining a User's Interest from his Photo Collections." Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising, Paris, France, (2009) Jun 28 – July 01
16. W. H. Hsu, L. S. Kennedy, S. Chang, "Reranking Methods for Visual Search," IEEE Transactions on Multimedia, vol.14, no. 3, (2007) p. 14-22
17. P. Xie, Y. L. Pei, Y. A. Xie, and E. P. Xing. "Mining User Interests from Personal Photos." Proceeding of the AAAI, Austin, USA (2015) Jun 25-29
18. Z. Yun, and M. Shah. "Visual Attention Detection in Video Sequences using Spatiotemporal Cues." Proceedings of the 14th ACM international conference on Multimedia, Santa Barbara, CA, USA, (2006) Oct. 23-27

Zhize Wu received a Ph.D. degree from the School of Computer Science and Technology at the University of Science and Technology of China. His current research interests include image processing, neural network, and computational advertising.

Shouhong Wan is currently an associate professor of the School of Computer Science and Technology at the University of Science and Technology of China. Her research interests include big data processing, computer vision, and remote sensing processing.

Ming Tan is currently a professor of the Department of Computer Science and Technology at Hefei University. His research interests include machine learning and computer vision.