Text and Object Detection on Billboards

Tripidok Intasuwan[*] Jakkraphatara Kaewthong[†] Sirion Vittayakorn[‡]
Faculty of Information Technology
King Mongkut's Instituteof Technology Ladkrabang
Bangkok, Thailand, 10520
Email: {57070040*,57070015[†]}@kmitl.ac.th, sirion@it.kmitl.ac.th[‡]

***Abstract*—To captivate people's attention in a blink of an eye, a billboard as a commercial advertisement must be attractive and informative. When space means money like on the billboard, much important information must be dropped out. Unfortunately, insufficient details and the difficulty of accessing those details might easily kill the customer's attention in a flash as well. In this work, we propose a system that automatically connects a billboard with the product website instantly. More specifically, given the billboard image as an input, the system will lead users immediately to the product website for more details, or contact information. The results demonstrate that our system is significantly outperform the Google image search baseline where the product website is always included in the top 10 websites. Finally, We posit that the system can save the customers' time and complexity to access their product of interest, resulting the increasing of product sale.**

***Keywords*—Text detection, Object detection**

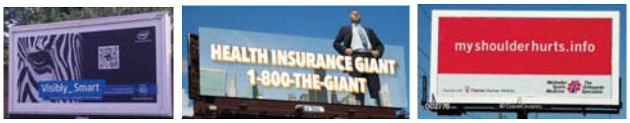
I. Introduction

Make it simple. Make it memorable. Make it inviting to look at - Leo Burnett, an American advertising executive and the founder of Leo Burnett Company, Inc.

Roadside billboard advertisements have been around for years due to their potential to reach massive numbers of people [1]. Table I shows the average rates of renting the standard 14 by 48 billboards for 4 weeks of different sized cities. The results show that by paying $3,500 a week, your advertisement will be displayed to 8 million people in New York city twenty-four seven. Moreover, the weekly impressions have shown that within a week, in a small sized city like Spokane, WA, everyone in town will have seen the advertisement. Although the weekly impression numbers decrease when the city size gets bigger, these numbers are still very high (80.50% for Atlanta, GA and 42.23% for New York, NY).

One goal of the billboard is directing people's attention to the product. However, on average, the driver on the street will have around 5-10 seconds to observe the billboard, read the text, and comprehend the message. Moreover, in those short periods of time, the driver could also be glancing back and forth between the billboard and the road depending on the severity of traffic. That makes them have even less time to spend on the advertisement. Thus, it is crucial to keep the message simple and short, a maximum of 7 words or less [2].

However, a short message cannot convey all the information about the product to the customer. Adding more textual content to the billboard is not only obviously useless, but also diminishes the attractiveness of the billboard. To mitigate this problem, many billboards provide the product website, phone number or QR-code, as shown in Fig. 1, as a link between the billboard and additional information such as details about the product, the price of the product, the location of the store or promotions about the product.



(a) (b) (c)

Fig. 1: An example of a billboard with (a) QR-code, (b) phone number and (c) website to provide additional information.

TABLE I: The average rates for renting a standard 14 by 48 billboard for 4 weeks for small, medium and large sized cities, as provided by Lamar Advertising.

	Cities		
	Spokane,WA	Atlanta, GA	New York, NY
Cost	$1,800	$3,000	$14,000
Adult Population	1,063,000	6,211,000	18,945,000
Weekly Impressions	1,747,267	5,000,000	8,000,000

Although the website, phone number or QR-code can lead the customers to product details that they might be interested in, remembering this additional information while driving or rushing from one place to another might not be practical, resulting in losing valuable potential customers. Unfortunately, the situation might be worse when the customers mistake the product with other brands.

To alleviate the difficulty for customers to access the details of their product of interest, we propose a pipeline that automatically connects a billboard with the potential product website instantly. More specifically, when the billboard image is forwarded through our system, the system will immediately lead you to the product's official website for more detail or contact information, even though this information is not available on the billboard.

The rest of the paper is organized as follows. First, we review related work (Sec II). Then, we describe both the textual and visual datasets that we used in this work (Sec III). Next, we describe our pipeline which extracts product information from both textual and visual content of the billboard (Sec IV) and evaluate our pipeline on different standard datasets (Sec V). Finally, we conclude our work and propose many interesting future directions. (Sec VI).

978-1-5386-4739-4/18/$31.00 ©2018 IEEE

II. Related work

This section reviews previous work relevant to Optical character recognition, object detection, and word representation.

A. Optical character recognition (OCR)

OCR has been utilized to analyze for light and dark areas in order to identify each alphabetic letter or numeric digit on the input image. The process involves photo-scanning of the text, analysis of the scanned-in image, and translates the character image into an editable text format for further processing tasks. One example of the OCR framework is Tesseract [3]. Tesseract uses adaptive thresholding strategies [4] to convert the input image into a binary image. It then utilizes connected component analysis to extract character layouts or blobs which will be forwarded into recognition process.

B. Object detection

Recently, Convolutional Neural Networks (CNNs) have been exploited for object detection tasks in many different ways. For example, Region-based CNN or R-CNN [5] generates region proposals using Selective Search [6] and forwards them into the classifier. To speed up the process [5], Fast R-CNN [7] uses the technique called Region of Interest Pooling during the region proposal process and jointly trains the classifier, and bounding box regressor in a single model. The Single Shot MultiBox Detector (SSD) [8] discretizes the proposed boxes with a set of default boxes over different aspect ratios and scales per-feature map location. The prediction scores of the proposal boxes is based on the similarity of each box to the default box of each category.

C. Word representation

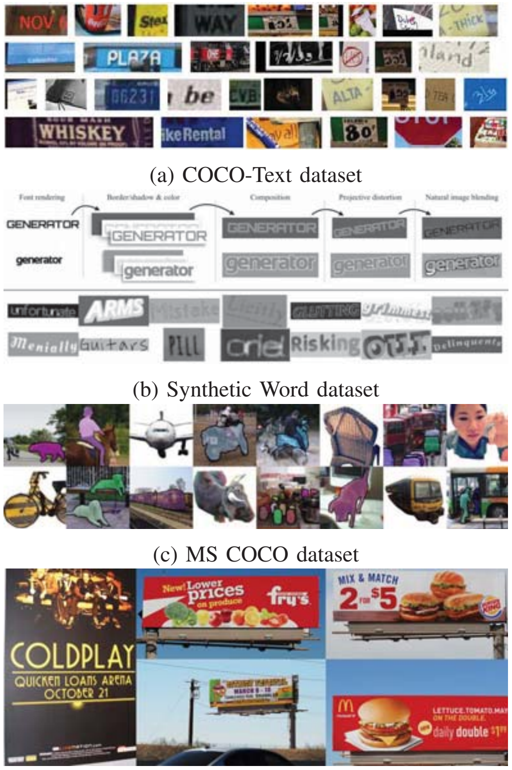
One well-known technique to determine the semantic similarity between two words is Word2Vec [9]–[11]. Word2vec is a computationally efficient predictive model for learning word embeddings from text. It comes in two approaches: the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. While CBOW predicts target words based on the source context words, the skip-gram does the inverse and predicts source context-words from the target words. Generally, CBOW turns to be more efficient for smaller datasets. However, since the skip-gram model treats each context-target pair as a new observation, this tends to do better for larger datasets.

III. Datasets

In this work, we use 4 different datasets including two standard textual datasets: 1) the COCO-Text [12] and 2) Synthetic Word dataset [13] for the text detection and recognition task, the MS COCO dataset [14] for the object detection task and finally the novel Billboard dataset.

A. COCO-Text dataset

The COCO-Text dataset [12] contains 173,589 text annotations from 63,686 images based on the MS COCO dataset [14], which contains images of complex everyday scenes containing a broad variety of text instances. For each text region, five different annotations are provided: (a) location in terms of a bounding box, (b) fine-grained classification into machine printed text and handwritten text, (c) classification into legible and illegible text, (d) script of the text and (e) transcriptions of legible text. The example images of COCO-Text dataset are shown in Fig. 2a.



(a) COCO-Text dataset

(b) Synthetic Word dataset

(c) MS COCO dataset

(d) Billboard dataset

Fig. 2: The example images from (a) COCO-Text, (b) Synthetic Word, (c) MS COCO and (d) Billboard dataset.

B. Synthetic Word dataset

The dataset consists of 9 million images covering 90k English words, and includes training, validation and test splits [13]. Example images from the dataset are shown in Fig. 2b.

C. The Microsoft Common Objects in Context dataset

The Microsoft Common Objects in Context (MS COCO) dataset [14] is a large-scale object detection, segmentation, and captioning dataset. The dataset contains 91 common object categories, resulting in 2,500,000 labeled instances and segmentations in 328,000 images. The example images are shown in Fig. 2c.

D. Billboard dataset

To evaluate our framework, we collected a novel dataset called the Billboard dataset. The dataset consists of 500 billboard and poster images crowdsourced from the Internet. Each image has the corresponding link to the official website of the product on the billboard which will provide additional details about the product. Example images from the novel dataset are shown in Fig. 2d.

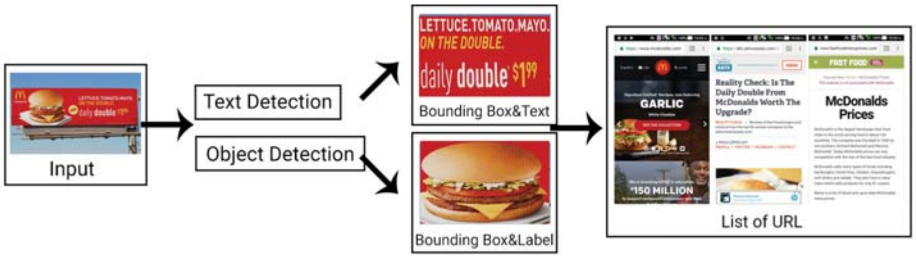


Fig. 3: Given the billboard image as the input, both text detection and object detection pipelines are applied. The results from both pipelines are forwarded into the search engine to find the corresponding websites of the input billboard.

IV. Experiments

To direct the potential customers from the billboard advertisement to the product website in one click, our proposed method starts by forwarding a billboard advertisement image from the camera to the dual pipelines: 1) object detection pipeline, and 2) text detection and recognition pipeline. The object detection pipeline will return the bounding boxes of objects in an image while the text detection and recognition pipeline will output the text labels for an input image. Both visual and textual information from the previous step are then used as a seed for searching for the corresponding websites of the product on the input image. Finally, the top 10 ranked websites, based on both text and object detection confidence, will be shown to the user as the output from our framework. The system overview of our framework is shown in Fig. 3.

A. *Object detection*

To detect objects in an input image, our framework exploits the detection convolutional neural network (CNN) via the Cloud Vision API [15] for this task. The network is based on two different well-known object detection structures: Fast R-CNN [7] and SSD [8]. Given the input image, the network will return the bounding boxes of all objects in the input image with their corresponding detection confidence scores.

B. *Text detection and recognition*

To extract the textual information from the input image, our framework applies the text detection and recognition network called Tesseract OCR [3] to the billboard image. The open-source OCR engines will return the bounding boxes of words in the billboard, the label of the word in the bounding box which is the recognition result, together with the confidence scores.

C. *Visual and textual search*

Given both the visual and textual information from the dual pipelines, we then use both pieces of information as a seeds to search for the product websites.

Visual search: The system starts from cropping out the object patches based on the detection bounding box. The patch with the highest detection score is then forwarded to the the image search (Google Custom Search API [16]) resulting in the relevant website ranking which is purely based on visual information. Only the top 10 websites will be considered as potential hits. After that, both the object detection confidence c_o and the ranking position score $r_p^v = \frac{(11-p)}{10}$, where p is the p-th position of the website from the top 10 image search ranking, will be taken into account to create the final ranking from the visual pipeline or R_v as follows:

$$R_v = \mu_1 c_o + \mu_2 r_p^v \tag{1}$$

where μ_1 and μ_2 are the weight for the object detection confidence and image search ranking. In our experiment, we use equal weight for both values: $\mu_1 = \mu_2 = 0.5$.

Textual search: Similarly to the visual search, the system forwards the textual label from the text detection and recognition pipeline which has the highest score to the text search (Google Custom Search API [16]). The output here is the relevant website ranking which is solely based on textual information. The top 10 websites will be considered as the potential ones. Next, both the recognition confidence c_t and the ranking position score $r_p^t = \frac{(11-p)}{10}$, where p is the p-th position of the website from the top 10 text search ranking, will be taken into account to create the final ranking from the textual pipeline or R_t as follows:

$$R_t = \mu_1 c_t + \mu_2 r_p^t \tag{2}$$

where here μ_1 and μ_2 are the weights for the text recognition confidence and text search ranking. Similarly to the previous pipeline, we use equal weight for both values: $\mu_1 = \mu_2 = 0.5$.

Finally, we incorporate both visual and textual information from the billboard. The final ranking, R_{final}, is the weighted linear combination of both visual and textual search:

$$R_{final} = w_1 R_v + w_2 R_t \tag{3}$$

where w_1 and w_2 are the weights for visual and textual based rankings which we defines as $w_1 = 0.6$ and $w_2 = 0.4$.

V. Experimental results

To evaluate the proposed framework, we explore several evaluation tasks including 1) object detection evaluation, 2) text detection and recognition evaluation and 3) visual and textual search evaluation from various sources.

TABLE II: The average precision (AP) of the object detection pipeline on 4 image categories from MS COCO dataset [14].

Category	Average Precision (AP)
Cat	0.70
Orange	0.80
Television	0.75
Teddy bear	0.68
Mean	0.73

TABLE III: The similarity scores between text recognition results and ground truth across 5 different image categories.

Category	Average Precision (AP)	Similarity score
Handwritten [12]	0.67	0.61
Synthetic word [13]	0.72	0.73
Good clarity	0.70	0.66
Poor clarity	0.50	0.49
Billboard	0.65	0.62
Mean	0.65	0.62

A. *Object detection evaluation*

In this task, we exploit the MS COCO dataset [14] to evaluate our object detection pipeline. In our experiment we sample a total of 1,000 images (250 images per category) from 4 different categories including cat, orange, television and teddy bear. The object bounding boxes with an intersection over Union (IoU) score [17] higher than 0.5 are considered as positive samples, while others are negative samples. The average precision (AP) of the object detection pipeline is shown in Table II. The results show that the performance of our object detection pipeline is reliable with a mean AP of 0.73 across all categories.

B. *Text detection and recognition evaluation*

To evaluate our text detection and recognition pipeline, we used 3,500 images from 5 different textual image categories from 3 datasets as follows:

Handwritten images: One thousand images of handwritten text were included from the COCO-Text dataset [12]. We believe the large diversity of handwritten texts makes this category challenging.

Synthetic word images: Since computer generated text is easier to read, it is widely used across many billboards. Another thousand images of computer generated text images from the Synthetic Word dataset [13] are included.

Good clarity images: In order to investigate the effect of text clarity on the text detection and recognition pipeline, 500 images of clear text from COCO-Text [12] are included.

Poor clarity images: Comparing with the previous subset, we also included 500 images of poor clarity text from COCO-Text [12] as well.

Billboard images: To evaluate our pipeline on the real world images: billboard images with complex backgrounds and being far away in distance, we finally included 500 images from our novel Billboard dataset into the test set.

Using the evaluation set from a variety of sources, we first computed the AP for the text detection task where the bounding boxes with the IoU score [17] higher than 0.5 are considered as positive samples, others are considered negative samples. The AP of the text detection task on 5 different image categories are shown in the second column of Table III. The results demonstrate that the detection framework performs well on both computer generated and good clarity text images with the AP of 0.71 ± 0.01 ; these are relatively easy categories compared to others. When the backgrounds are more complex or the objects are small (e.g., billboard images), or there are the inconsistencies of the character appearance (e.g., handwritten images), the AP decreases to 0.66 ± 0.01. The pipeline struggles the most when the clarity of the text is low, and images have poor clarity, where the AP drops to 0.5. However, the performance of the text detection is still reasonable with the mean AP of 0.65.

Moreover, we evaluate the text recognition task as follows: 1) converting the textual labels into Word2Vec vectors [9]–[11] and 2) computing the L2-distance between the recognition result and the ground truth label. The average similarity of the recognition results and the ground truth labels from 5 image categories are shown in the third column of Table III. The scores demonstrate that although the low clarity of the text affects the performance of the recognition pipeline, with average similarity score of 0.49, the pipeline works well in most of the categories including the challenging ones (e.g., Billboard images with complex backgrounds or handwritten images) and achieves a mean similarity score of 0.62.

C. *Visual and textual search evaluation*

Finally, we evaluate the final ranking of our framework compared to a Google image search baseline using 500 images from the Billboard dataset as shown in Fig. 4. The results from our framework show that 50.8% of the ground truth websites appear at the second position of the final ranking, and 98% of the ground truth websites appear in the top 5 of the final ranking. Although we notice that 9.4% of the ground truth websites appear at the first position of the ranking, the rest of the first positions are the Wikipedia pages of the product company which are highly related to the product as well. The same trend appears with the websites on the top positions of the ranking which are highly related to the product e.g., unofficial websites of the product, articles about the product, unofficial retailers of the product, etc. Moreover, the results show that the ground truth websites always appear in the top 10 of the final ranking (the lowest position of the ground truth website is 9th) where the average position of the ground truth website is 2.58.

Unlike our approach, Google image search [16] is not successful in retrieving the official website of the product using a billboard image as the input. Fig. 4 shows that 97.4% of the ground truth websites appear at more than 80th position of the retrieval ranking and the highest position is 30th. We believe that is because Google image search is designed to retrieve images visually similar to the input. Many times the billboard advertisement and the image advertisement on the official website are totally different; that's why the baseline struggles to retrieve the correct one.

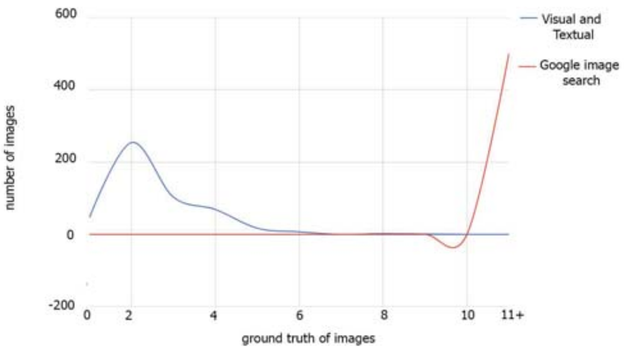


Fig. 4: The histogram of the ground truth website positions from the final ranking of our framework (blue line) compared with Google image search (red line) on Billboard dataset.

An example of qualitative results from our framework are shown in Fig. 5 where the green bounding box indicates the most relevant product website or official product website. The result from the first row shows that the proposed pipeline correctly redirects the user to the correct web page: not just the Burger King's official website but the Burger King's whopper sandwich promotion (the exact sandwich which appears on the billboard). Moreover, other websites in the top rank are also highly related to the product: 3 out of 4 websites are about Burger King's burgers while the last website provides Burger King's coupon. Similarly to the first row, the other results show the same trend where the top 5 not only include the corresponding website but also other websites which are highly related to the product on the input billboard.

VI. Conclusion and Future work

In this work, we propose a system that takes a billboard image as an input and automatically forwards the user to a website ranking which are highly related to the product on the billboard. Our framework, which is implemented as an Android application, utilizes both visual and textual information from the input and redirect potential customers to the product website in a single click. To evaluate our framework, we propose a novel dataset of billboard advertisement images with their corresponding official product website. The results demonstrate that our framework is outstanding. It significantly outperforms the Google image search baseline, where the ground truth websites always appear in the top 10 of the final ranking with the average position of 2.58. We posit that the system will increase sales of the product by automatically redirecting valuable customers to the relevant product website in the blink of an eye.

There are several interesting directions to extend our work including 1) extend the billboard dataset to include more challenging scenarios such as partially occluded billboards or billboard images taken during the night, 2) navigation assistance to the closet store, and 3) a recommendation system. Although the users can access the official website of the product via the billboard advertisement using our current framework, directions to the closest store are still crucial in many situations. For example, hungry users are driving on the highway when they see the billboard advertisement of buy one get one free pizza. Guiding them to the official pizza store where they can order the pizza on-line might not be the optimal solution if the local store is only minutes away from their current location. To handle these situations, navigation assistance that can guide users from their current location to the closest store is required. Finally, we are interested in expanding our framework with a recommendation system based on the user's product of interest. For instance, given the billboard image of Toyota Yaris, we know that the user is interested in an economy car. The recommendation system should provide a list of economy car websites from different brands which have similar features to Toyota Yaris e.g., color, displacement, price, seating capacity, production year, etc.

References

[1] M. Aland, "How much does billboard advertising cost?" accessed 2018-02-11. [Online]. Available: https://fitsmallbusiness.com/how-much-does-billboard-advertising-cost/

[2] H. Morones, "5 rules of effective billboard design and advertising," accessed 2018-02-11. [Online]. Available: https://99designs.com/blog/tips/billboard-design-tips/

[3] R. Smith, "An overview of the tesseract ocr engine," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ser. ICDAR '07, 2007, pp. 629–633.

[4] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *Journal of graphics tools*, vol. 12, no. 2, pp. 13–21, 2007.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*. IEEE, 2014, pp. 580–587.

[6] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[7] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*. Springer, 2016, pp. 21–37.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 3111–3119.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*.

[11] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 746–751.

[12] A. Veit, T. Matera, L. Neumann, J. Matas, and S. J. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*.

[13] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[15] "Cloud vision api," accessed 2018-04-05. [Online]. Available: https://cloud.google.com/vision/

[16] "Google custom search," accessed 2018-04-05. [Online]. Available: https://developers.google.com/custom-search/

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

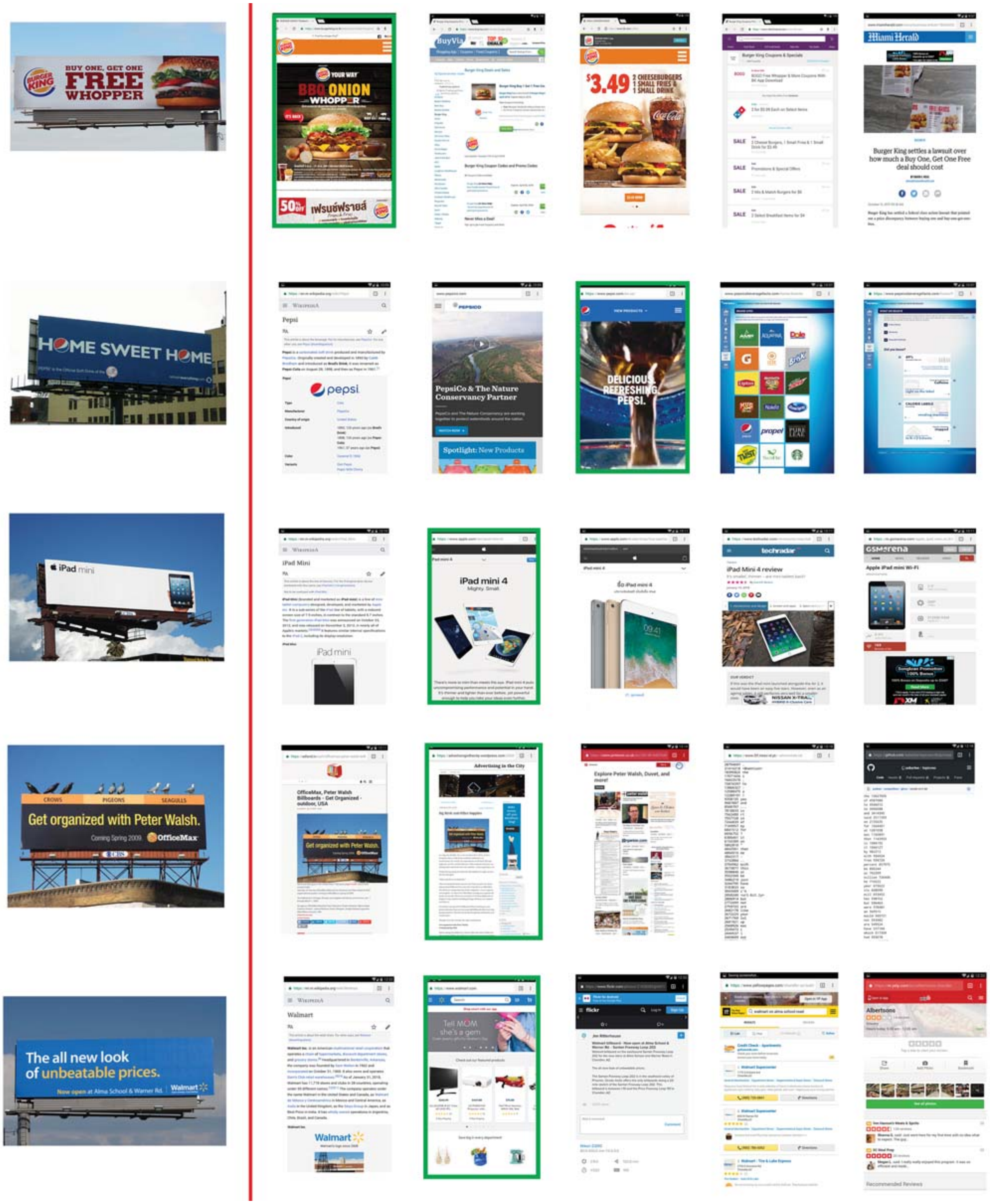


Fig. 5: Given the billboard image on the leftmost column as the input, the top 5 (left to right) corresponding product websites as shown on the right where the green bounding box indicates the most relevant product website or official product website.