Advertisement Detection, Segmentation, and Classification for Newspaper Images and Website Snapshots

Wei-Ta Chu
National Chung Cheng University
Chiayi, Taiwan
wtchu@ccu.edu.tw

Han-Yuan Chang
National Chung Cheng University
Chiayi, Taiwan
chyuan101m@cs.ccu.edu.tw

***Abstract*—Advertisement plays an important role in the human society. Advertisement studies are related to many important issues in economics, social science, and marketing. In this paper, we propose a system to detect, segment, and classify advertisements from newspaper images and website snapshots, in order to facilitate advertisement studies. First, we detect advertisement candidates based on a connected components method. We then design rule-based filters and learning-based filters to remove non-advertisement candidates. From the remained advertisement candidates, we extract visual features and construct classifiers to classify them into predefined advertisement categories. Based on the advertisement categories published over years, we uncover several interesting statistics derived from newspaper front pages and website snapshots.**

***Keywords*-advertisement; image segmentation; image classification; statistics;**

I. Introduction

In the modern ages, advertisements are widely used to actively persuade customers to buy products or service. There is thus a deep relationship between advertisements and the human society. There are three main viewpoints to describe such relationship [1]. The first one is that advertisements create demands. Some say that advertisements create a lot of useless demands, but the opposite side says that these demands already exist but people didn't know before. The second viewpoint is that advertisements affect people's social behaviors, while the opposite side thinks that such affection is weak. The third viewpoint is that advertisements cause overspending. But some researchers think that advertisements only reflect reality of the society.

From these arguments, we realize that there are complex relationships between advertisements and the human society. Advertisement studies thus play an important role to realize macro human behaviors. However, tremendous amounts of advertisements in various forms impede efficient advertisement studies. In this work, we develop a visual analysis system that can be used to analyze large-scale advertisements and thus facilitate advertisement studies. We propose a framework to conduct advertisement detection, segmentation, and classification. Focusing on newspaper images and website snapshots, we segment images and detect advertisement regions, followed by classifying them into one of the predefined advertisement classes. Based on classification results, we unveil several interesting statistics linking advertisements and real-world events in the human society.

Contributions of this paper are summarized as follows.

- We propose a framework for analyzing advertisements in newspaper images and website snapshots. We first segment and extract advertisement candidates, and then classify advertisements based on visual analysis.
- We show several statistics of advertisement classification results, and find some interesting trends that may be utilized in many potential applications.

The rest of this paper is organized as follows. Related works will be surveyed in Sec. II. In Sec. III, we introduce the collected dataset, and provide details of advertisement detection, segmentation, and classification. Experimental results corresponding to each part of the framework will also be shown. In Sec. IV, we present statistics of advertisements, and connect these results with real-world events. Conclusion and future works will be given in Sec. V.

II. Related Works

A. Newspaper Layout Segmentation

Gatos et al. [3] extracted image components like line, image and drawing, and title blocks, and then presented a rule-based approach for article identification. Liu et al. [4] proposed a bottom-up algorithm for newspaper layout analysis. They first detected connected components and then classified them into basic components, line components, text components, or graph components. Components were merged by a heuristic rule considering component attributes. Mitchell and Yan [5][6][7] proposed a series of bottom-up approaches to segment components in newspaper images. Patterns of these components are composed of adjacent rectangular regions. In contrast to rule-based methods, Bansal et al. [8] proposed a novel machine learning framework to learn the structure and layout of newspaper documents. They used the fixed point model proposed in [9] to classify detected blocks and determine the relationship between neighboring blocks.

B. Web Page Segmentation

Hua et al. [10] utilized the Document Object Model (DOM) to build a syntax tree based on HTML tags. They then extracted semantic knowledge from web pages. Chakrabarti et al. [11] formulated webpage segmentation as an optimization problem on weighted graphs, and showed substantial improvement over rule-based or heuristic segmentation methods. Fauzi et al. [12] proposed a webpage segmentation algorithm that extracts web images and their contextual information based on where they locate on webpages. Cai et al. [13] proposed the VIsion-based Page segmentation (VIPs) algorithm that utilized both the DOM tree and visual information in web pages. Song et al. [14] used VIPs to segment web page and identified importance of regions by support vector machines (SVM) and neural network models. Wu et al. [15] used segmentation results of VIPs to evaluate visual quality or aesthetics of web pages.

C. Advertisement Classification

Few works have been done specifically for advertisement classification. Based on text information, Peleato et al. [16] combined naive Bayes classifiers and the information associated with advertisements to classify newspaper advertisements into one of four classes: real estate, vehicles, employment, or others. Yin et al. [17] presented an algorithm automatically categorizing web elements based on random walks. It classified web elements into five functional categories: content, related links, navigation and support, advertisement, and form.

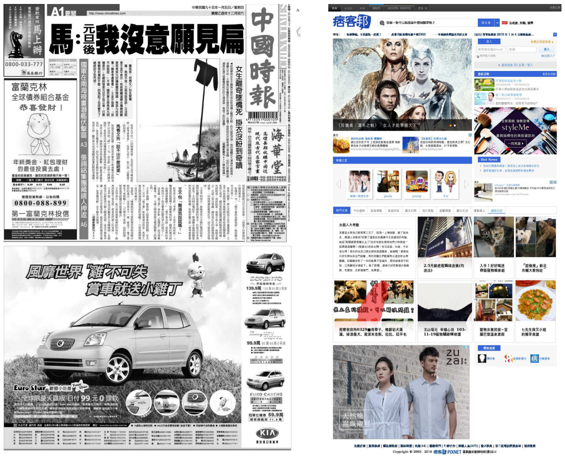
III. ADVERTISEMENT DETECTION, SEGMENTATION, AND CLASSIFICATION

The proposed system consists of three components: advertisement candidate detection, non-advertisement filtering, and advertisement classification. We first use image segmentation techniques to detect advertisement candidates from the source image. Non-advertisement filters are then proposed to remove unlikely candidates. Finally, advertisement regions are classified into one of the predefined classes. Based on classification results, we are able to explore some interesting statistics of advertisements.

A. Databases

Newspaper Image Database. We collect newspaper images from the Document Digitalization Project initiated by National Central Library in Taiwan. We collect the front page of China Times, in the form of scanned images, published from January 16, 2002 to October 30, 2015, and have 5,338 gray-level scanned newspaper images in total. Figure 1(a) shows a sample front page of China Times.

Website Snapshot Database Another important target of our work is website snapshots, because many advertisements are published on portal websites. We collect snapshots of three portal websites: Yahoo! News, ETtoday, and PIXNET, which are three most popular websites in Taiwan. We catch website snapshots every six minutes from April 9, 2016 to June 30, 2016. We totally collect 59,310 website snapshots. Figure 1(b) shows one website snapshot from PIXNET.



(a) China Times (b) PIXNET

Figure 1. (a) A sample front page from China Times; (b) a sample snapshot from PIXNET.

B. Advertisement Detection

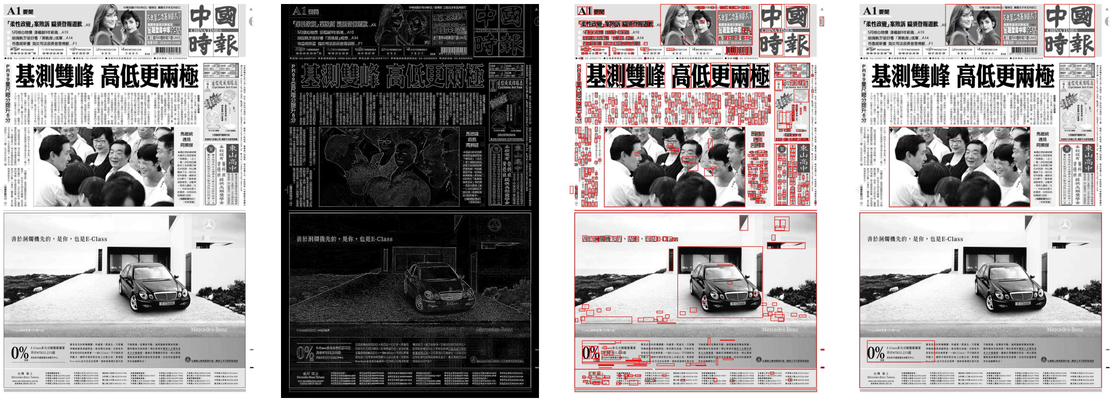
Newspaper images in the database are scanned images, and we thus first employ histogram equalization to ease deviations caused by different scanners. The Laplacian of Gaussian filter is used to detect and enhance edges in the image, as shown in Figure 2(b). We then use the Otsu's method to binarize the edge image, and find 8-connectivity connected components from the binary map, as shown in Figure 2(c). As can be seen, many connected components shown in Figure 2(c) are not advertisements.

Two types of filters are designed to remove non-advertisement regions. First, because the press charges fee according to location and area of advertisement to be released, advertisements would be displayed in specific sizes. We thus check whether both height and width of each candidate region meet the following criteria:

$$CC_w > \alpha \times I_w, \quad \text{and} \quad CC_h > \alpha \times I_h, \tag{1}$$

where CC_w and CC_h are respectively the width and height of a connected component, and I_w and I_h are respectively the width and height of the original image (front page of the newspaper). We set the ratio α as 0.08 according to our preliminary statistics.

Another rule comes from that most advertisements are well bounded, i.e., they usually have at least one or two boundary edges. We thus apply the Sobel operator to detect edges of an advertisement candidate, and project edge pixels onto the vertical axis and the horizontal axis, respectively. From the horizontal profile P_h and vertical profile P_v, we



(a) Original image (b) Edge detection (c) Connected components (d) After rule-based filtering

Figure 2. Sample results of the advertisement candidate detection process.

check if there is any peak in them. For example, peaks in the horizontal projection peaks are detected by checking

$$\begin{cases} P_h[i] > P_h[i-1], & \text{and} \\ P_h[i] > P_h[i+1], & \text{and} \\ P_h[i] > 0.8 \times CC_w, \end{cases} \tag{2}$$

where $P_h[i]$ is the projection value of the ith row of the horizontal projection. We especially check the rows in the ranges $[1, \frac{CC_h}{5}]$ and $[\frac{4 \times CC_h}{5}, CC_h]$ for that they are close to region borders. If no peak is found, this connected component is not viewed as an advertisement region. A sample result after the aforementioned rule-based filtering is shown in Figure 2(d).

To further filter out non-advertisement candidates, we extract convolutional neural network (CNN) features to describe candidate regions, and construct a support vector machine (SVM) classifier to discriminate advertisements and non-advertisements. We extract CNN features by the MatConvNet toolbox [18] with the vgg-f per-trained model [19]. Results of the sixth layer, i.e., the first fully connected layer, are taken as the image region representation, which is 9,216-dimensional. We then utilize the libSVM package [20] to train the SVM model. In our preliminary experiments, the classification accuracy is around 96.60%.

C. Advertisement Classification on Newspaper Images

We define seven advertisement categories for newspaper images, according to the business categories table defined by Taiwan government. They are *manufacturing*, *whole sale and retail trade*, *finance and insurance*, *real estate*, *education*, *announcement*, and *politics*. In addition to extracting CNN features $\boldsymbol{v}_c$ to describe visual information, we also utilize the VIREO374 package [21] to detect semantic concepts embedded in advertisement regions. The VIREO374 concept detectors estimate the confidence of each of 314 common semantic concepts from the advertisement region, and form a 374-dimensional semantic vectors $\boldsymbol{v}_s$. The vectors $\boldsymbol{v}_c$ and $\boldsymbol{v}_s$ are respectively used to construct a multiclass SVM classifier to classify any given advertisement region into one of the seven categories. For a given advertisement region, the final classification result i^* is determined by

Table I
AVERAGE CLASSIFICATION ACCURACY BASED ON DIFFERENT FEATURES.

Features	Accuracy
Only CNN	85.54%
Only semantics	84.65%
Both	94.04%

$$i^* = \arg\max_i (p_i^c(\boldsymbol{v}_c) + p_i^s(\boldsymbol{v}_s)), \tag{3}$$

where $p_i^c(\boldsymbol{v}_c)$ and $p_i^s(\boldsymbol{v}_s)$ are the probability of the ith category estimated by the SVM built based on CNN features and semantic features, respectively.

We evaluate the classification process based on the five-fold cross validation scheme. Table I shows the average classification accuracy based on different features, which is quite promising, i.e., over 90% accuracy can be achieved if both features are used. Table II shows the confusion matrix of advertisement classification, with each entry showing the number of advertisement regions classified into one category. It can be seen that we can obtain very good classification results for most categories, except for politics advertisements. The reason is that the number of politics advertisements is much fewer than others, and we don't have enough data for training.

D. Advertisement Classification on Website Snapshots

We use the same framework and setting as newspaper to do advertisement classification for website snapshots, except for advertisement categories. Because there are many manufacturing advertisements on the web, we further divide it into *clothes*, *vehicle*, *food and drink*, *medical beauty products*, *3C products*, and *appliances*. We finally have

Table II
THE CONFUSION MATRIX OF ADVERTISEMENT CLASSIFICATION, BASED ON BOTH FEATURES.

	MA	WR	FI	RE	ED	AN	PO
MA	58	0	2	0	1	1	0
WR	0	69	4	1	0	1	0
FI	2	1	265	1	0	0	0
RE	0	0	3	54	0	0	1
ED	0	0	2	0	57	4	0
AN	1	0	2	1	1	115	1
PO	1	0	5	0	0	4	7

thirteen advertisement categories, including the six manufacturing advertisements, *wholesale and retail trade*, *travel transportation*, *information communication*, *game industry*, *finance and insurance*, *real estate*, and *education*. Also based on the five-fold cross validation scheme, the average classification accuracy is 97.07%. We omit the classification confusion matrix due to space limitation.

IV. ADVERTISEMENT ANALYSIS

A. Statistics of Newspaper Advertisements

We detect and classify China Times's advertisements published from January 16, 2002 to December 31, 2014, which totally consist of 4983 images. Figure 3(a) shows the pie chart of advertisement amounts of each category. According to this pie chart, we can realize that the main customer of newspaper front page advertisements is finance and insurance. Finance and insurance is one of the most thriving industries in Taiwan, and there are many announcement advertisements because front page is the most notable place in newspapers. There are much fewer politics advertisements because they only appear in the period of election.

In newspaper front pages, advertisements can be roughly categorized into half-page advertisements and smaller-area advertisements. Figure 3(b) shows the ratios of half-page advertisements in each category, and Figure 3(c) shows the ratios of small area advertisements. We can see that manufacturing and real estate prefer to publish their advertisement in half pages. Whole sale and retail trade, and education prefer to publish advertisements using smaller areas. Manufacturing and real estate need to provide more information about their products, and thus they need lager areas to convey more text and fine graphs. Although whole sale and retail trade have many products, the main of their advertisements is disseminating special offer messages to attract people entering stores. The education advertisements are usually used to announce student recruiting messages, and large-size advertisements are not needed.

We then explore the relationship between advertisements and time. As can be seen in Figure 4, the total advertisement amount sharply decreases from 2008. This may be due to that the worldwide financial crisis occurred in 2007 and 2008. This crisis was out of control in September, 2008. The second reason may be the fading of newspaper publication.

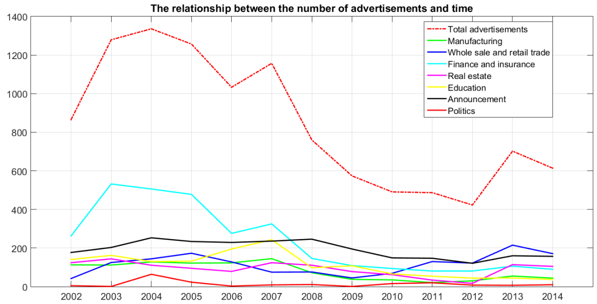
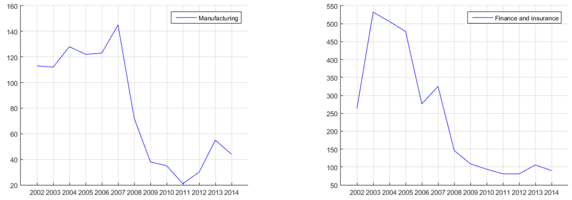


Figure 4. The relationship between the number of advertisements and time.



(a) Manufacturing

(b) Finance and Insurance

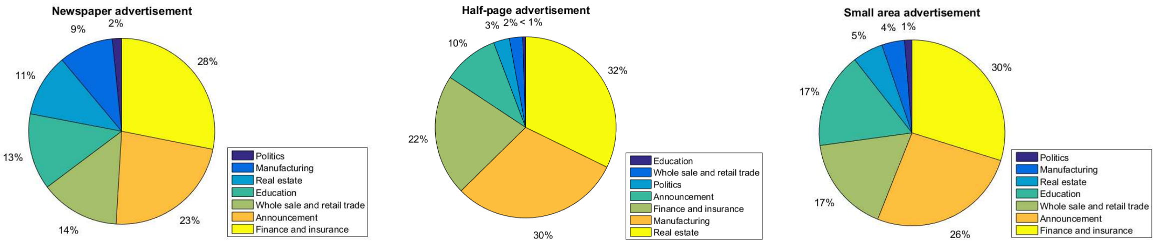
Figure 5. The relationship between the number of manufacturing and finance advertisements and time, respectively.

Because of the growth of internet, more and more people read news and search information from internet rather than newspaper, and this causes recession of newspaper as well as advertisements.

More particularly, Figure 5(a) shows the number of manufacturing advertisements decreases quickly in 2008 because of the financial crisis. And then in 2011, Taiwan food scandal causes another drop. In Figure 5(b), the number of finance and insurance advertisements reflects the situation of financial crisis and the credit/cash card crisis in Taiwan. Credit/cash card crisis come from loose management of the application processes in 2000. The amounts of cards application increased quickly from 2001 to 2004. Because too many people, who can't afford the fee, apply cards, the crisis began in 2005. Finally, Taiwan's government and legislature interfered and solved these problems in 2006. Although finance and insurance recovered a little bit in 2007, the financial crisis caused another recession.

B. Statistics of Website Advertisements

We detect and classify website snapshot advertisements coming from three websites, i.e., Yahoo! News, ETtoday, and PIXNET, from April 9, 2016 to June 13, 2016. The pie charts shown in Figure 6 visually show the numbers of regions of each advertisement category in website snapshots. We can see that the ratios of website advertisements are obviously different in different websites. In ETtoday, the main advertisement categories are travel transportation, medical beauty products, information communication, and



(a) Overall

(b) Half-page advertisement

(c) Smaller-area advertisement

Figure 3. The pie chart showing the numbers of regions of each advertisement category.

game industry. They are relatively more with entertainment. In PIXNET, more than half amounts of advertisements are medical beauty products, and the second one is clothes. The advertisements in PIXNET are more relatively with manufacturing products. In Yahoo! News, the amounts of different categories are nearly equal, except for finance and insurance and education. Comparing to ETtoday and PIXNET, Yahoo! News advertisements are more diverse. This may be because Yahoo! is a general portal covering a wide range of topics.

We then explore the relationship between advertisement and time. As shown in Figure 7(a), there are some relationships between time and the number of medical beauty advertisements. In 2016, Mother's Day is in May 8. In ETtoday, the amount of medical beauty advertisements increases quickly during this period. The increasing is not so apparent in Yahoo! News and PIXNET. In PIXNET medical beauty advertisements commonly appear all the time (Figure 6(c)), and thus the increasing is not very apparent.

We also can check characteristics of advertisements published on a day. An interesting statistics can be found in Figure 7(b). The amount of game advertisements sharply decreases before dawn, and recovers after the noon. The amount reaches to a peak at midnight.

V. Conclusion

We have presented an advertisement detection, segmentation, and classification framework to facilitate advertisement studies in newspaper images and website snapshots. We classify advertisement based on visual analysis that attracted little attention before. The evaluation results show that the proposed method performs well for both newspaper images and website snapshots. After analyzing statistical results of newspaper advertisements and website advertisements, we find several interesting characteristics between advertisements and business policies.

We face some limitations needed to be solved in the future. One of them is the classification accuracy of politics advertisements. Much more data should be collected to improve this part.

Acknowledgement. The work was partially supported by the Ministry of Science and Technology in Taiwan under the grant MOST103-2221-E-194-027-MY3, MOST104-2221-E-194-014, and MOST105-2628-E-194-001-MY2.

References

[1] S. Moriarty, N. Mitchell, and W. D. Wells, "Advertising & imc: Principles and practice," *Pearson*, 2014.

[2] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519, 2014.

[3] B. Gatos, S. Mantzaris, K. Chandrinos, A. Tsigris, and S. Perantonis, "Integrated algorithms for newspaper page decomposition and article tracking," *Proceedings of International Conference on Document Analysis and Recognition*, pp. 559–562, 1999.

[4] F. Liu, Y. Luo, M. Yoshikawa, and D. Hu, "A new component based algorithm for newspaper layout analysis," *Proceedings of International Conference on Document Analysis and Recognition*, pp. 1176–1180, 2001.

[5] P. Mitchell and H. Yan, "Newspaper document analysis featuring connected line segmentation," *Proceedings of International Conference on Document Analysis and Recognition*, pp. 1181–1185, 2001.

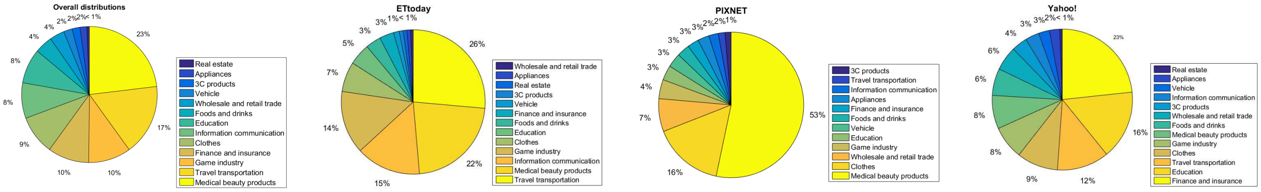
[6] ——, "Newspaper layout analysis incorporating connected component separation," *Image and Vision Computing*, vol. 22, no. 4, pp. 301–317, 2004.

[7] ——, "Connected pattern segmentation and title grouping in newspaper images," *Proceedings of International Conference on Pattern Recognition*, pp. 397–400, 2004.

[8] A. Bansal, S. Chaudhury, S. Roy, and J. B. Srivastava, "Newspaper article extraction using hierarchical fixed point model," *Proceedings of IAPR Workshop on Document Analysis Systems*, 2014.

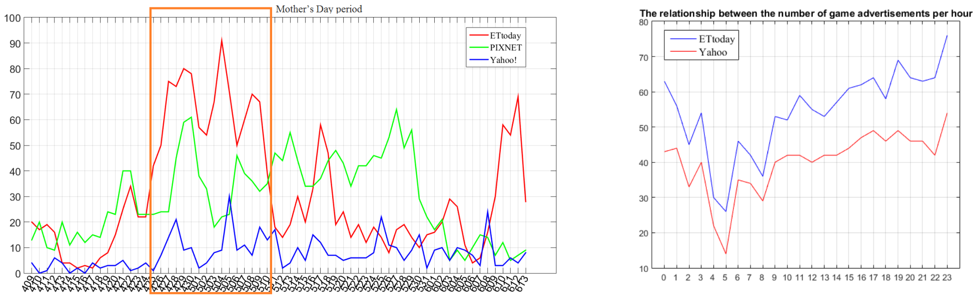
[9] Q. Li, J. Wang, D. Wipf, and Z. Tu, "Fixed-point model for structured labeling," *Proceedings of International Conference on Machine Learning*, 2013.

[10] Z. Hua, X.-J. Wang, Q. Liu, and H. Lu, "Semantic knowledge extraction and annotation for web images," *Proceedings of ACM International Conference on Multimedia*, pp. 467–470, 2005.



(a) Overall (b) ETToday (c) PIXNET (d) Yahoo!

Figure 6. The pie chart showing the numbers of regions of each advertisement category.



(a) Medical Beauty (b) Online Game

Figure 7. The relationship between the number of web advertisements and time.

[11] D. Chakrabarti, R. Kumar, , and K. Punera, "A graph-theoretic approach to webpage segmentation," *Proceedings of International Conference on World Wide Web*, pp. 377–386, 2008.

[12] F. Fauzi, J.-L. Hong, and M. Belkhatir, "Webpage segmentation for extracting images and their surrounding contextual information," *Proceedings of ACM International Conference on Multimedia*, pp. 649–652, 2009.

[13] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Vips: A vision based page segmentation algorithm," *Technical Report MSR-TR-2003-79, Microsoft Research*, 2003.

[14] R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma, "Learning block importance models for web pages," *Proceedings of International Conference on World Wide Web*, pp. 203–211, 2004.

[15] O. Wu, Y. Chen, B. Li, and W. Hu, "Evaluating the visual quality of web pages using a computational aesthetic approach," *Proceedings of ACM International Conference on Web Search and Data Mining*, pp. 337–346, 2011.

[16] R. Peleato, J.-C. Chappelier, and M. Rajman, "Using information extraction to classify newspapers advertisements," *Proceedings of International Conference on the Statistical Analysis of Textual Data*, 2000.

[17] X. Yin and W. S. Lee, "Understanding the function of web elements for mobile content delivery using random walk models," *Proceedings of International Conference on World Wide Web*, pp. 1150–1151, 2005.

[18] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," *Proceedings of ACM International Conference on Multimedia*, pp. 689–692, 2015.

[19] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *Proceedings of British Machine Vision Conference*, 2014.

[20] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.

[21] Y. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.