Worm-like image descriptor for signboard classification

Aleksei Samarin
VK Research
Saint-Petersburg State University
Saint-Petersburg, Russia
Email: aleksei.samarin@vk.com

Valentin Malykh
Kazan Federal University
Kazan, Russia
Email: valentin.malykh@phystech.edu

***Abstract*—We introduce a special image descriptor that is well suited for classification of images containing various inscriptions. In order to demonstrate effectiveness of the proposed solution we provide evaluation of a system based on the introduced descriptor on commercial building facade photographs grouping problem according to the type of services provided. Our system achieved state of the art performance (0.28 in averaged F_1) over classical CNN-based methods and a composite baseline.**

***Index Terms*—image descriptor, image classification, signboard recognition, visual characteristics**

I. Introduction

Currently, in the field of applied marketing, problems related with advertising signs recognition are urgent [10], [16], [17]. One of these issues is the problem of photographs of advertising posters classification by type of a provided services. The problem is quite difficult because of unique fonts and colors, and different label sizes, and also various shooting conditions. A decision on whether a photograph of an advertising sign belongs to one or another category can be obtained on the basis of both textual information located on this sign and pure visual features extracted from the photograph [10]. Up to this day, a lot of methods have demonstrated their effectiveness in solving the problem of general image classification [5], [6], [13], although classification of general objects photographs is quite different from the classification of photographs of advertising posters. One important feature of advertising signs which differs heterogeneous objects (like the ones depicted in [3], [8]) and significantly complicates their classification is the absence of the convex elements. This feature leads to lower efficiency of the heterogeneous images classification methods to distribute signboard photographs by groups [10]. Another important feature of photographs of advertising posters is presence of a textual information. In some cases, the text shown on the signboard may contain information of key importance for classifying an image. There are a lot of document image classification methods that are based on the prior text recognition [11], [14], [18]. Such methods demonstrate high effectiveness for scanned document classification. However, such factors as the possible lack of sufficient information in the text of the advertising sign and the difficulty of solving the problem of optical text recognition with variable angles, fonts, styles of signage and lighting (Fig. 1), that is typical for photographs of advertising signs makes pure extracted text based signboard photograph classification approaches insufficient. We should also mention another type of methods that use combined classifiers that retrieve textual information as well as pure visual features in order to achieve better performance in signboard photographs classification, e.g. [10]. These methods also suffer from poor OCR quality. In order to improve such deficiency, we developed a new solution avoiding explicit text retrieval and replacing it with special visual features extraction from image patches with text.

We propose a neural network method based on the extraction of several types of general visual features and the special image descriptor analysis. This method shows better efficiency than methods that use only visual information or based only on the analysis of the text recognized during photo processing and also combined methods that uses visual features and explicit text information retrieval [10].

II. Problem Statement

In this work we investigate a special image descriptor effectiveness for the problem of advertising sign photograph classification by the type of provided services. The problem can be formulated as follows. An input photograph containing signboard Q should be assigned to one of the classes $C = C_i$, where $i \in [0, N]$.

In addition to the formal statement of the problem we use the following restrictions. Images are captured by a camera fixed on a car, following along the roadway [15], hence: a) may contain visual defects - sun glare, noise, including those that greatly impede optical text recognition; b) angle, framing, lighting and colour balance are unknown and can vary significantly from shot to shot; c) the relative size and placement of signboard in snapshot can also vary greatly (Fig. 1).

III. Proposed method

The proposed system contains several modules: a visual features extraction module, text detection one, and special text-containing image descriptor module. The general architecture of our solution is presented on Fig. 2.

The proposed scheme contains a module of visual features extraction. It is CNN-based, since such image features extractors is effective in solving problems of classifying im-

a) b) c) d)

Figure 1. *Considered dataset illustration: a) photograph of a hotel; b) photograph of a store facade; c) image of a restaurant signboard; d) photograph of a signboard that does not belong to categories listed above.)*

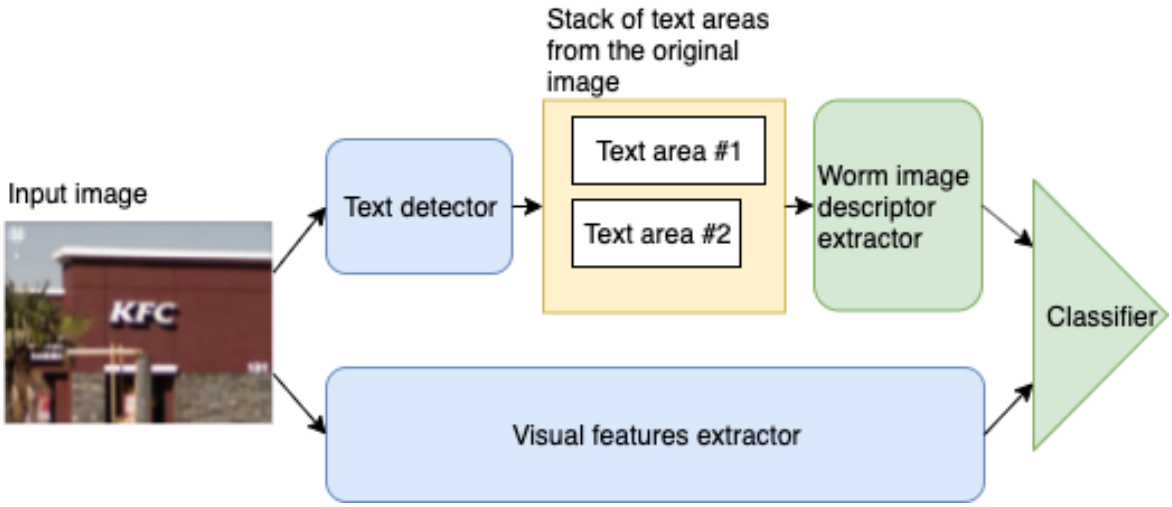


Figure 2. General architectural diagram of the combined advertising signboards classifier engine.

ages of heterogeneous objects [5], [6], [13]. However classifiers that are based only on CNN extracted features do not achieve a great performance in signboard photographs classification [10]. The reason for this phenomenon is the specificity of advertising posters in terms of general image features. In order to improve efficiency of our solution we introduce the additional features type that is obtained from areas of an input image that contain text. We call it *worm-like descriptor*. The output of the post-processing result of evaluated *worm-like descriptor* is concatenated with CNN features obtained from the whole original image. The result of the concatenation is projected onto a space of dimension 4 (according to the number of classes). Then we apply SoftMax function to the obtained vector and interpret result values as the probabilities of target classes.

A. General image features extractor

Following [10], we use MobileNet [5] as general image descriptor. The features are extracted from a whole input image in order to retrieve significant information from background which helps to establish the type of services provided. MobileNet model itself is a sequence of convolutional, fully connected layers and residual connections [9].

B. Text detector

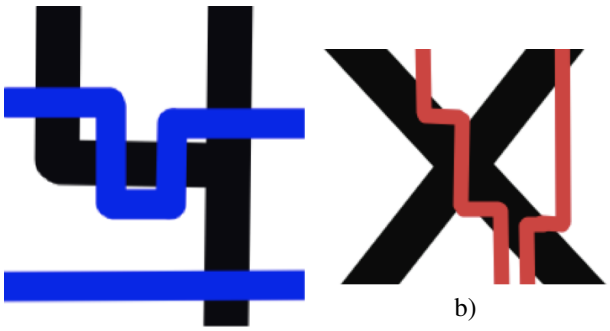
Following [10], we use EAST text detector [18] to localize the text position of the scene space is a fast multi-channel CNN-based architecture resistant to a varying angle.

C. Worm-like image descriptor

We introduce a special type of descriptor for images containing textual information. We aim to develop a method, which could be of low computational complexity and could be applied in parallel.

The most important feature of this descriptor is based on the idea of obtaining the maximal information from the mutual arrangement of regions with the maximum brightness variation. The second feature of poster images used is the repeating nature of the characters. Thus, local differences between the expression of the first and last characters of a word can be described independently and in the same terms. Using these considerations, we construct a picture descriptor as a trace of a certain number of agents (we call them *worms*) moving from given initial positions on the picture in directions that maximize the brightness variance at each step. The sample agent traces presented as Fig. 3.

It should be noted that each *worm* has a predefined movement direction to avoid displacement of the main direction of movement in the direction of contours that are not related to



a) b)

Figure 3. *Samples of worm traces (original sign is showed in black): a) example of horizontal traces (marked with blue color); b) example of vertical traces (marked with red color).*

symbol images (for example, poster borders). Summing up the above, the main component of the descriptor of an image is the trace of an agent:

$$T^v(x_0, y_0) = (m_1^v(x_0, y_0), ..., m_N^v(x_0, y_0)),$$

where $T^v(x_0, y_0)$ - trace of a *worm* with priority movement direction v and initial position (x_0, y_0). $m_i^v(x_0, y_0)$ stands for a movement direction (couble be $\{up, down, left, right\}$) for a step number i with priority direction v and initial position (x_0, y_0). We select each movement type according to the following expression:

$$m_{i+1}^v = \arg\max_{m \in M}(\mathrm{Var}[I[x, y]] + c(m, v)),$$

where

$$(x, y) \in P((x_i, y_i), s), \;\; v \in \{up, down, left, right\},$$

and $m_i^v(x_0, y_0)$ stands for a movement direction for a step number i with priority direction v and initial position (x_0, y_0). N is equal to the number of traced steps.$P((x_i, y_i), s)$ stands for a set of coordinates that can be achieved with one step from position (x_i, y_i) with a step size that is equal to s pixels. And M stands for a set of possible step types ($\{start, finish, up, down, left, right\}$). Thus we evaluate general descriptors for an each movement direction:

$$\begin{aligned} T^{up} &= (T^{up}(x_0, H), ..., T^{up}(x_A, H)), \\ T^{down} &= (T^{down}(x_0, 0), ..., T^{down}(x_A, 0)), \\ T^{left} &= (T^{left}(W, y_0), ..., T^{left}(W, y_B)), \\ T^{right} &= (T^{right}(0, y_0), ..., T^{right}(0, y_B)), \end{aligned}$$

where A and B stands for horizontal and vertical *worms*) number, W and H denotes an input image width and height. Finally we merge descriptors from each direction into result image descriptor that can be described as follows:

$$T = (T^{up}, T^{down}, T^{left}, T^{right}).$$

From the above description, it is easy to construct an algorithm that calculates *worm-like* descriptor for the number of steps $O(w + h)$, where w and h stands for an input image width and height correspondingly, whereas a lot of basic image descriptors implementations (HOG [7] and LBP [12]) requires $O(w * h)$ steps. It should also be noted that the procedure for calculating our descriptor is parallelized with a small effort.

IV. Experiments

A. Dataset

We trained the proposed classifier on a dataset, presented in [15]. The dataset contains 357 advertising signs photographs that are taken using a camera fixed on a car. All of the images were obtained under different lighting conditions and camera angles. Signboards contain textual information decorated with different fonts styles and colors. We also use the additional markup presented in [10]. All photographs from the dataset were split into 4 classes according to the type of services provided (hotels, shops, restaurants, and "other"). All of listed classes contains approximately the same number of samples.

B. Baselines

We compare performance of our solution with a model proposed in [10]. That model is based on a classifier that merges visual and textual features to enrich the image embedding. The main scheme of this baseline is similar to our one, the difference is in usage of an image descriptor, where authors involve OCR to produce a noisy text from the detected text regions, and then embed this text with special character-level vector model.

We also provide results of comparison with other combined methods, where text region descriptor is either LBP [12] or HOG [7] based. That experimental models are based on the same architecture as our one with the only and differ only in the type of used descriptor. We chose these two baselines, since they have proven their effectiveness in image classification problems [1]. Local binary pattern (LBP) descriptor uses a binary string representation for demonstration of the spatial relationship between the local neighboring pixels [2]. Histogram of oriented gradients (HOG) descriptor is based on the histogram of pixel gradients neighbors for image blocks [4].

V. Results

As a quality measure we use F_1 metric following [10]. It is formulated as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP + FP}, \\ Recall &= \frac{TP}{TP + FN}, \\ F_1 &= \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \end{aligned}$$

where TP is a number of objects are correctly marked up by a model to belong to the class; FP is a number of objects that are incorrectly marked up by a model to belong to the class; and FN is a number incorrectly marked up by a model to not

belong to the class. The F_1 metric defined above describes quality only for one class. In order to get the final F_1 metric for all classes we average their scores. Each configuration was trained ten times. The results of comparison with the baselines (F_1 score mean and variation values) are given in Tab. I.

Table I
PERFORMANCE COMPARISON OF INVESTIGATED MODELS

Model	F_1
Malykh & Samarin, 2019 [10]	0.24 ($\pm$0.0023)
Combined classifier + HOG	0.23 ($\pm$0.0007)
Combined classifier + LBP	0.26 ($\pm$0.0011)
Combined classifier + *Worm-like (ours)*	**0.28 ($\pm$0.0010)**

As one can see, LBP-based model shows higher results than previous best model, although the proposed worm-like description model shows even better performance in this task.

CONCLUSION

We propose a special image descriptor for an implicit semantic information extraction from a photograph of signboard. We also show the effectiveness of a combined model configured to use introduced *worm-like* descriptor in the context of advertising sign photograph classification problem. The introduced model has demonstrated better efficiency in comparison to methods based on the use of only visual or combined visual and explicit textual features method. In the problem of signboard photographs classification our model achieves new state of the art result (0.28 in averaged F_1 score against 0.24 of previous best model). In addition to efficiency in solving the problem under consideration, the proposed method is more lightweight than its analogues, and contains modules that are portable to mobile devices. Among the disadvantages of the proposed method we highlight some general heaviness of the modular architecture, that does not allow its usage in real time on mobile devices. Basing on the obtained results, the further research could be focused on the study of more optimal strategies for getting traces of *worm-like* agents and the whole model performance optimization.

REFERENCES

[1] Ashutosh Bachchan, Apurba Gorai, and Phalguni Gupta. Automatic license plate recognition using local binary pattern and histogram matching. pages 22–34, 07 2017.

[2] Ayan Bhunia, Shuvozit Ghose, Partha Roy, and Subrahmanyam Murala. A novel feature descriptor for image retrieval by combining modified color histogram and diagonally symmetric co-occurrence texture pattern. *Pattern Analysis and Applications*, 03 2019.

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[4] Tamarafinide Dittimi and Ching Suen. Modified hog descriptor-based banknote recognition system. *Advances in Science, Technology and Engineering Systems Journal*, 3, 10 2018.

[5] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. 04 2017.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[7] Chunde Huang and Jiaxiang Huang. A fast hog descriptor using lookup table and integral image. 03 2017.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[9] Tianyi Liu, Shuangsang Fang, Yuehui Zhao, Peng Wang, and Jun Zhang. Implementation of training convolutional neural networks. *CoRR*, abs/1506.01195, 2015.

[10] Valentin Malykh and Aleksei Samarin. Combined advertising sign classifier. In *Analysis of Images, Social Networks and Texts*, pages 179–185, Cham, 2019. Springer International Publishing.

[11] R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, Sep. 2007.

[12] Junding Sun, Zhu Shisong, and Wu Xiaosheng. Image retrieval based on an improved cs-lbp descriptor. pages 115 – 117, 05 2010.

[13] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[14] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 56–72, Cham, 2016. Springer International Publishing.

[15] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1457–1464, Washington, DC, USA, 2011. IEEE Computer Society.

[16] Alok Watve and Shamik Sural. Soccer video processing for the detection of advertisement billboards. *Pattern Recogn. Lett.*, 29(7):994–1006, May 2008.

[17] Jiang Zhou, Kevin McGuinness, and Noel E. O'Connor. A text recognition and retrieval system for e-business image management. In Klaus Schoeffmann, Thanarat H. Chalidabhongse, Chong Wah Ngo, Supavadee Aramvith, Noel E. O'Connor, Yo-Sung Ho, Moncef Gabbouj, and Ahmed Elgammal, editors, *MultiMedia Modeling*, pages 23–35, Cham, 2018. Springer International Publishing.

[18] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, July 2017.