Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 1, January 2019, pp. 281-287
DOI: 10.23940/ijpe.19.01.p28.281287

Can Machine Automatically Discover Text Image from Overall Perspective

Wei Jiang[a], Jiayi Wu[a], and Chao Yao[b,*]

[a]*School of Software, North China University of Water Resources and Electric Power, Zhengzhou, 450045, China*
[b]*School of Automation, Northwestern Polytechnic University, Xi'an, 710071, China*

Abstract

Recently, more and more researchers have focused on the problem about how to automatically distinguish text images from non-text ones. Most of previous works have originated from local features, which are computational expensive, and usually employ GPU in their procedure. To address this problem, we propose a new and simple but effective scheme from an overall perspective. In the proposed scheme, a sort of holistic feature is first extracted from Fourier spectrum, which describes the characteristic of the image or the sub-image as a whole without local feature extraction; then, random forests are utilized to classify images into text and non-text ones. Experimental results in several public datasets demonstrate that this scheme is efficient and effective.

Keywords: natural images; holistic feature; text/non-text image classification; random forests

(Submitted on October 12, 2018; Revised on November 11, 2018; Accepted on December 23, 2018)

1. Introduction

Text/non-text image classification is a helpful and significant problem, which can be applied into image or video retrieval and management, road navigation, and so on. But, the problem is still open and challenging; it is gaining more and more focus from researchers all over the world.

Text in the natural image usually carries a large amount of information, which could be useful in many applications, such as image retrieval, scene analysis and so on. Therefore, text detection and recognition in the natural image have always been hot research areas in computer vision. Since 2015, a new problem has been proposed; that is, how to automatically distinguish text images from non-text ones in natural scene. It is significant and valuable to distinguish text image from non-text images in natural scene. In social network, there are merely 10-15% images containing text [1]; therefore, it wastes a large amount of time and expensive computational power to detect and recognize text directly in the image. If non-text images are removed from the natural image with limited time and computational resource, a lot of time and computational resource will be saved.

To tackle the text/non-text image classification problem, some attempts have been made. In light of a different category of image, previous work could be divided into three parts: document image, video image and natural image.

For the document image, Alessi [2] tried to detect text candidate block and then discriminate text documents from non-text documents with setting threshold. Indermuhle [3] and Vidya [4] both proposed a scheme to address text/non-text regions classification problem in handwritten documents. The works mentioned above are only designed for the document image, not for natural image.

For the video image, Shivakumara [5-7] proposed the methods that the video image was first divided into several blocks, which were classified into text or non-text through clustering by wavelet or edge feature. Shivakumara's works are

* Corresponding author.
E-mail address: yaochao@nwpu.edu.cn

important because the following works on the natural image are all inspired from them.

For the natural image, Zhang [8] was first to propose the problem. He also published a large dataset named TextDis. Zhang's method first utilized MSER to extract region and then extracted from a bag of words with CNN, which was eventually input into SVM to distinct text images from non-text ones. Bai [2] designed and implemented a novel Convolutional Neural Network (CNN) variant called Multi-scale Spatial Partition Network (MSP-Net), which completely managed the text/non-text natural images classification. Lyu [9] presented a method to extract 2-dimensional feature maps of image blocks with Deep Convolutional Neural Network (DCNN) from the original image, which was input into Multi-Dimensional Recurrent Neural Network (MDRNN) to capture spatial context for predicting a label for each block (text or not) in following step.

Of the above three parts, the problem in the document image is simplest due to the constrained environment. The problem in the video image is a little easier than in the natural for relatively fixed position and style of superficial text in the video. In light of text diversity, background complexity, and strong interference factors, the problem in the natural image is the most challenging. Furthermore, text/non-text natural images classification could be widely applied in numerous occasions such as image retrieval, self-driving and image understanding. Although it is challenging and useful, text/non-text natural images classification is gaining more and more attention from researchers in the community.

As shown in Figure 1, text/non-text natural images classification is a preprocessing step before text detection and recognition in the natural image, so it must be effective and efficient. Nowadays, the works [3, 9-10] utilized CNN or CNN variants to extract local features, which is effective and powerful. Nevertheless, CNN or CNN variants is computational expensive, which works either slowly on CPU with taking massive time or efficiently on GPU with consuming massive money. In consideration of cost and power, GPU is not able to be applied in some occasions, e.g. mobile phone. To address the problem, we propose an alternative scheme that is new and simple but effective without local feature. Different from previous works, we argue that the holistic feature can describe the characteristic of image or sub-image in the paper. The scheme with the holistic feature from Fourier spectrum and random forest is subsequently provided to distinguish text ones from non-text ones in natural images dataset.

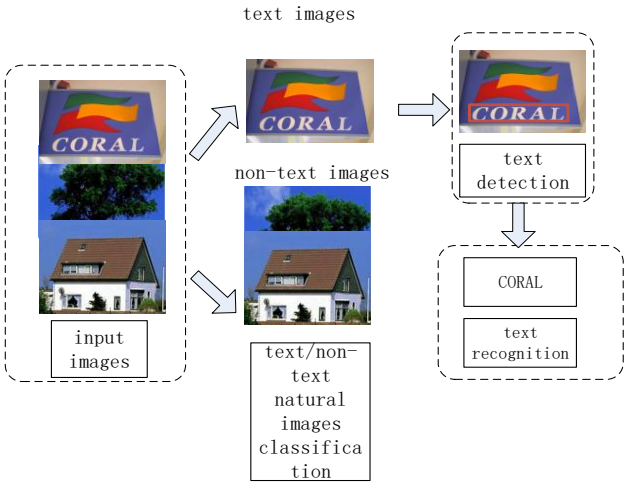


Figure 1. The overall flowchart of text in natural image processing

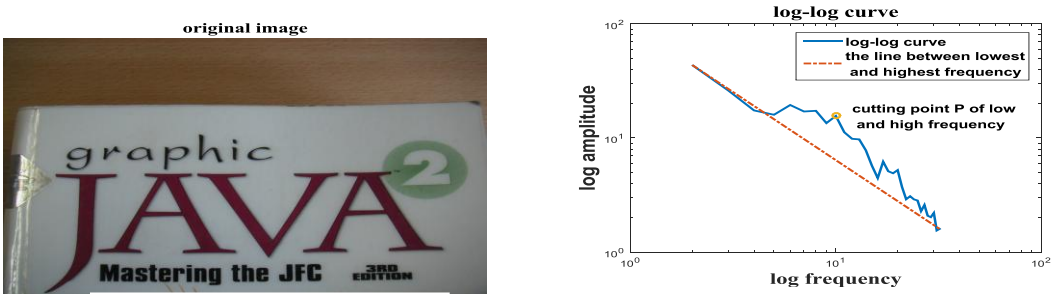
2. The Holistic Feature

The holistic feature is motivated from log-log spectrum and $1/f^{\alpha}$ law. Log-log spectrum and $1/f^{\alpha}$ law are scale-invariant in image statistics, whereas local features are scale-sensitive besides expensive computation.

2.1. Log-Log Spectrum

In image statistics, the Fourier amplitude spectrum of the ensemble of natural images after averaging over orientations lies approximately on a straight line on a log-log scale [10]. But through observation, we argue the log-log curve can be much more approximated by two straight lines separated by the cutting point P (the point marked by orange circle in Figure 2(b), on the left of which low frequency is on the other side of which high frequency is, as shown in Figure 2. The cutting point P

is the point farthest from the line (dash line in Figure 2(b) connecting lowest frequency and highest frequency. Figure 2(a) is a natural text image. The solid curve in Figure 2(b) is log-log curve.



(a) Original image (b) Log-log curve

Figure 2. Log-log curve of Fourier amplitude

Based on the observation described above, we have an experiment on a dataset with 1159 text natural and 1485 non-text natural images to verify whether the fact discovered above is effective for text/non-text natural image classification. There are 3 variables to compute: k1 is the slope of line fitting low frequency part of log-log curve, k2 is the slope of line fitting high frequency part of log-log curve, and k3 is the slope of line fitting the whole log-log curve. The experiment result shown in Figure 3 is that the text natural images (“+” in Figure 3) and the non-text natural images (“○” in Figure 3) are almost separable [11]. In section 2.1, Fourier amplitude spectrum and Fourier energy spectrum are both employed for log-log curve; so, 6-dimensional holistic feature is computed, which are respectively the slope of low frequency line of log-log amplitude spectrum, the slope of high frequency line of log-log amplitude spectrum, the slope of the whole line of log-log amplitude spectrum, the slope of low frequency line of log-log energy spectrum, the slope of high frequency line of log-log energy spectrum and the slope of the whole line of log-log energy spectrum.

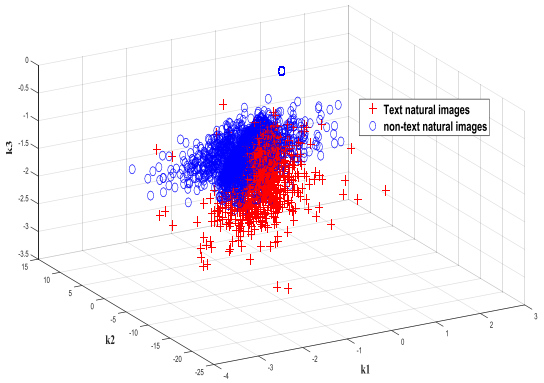


Figure 3. Separability of the holistic feature on log-log spectrum

2.2. $1/f^{\alpha}$ *Law*

$1/f^{\alpha}$ law is widely known as a property in natural image statistics. It is noteworthy that $1/f^{\alpha}$ law is derived from log spectrum rather than log-log spectrum. It states that the averaged Fourier spectrum of the Amplitude $A(f)$ for the ensemble of natural images obeys a distribution:

$$E\{A(f)\} \propto 1/f^{\alpha} \tag{1}$$

For the convenience of computation, we change the Equation (1) into Equation (2) such as below, and divide the log spectrum into two parts with cutting point P proposed in section 2.1. We compute respectively the means and the variances of C and α of text and non-text natural images in the dataset described in section 2.1, which are plotted in Figure 4. Figure 4 shows that C and α of text natural images are obviously different from non-text natural images.

$$E\{A(f)\} = 1/(C \cdot f^{\alpha}) \tag{2}$$

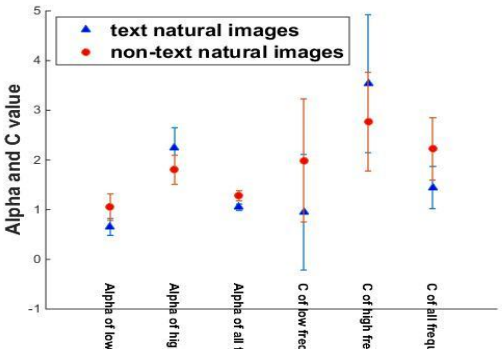


Figure 4. Means and variances of *C* and alpha

In summary, we compute *C* of low frequency of Fourier amplitude spectrum, *C* of high frequency of Fourier amplitude spectrum, *C* of Fourier amplitude spectrum, alpha of low frequency of Fourier amplitude spectrum, alpha of high frequency of Fourier amplitude spectrum, and alpha of of Fourier amplitude spectrum as 6-dimensional holistic features and also 6-dimensional holistic features like above in Fourier energy spectrum. There are 12-dimensional holistic features in total.

In summary, 18 (6+12) dimensional holistic features is employed to distinguish text natural images from non-text ones.

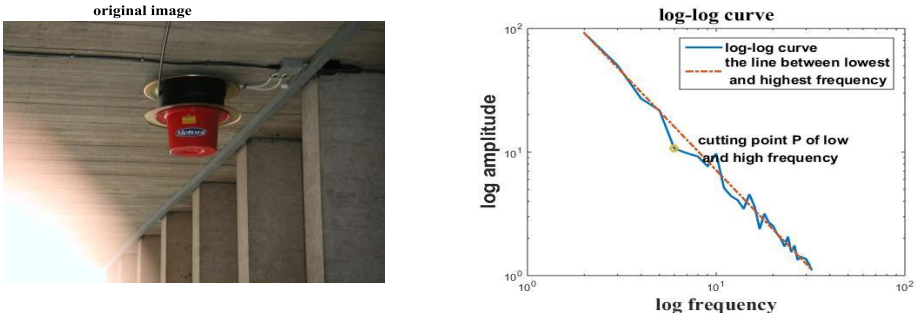
3. Methodology

The feature proposed is named holistic feature for two reasons. First, the feature is extracted based on Fourier transform, which transforms the whole image from space domain into frequency domain. Second, the feature is the descriptor for overall characteristic of the whole image. Therefore, the holistic feature is able to be utilized to discriminate whether the whole image contains text, regardless of both the location of text and the completeness of text in the image. Furthermore, the holistic feature can even be used to classify the image with much larger size than the text area in the image if the text area relatively is not too small.

In Figure 5, image patches have been found by Random Forests with the holistic feature and sliding-window traversed through training images from TextDis dataset. It is clear that there are parts of character, multiple characters, distorted character, and skewed character. So, the holistic feature is not dependent on the completeness alignment of text image. And it is also obvious that the holistic feature is language-independent since several kinds of language characters are found in the image patches.

Figure 5. Image patches discovered with the holistic feature

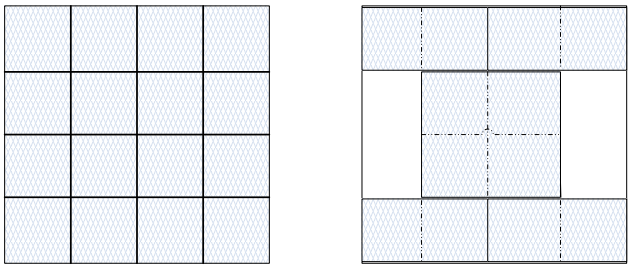
The traditional hand-crafted feature, after getting extracted from image, is directly input into classifier for training. But the holistic feature faces a special problem: the characteristic of text natural image will be ambiguous if the text only occupies a small portion of the whole natural image, because the information of background veils the information of text. The holistic feature directly extracted from the whole image will not work well in this case, as shown in Figure 6. For the original image in the Figure 6(a), the text is so small that the two characteristics of the log-log curve is hard to visualize and the classifier will be confused by this kind of text/non-text natural image.



(a) Original image (b) Log-log curve

Figure 6. An image with small portion of text

In the training, to alleviate the situation, we normalize the size of each image into 256×256, and divide the image into several sub-images. This is shown in Figure 7, which makes the text a larger percentage of the image or sub-image to avoid failure of the holistic feature. Sub-images with text area less than *t* are labeled as non-text; the others are known as text. *t* is an empirical value set as 0.2. In the following step, we extract the holistic feature from each sub-image, and input them into random forests for training.



The shadow regions are sub-images, and solid lines are boundaries of sub-images.
Some large sub-images are made up of several small sub-images, and dash lines are boundaries of small sub-images.
Figure 7. The division of natural image.

In the classification, the process is just similar to that of the training. The difference is that the holistic feature of each sub-image is input into random forests to classify. Any sub-image is classified as text, and the whole image will be deemed as a text natural image.

4. Experiments and Analysis

We evaluate the holistic feature and the scheme for text/non-text natural images classification on three datasets: TextDis, ICDAR2003 and Hua's dataset. The results show the effectiveness of the holistic feature and the scheme.

4.1. Datasets

Among the three benchmarks, only TextDis is specially established for the task of text/non-text natural images classification by Zhang [8]. The dataset is mainly composed of natural images which is majority, born-digital images and scanned document images. There are 7302 text images and 8000 non-text images in the training set; meanwhile, there are 2000 text images and 2000 non-text images in the testing set.

ICDAR2003 is a public scene text dataset with more than 250 images, both for training and testing for text detection. Hua's is a video frame dataset with 45 video images.

TextDis is the most difficult, with complex background and variable condition. In the next section, we will show the performances of our method in the three datasets. The proposed method does worse in TextDis than the other two.

4.2. Performances Evaluation

To discuss the performances evaluation of the method, we should take two factors into consideration: metric and dataset. We adopt different metrics for convenience to compare with other researchers; the three datasets are listed in the last section.

The evaluation, listed as Table 1, in Hua's dataset aims to demonstrate the generalization of our scheme in the video frame. The difficulties in video frames are relatively simpler than those in natural scene images. So, compared with other methods, our method has a reasonable result.

Table 1. Evaluation in Hua's dataset

Methods	Text (%)	Non-text (%)
Proposed Method	100	100
Bai	100	100
Shivakumara [12]	97.62	100
Shivakumara [5]	75.54	24.46

The evaluation listed as Table 2 in ICDAR2003 shows the result in focused scene text dataset. ICDAR2003 is more complicated and difficult than Hua's and all the schemes performance worth than in Hua's. Our method does better than Shivakumara's two methods, but fails comparing with Bai's method with CNN.

Table 2. Evaluation in ICDAR2003 dataset

Methods	Text (%)	Error (%)
Proposed Method	82.21	17.79
Bai et al. [1]	89.2	10.8
Shivakumara et al. [12]	80.97	19.03
Shivakumara et al. [5]	81.12	18.88

The evaluation, listed as Table 3, in TextDis dataset is a little challenging for our method. It is just a bit better than the other three methods but can hardly compare with Bai's method. However, we still argue our method has its own advantages; it is simple and efficient. The method with CNN usually works with GPU, which is expensive and needs high power. Our method implemented in Matlab without any optimization is experimented on the PC with I3 1.9G CPU and 4G RAM. It takes 2.1s to process per image, which can be improved.

Figure 8 shows the classification result of the three datasets; our method obtains reasonable performances.

Table 3. Evaluation in TextDis dataset

Methods	Precision (%)	Recall (%)	F-Measuremen
Proposed Method	80.5	91.3	85.5
Bai [1]	93.7	95.4	94.6
Zhang [8]	75.4	97.9	85.1
Yao [13]	80.8	90.2	85.3
Neumann [14]	52.5	98.4	68.5

(a) Some samples of Hua's dataset (b) Some samples from ICDAR2003 dataset (c) Some samples of TextDis dataset

Figure 8. Classification result of proposed method

5. Conclusion

This paper presents a new, simple but effective kind of the holistic feature, which can distinguish the text natural images from the non-text ones. This paper also proposes a scheme with the holistic feature. It makes the machine discover the text image from an overall perspective of an image.

The holistic feature and the scheme are simple and efficient, and both can work without GPU. Thus, they will be an alternative solution in some cases where the hardware is low power or low cost.

Acknowledgement

This work was primarily supported by National Natural Science Foundation of China (NSFC) (No.61601184) and Key scientific research project of Education Department of Henan Province (No.16A520018). The authors thank the anonymous reviewers for their helpful suggestions.

References

1. X. Bai, B. Shi, C. Zhang, X. Cai, and L. Qi, "Text/Non-Text Image Classification in the Wild With Convolutional Neural Networks," *Pattern Recognition*, Vol. 66, No. 6, pp. 437-446, 2017
2. N. G. Alessi, S. Battiato, G. Gallo, M. Mancuso, and F. Stanco, "Automatic Discrimination of Text Images," in *Proceedings of SPIE*, pp. 351-359, 2003
3. E. Indermuhle, H. Bunke, F. Shafait, and T. Breuel, "Text Versus Non-Text Distinction in Online Handwritten Documents," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, pp. 3-7, 2010
4. V. Vidya, T. R. Indhu, and V. K. Bhadran, "Classification of Handwritten Document Image into Text and Non-Text Regions," in *Proceedings of the Fourth International Conference on Signal and Image Processing*, pp. 103-112, 2012
5. P. Shivakumara, A. Dutta, T. Q. Phan, C. L. Tan, and U. Pal, "A Novel Mutual Nearest Neighbor based Symmetry for Text Frame Classification in Video," *Pattern Recognition*, Vol. 44, No. 8, pp. 1671-1683, 2011
6. P. Shivakumara, A. Dutta, and C. L. Tan, "A New Symmetry based on Proximity of Wavelet-Moments for Text Frame Classification in Video," in *Proceedings of International Conference on Pattern Recognition*, pp. 129-132, 2010
7. P. Shivakumara and C. L. Tan, "Novel Edge Features for Text Frame Classification in Video," in *Proceedings of International Conference on Pattern Recognition*, pp. 3191-3194, 2010
8. C. Zhang, C. Yao, B. Shi, and X. Bai, "Automatic Discrimination of Text and Non-Text Natural Images," in *Proceedings of International Conference on Document Analysis and Recognition*, pp. 886-889, 2015
9. P. Lyu, B. Shi, C. Zhang, and X. Bai, "Distinguishing Text/Non-Text Natural Images with Multi-Dimensional Recurrent Neural Networks," in *Proceedings of International Conference on Pattern Recognition*, pp. 3981-3986, 2016
10. X. D. Hou and L. Q. Zhang, "Saliency Detection: A Spectral Residual Approach," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007
11. W. Jiang, Z. Y. Lu, J. Li, X. P. Liu, and C. Yao, "Visual Saliency and Text Confidence Map based Background Suppression for Scene Text," *Chinese Journal of Electronics*, Vol. 43, No. 1, pp. 62-68, 2015
12. N. Sharma, P. Shivakumara, U. Pal, M. Blumenstein, and C. L. Tan, "Piece-wise Linearity based Method for Text Frame Classification in Video," *Pattern Recognition*, Vol. 48, No. 3, pp. 862-881, 2015
13. C. Yao, X. Bai, and W. Y. Liu, "A Unified Framework for Multi-Oriented Text Detection and Recognition," *IEEE Transaction on Image Processing*, Vol. 23, No. 11, pp. 4737-4749, 2014
14. L. Neumann and J. Matas, "A Method for Text Localization and Recognition in Real-World," in *Proceedings of ACCV*, pp. 770-783, 2010

Wei Jiang received the PH.D. degree from Xidan University, Xi'an, China, in 2014. He now works as a senior lecturer in School of Software, North China University of Water Resources and Electric Power in Zhengzhou, China. His interest is scene text detection and recognition.

Jiayi Wu received her master degree from the University of Warwick, England, in 2010. She now works for School of Software, North China University of Water Resources and Electric Power in Zhengzhou, China.

Chao Yao received the PH.D. degree from Xidan University, Xi'an, China, in 2014. He has visited Concordia University in Montreal, Canada as joint PH.D. Student from 2011 to 2012. He has finished post-doc in 2017 and works as assistant professor in School of Automation, Northwestern Polytechnic University, in Xi'an, China. His interest is text recognition and dimension reduction.