Multi-scale Graph-matching Based Kernel for Character Recognition from Natural Scenes

SHI Cun-Zhao[1] WANG Chun-Heng[1] XIAO Bai-Hua[1] ZHANG Yang[2] GAO Song[1]

Abstract Recognizing characters extracted from natural scene images is quite challenging due to the high degree of intraclass variation. In this paper, we propose a multi-scale graph-matching based kernel for scene character recognition. In order to capture the inherently distinctive structures of characters, each image is represented by several graphs associated with multi-scale image grids. The similarity between two images is thus defined as the optimum energy by matching two graphs (images), which finds the best match for each node in the graph while also preserving the spatial consistency across adjacent nodes. The computed similarity is suitable to construct a kernel for support vector machine (SVM). Multiple kernels acquired by matching graphs with multi-scale grids are combined so that the final kernel is more robust. Experimental results on challenging Chars74k and ICDAR03-CH datasets show that the proposed method performs better than the state of the art methods.

Key words Character recognition, structure, graph-matching, energy, kernel, histograms of oriented gradients (HOG), support vector machine (SVM)

Citation Shi Cun-Zhao, Wang Chun-Heng, Xiao Bai-Hua, Zhang Yang, Gao Song. Multi-scale graph-matching based kernel for character recognition from natural scenes. *Acta Automatica Sinica*, 2014, **40**(4): 751−756

DOI 10.3724/SP.J.1004.2014.00751

With the rapid growth of intelligent mobile phones and portable devices, camera-based applications such as automatic sign reading, license plate recognition and navigation, are holding a huge market. Detecting and recognizing text in the natural images is indispensable for these applications. Although a lot of work has been done on text detection and some methods are shown to have achieved satisfactory performance[1−4], the performance of character recognition from natural scenes is still far from enough, due to the high degree of intraclass variation caused by various font styles, complex background, geometric distortions, unconstrained illuminations and different resolutions[5]. Therefore, aiming to make some contribution in this area, in this paper, we focus on the recognition of characters taken from scene images.

Generally speaking, the existing scene character recognition methods could be roughly classified into two categories: traditional optical character recognition (OCR) based and object recognition based methods. For traditional OCR based methods[6], various approaches have been proposed to get the binarized image which is directly fed into the off-the-shelf OCR software. However, since characters in scene images differ from text in traditional scanned document in terms of resolution, illumination conditions and font style, OCR system performs badly due to the unsatisfactory binarization result. On the other hand, object recognition based methods assume that scene character recognition is quite similar to object recognition with a high degree of intraclass variation. Campos et al.[5] benchmarked the performance of various features based on a bag-of-visual-words (BOW) representation to assess the feasibility of posing the problem as an object recognition task and showed that geometric blur[7] and shape context[8] in conjunction with nearest neighbor (NN) classifier, performed better than other methods. Wang et al.[9] proposed to use histograms of oriented gradients (HOG)[10] in conjunction with an NN classifier and reported better performance. Newell and Griffin[11] proposed two extensions of HOG descriptor to include features at multiple scales and their method achieved best recognition results so far on two datasets, Chars74k[5] and ICDAR03-CH[12], using the same evaluation framework as Campos[5].

As BOW discards all spatial information while HOG enforce certain degree of spatial consistency, HOG based methods perform better than BOW based methods[9, 11], which means structure information is quite important. However, HOG or multi-scale HOG features could not describe character images from natural scenes very well especially when the images have different font styles, geometric distortions or deformations. Fortunately, most characters are designed differently from each other in terms of shape or structure for reading. Therefore, structure-based representations that are invariant to deformations or distortions should be appropriate for characters. Thus, inspired by the recent progress in object recognition[13], we use a graph to represent each character image and propose a multi-scale graph-matching based kernel to recognize scene characters. Concretely, each character image is represented by a graph whose nodes correspond to a set of image blocks and edges reflect the geometric structure of the image[13]. The similarity between two character images are formulated as the optimum energy for matching the two graphs, which finds the best match for each node while also preserving the spatial consistency across adjacent nodes. The similarity could be directly used for NN classifier, or as the kernel for SVM based classifier. In order to make the final kernel more robust, multiple kernels computed by matching graphs with multi-scale image grids are optimally combined. Experimental results on both Chars74k and ICDAR03-CH datasets show promising performance.

1 The proposed method

The flowchart of the proposed method is shown in Fig. 1. For training, first, all the training images are represented by graphs associated with multi-scale image grids and similarity based kernels are acquired by graph matching for any two images. Then, kernels associated with graphs of multi-scale grids are combined to train a kernel-based SVM. For testing, given the test image, the similarity between the

Manuscript received May 22, 2012; accepted September 27, 2013
Supported by National Natural Science Foundation of China (60933010, 61172103, 61271429)
Recommended by Associate Editor DAI Qiong-Hai
1. The State Key Laboratory of Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China 2. Kuyun Interactive Technology Limited, Beijing 100007, China

test image and each of the training images is calculated by optimizing the energy function of matching the two graphs.

Then, the similarities are used to construct a kernel and kernels of different scales are combined to form the final kernel for SVM classification.

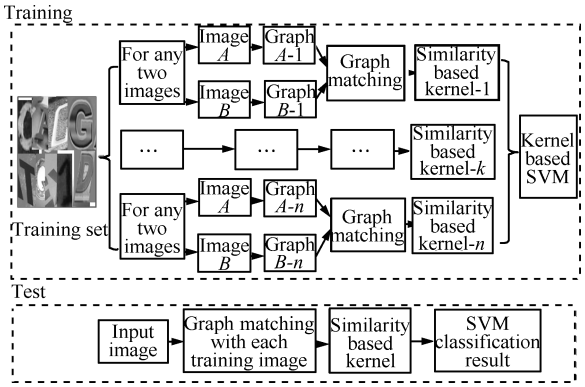


Fig. 1 Flowchart of the proposed method

1.1 Representing images by graphs associated with multi-scale grids

Each character image is represented by several graphs associated with image grids of different sizes. An undirected graph G is composed of nodes corresponding to a coarse image grid and each node is connected with its four neighbors. As shown in Fig. 2, the nodes are not pixels but the image grid and are indexed by the position of the grid. Each node n in G is represented by a feature vector F_n related to the corresponding image region. Since previous paper[9,11] has proved that HOG[10] is a better choice for character recognition, we choose HOG as the local region descriptor. 8 orientations of HOG are used and thus each node is represented by a 8-dimension vector. Since each character has its own structure, the image grid suitable for one character might not be proper for another. Thus, we use multiple levels of image grids to represent each image so that different image grids could complement each other. In this paper, we use four graphs with multi-scale image grids as shown in Fig. 2.

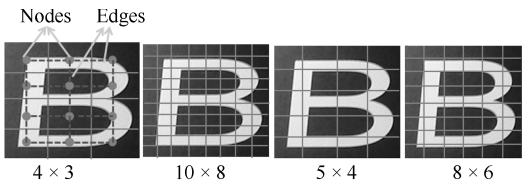


Fig. 2 The construction of the graph and the four grids used in the paper

1.2 Getting the similarity by graph matching

To match the images, we need to match the two graphs representing the images. We use the graph matching algorithm proposed by Duchenne et al.[13], which is fast and suitable to the grids of moderate size considered here. In their paper, to match the two graphs, the first graph G is distorted to the other one $G^{'}$ while enforcing spatial consistency across adjacent nodes. Concretely, given a node n in G and some displacement d_n, n is matched to the node $n^{'}$ in $G^{'}$ and the best matching is the one that maximizes the following energy function

$$E_{\rightarrow}(\boldsymbol{d}) = \sum_{n \in V} U_n(d_n) + \sum_{(m,n) \in \varepsilon} B_{m,n}(d_m, d_n) \quad (1)$$

where V and ε represent the set of nodes and edges of G, $\boldsymbol{d}$ is the vector formed by the displacements associated with all the elements of V, and U_n, $B_{m,n}$ denote unary and binary potentials respectively. For each node n, we fix a maximum displacement K in each direction, leading to a total of K^2 possible displacements. Thus, the energy function defined in (1) is a multi-label Markov random field (MRF) where the labels correspond to the displacements. In the experiment, K is set to 8.

The unary potential $U_n(d_n)$ is defined as the negative χ^2 distance between $\boldsymbol{F}_n$ and $\boldsymbol{F}_n^{'}$, where $\boldsymbol{F}_n$ and $\boldsymbol{F}_n^{'}$ are the feature vectors representing node n in graph G and $G^{'}$ respectively. The binary potential $B_{m,n}$ enforces spatial consistency and acts as a spring:

$$B_{m,n} = -\lambda ||d_m - d_n|| \quad (2)$$

where λ is the positive spring constant. l_1 distance is used to be robust to sparse distortion differences.

To get the similarity, we need to calculate the optimum energy for matching graph G to graph $G^{'}$ and the one for matching graph $G^{'}$ to graph G. The similarity between two images is defined as

$$S_{m,n} = \max\{\max_{\boldsymbol{d}_1} E_{\rightarrow}(\boldsymbol{d}_1)), \max_{\boldsymbol{d}_2} E_{\leftarrow}(\boldsymbol{d}_2)\} \quad (3)$$

where $E_{\rightarrow}(\boldsymbol{d}_1)$ and $E_{\leftarrow}(\boldsymbol{d}_2)$ is the energy function for matching G to $G^{'}$ and matching $G^{'}$ to G, respectively. We use the 2-step curve expansion[13] to optimize the energy function. The similarities between all the pairs of images are used to construct a kernel suitable for support vector machine (SVM).

Compared to traditional feature matching methods, in which the feature for each block could be only compared with the corresponding block, the proposed graph matching scheme is more robust to image rotations, since each block (node) could be compared with multiple blocks (nodes) and thus the chances of finding the corresponding block in the rotated image is larger. Moreover, for rotated images, the structure of the image does not change a lot. Since the binary potential enforces spatial consistency, the graph matching could find the best match for each block while also preserving the spatial consistency across adjacent nodes. Fig. 3 shows the difference between the traditional feature matching and the graph matching proposed in this paper. As we can see in Fig. 3 (a), for the traditional feature matching, since each block of the image is only compared with the corresponding block of the other one, the matching is sensitive to image distortions. While for graph matching, as we can see in Fig. 3 (b), the block x at (row 1, column 1) in the first image is matched to block x' at (row 2, column 1) in the rotated image, which is acquired by comparing x with other nodes and choosing the best match x' while also preserving the spatial consistency across adjacent nodes.

1.3 Combining multi-scale kernels

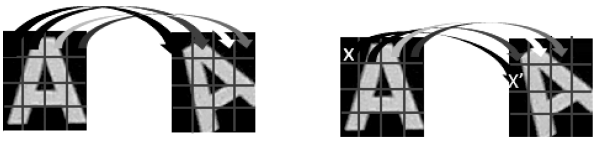
Since each image is represented by multiple graphs with multi-scale image grids, we get multiple kernels by matching graphs with different grids. Now we need to combine the kernels so that complementary information by matching graphs with multi-scale grids could be optimally incorporated into the final kernel. Campos et al.[5] proved that multiple kernel learning (MKL) which combines all the

features, performed better than each single feature. However, as shown by Gehler and Nowozin[14], simple kernel combination methods, such as averaging kernel in (4) and product kernel in (5), which are almost always left out in comparisons yield equally good results but are magnitudes faster than other combination methods, such as MKL[15], CG-Boost and LP-B[14]. Thus, we compare the recognition results with averaging kernel, product kernel as well as the corresponding single kernel.

$$k_a^*(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{F} \sum_{m=1}^{F} k_m(\boldsymbol{x}, \boldsymbol{x}') \tag{4}$$

$$k_p^*(\boldsymbol{x}, \boldsymbol{x}') = (\prod_{m=1}^{F} k_m(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{F}} \tag{5}$$

where k_m is the kth kernel and F is the total number of kernels.



(a) (b)

Fig. 3 The difference between the traditional feature matching and graph matching ((a) Traditional feature matching in which each block is only compared with the corresponding block; (b) Graph matching in which each block is compared with multiple blocks to get the best match while also preserving the spatial constraint)

2 Experiments and results

2.1 Datasets

Two most commonly used datasets, the Chars74k dataset[5] and ICDAR03-CH[12], are used to evaluate the proposed approach. The Chars74k dataset contains 62 classes consisting of digits, upper and lower case letters. Some images from the dataset are shown in Fig. 4. As we can see, the characters have different font styles, various distortions and different degrees of rotations and some images have part of other characters in addition to the main character. The second dataset, ICDAR03-CH, is the robust character recognition dataset from ICDAR 2003, which is quite similar to Chars74k dataset. Fig. 4 shows some examples.

Fig. 4 Some examples from the Chars74k dataset (left) and the ICDAR03-CH dataset (right)

2.2 Dataset splits

In order to compare our method with other methods, we use the same evaluation framework as Campos[5], in which the training images for Chars74k are 5 and 15 per class, referred to as Char74k-5 and Chars74k-15 respectively and the training images for ICDAR03-CH dataset is 5 per class.

For the Chars74k dataset, we randomly select 30 image per class from which to get the training and test sets. The test set is 15 images per class while the training set varies from 5 to 15 images per class. For the ICDAR03-CH dataset which consists of training and test sets, we randomly select 5 images per class from the training set and perform testing on the whole test set.

2.3 Kernel selection

In this experiment, we first evaluate the sensitivity of the proposed method to the levels of image grids. we test the performance of the proposed method with image grids 3×2, 4×3, 5×4, 8×6, 10×8, 20×16 using the same training and test set. The results show that when the image grid is too coarse, such as 3×2, the performance drops significantly. The reasons lie in: 1) Since we only use 8 orientations of HOG to describe each block, when the grid is too coarse, the extracted 8-d features could not represent the region very well, and 2) The graph matching does not work well at such coarse scale. When the image grid is 20×16, the performance also drops a lot. The reason lies in the fact that the size of each region representing each node is so small that it could not carry enough information for matching the graphs. Thus, in the paper, we choose the image grids which are not too coarse or too fine. In the following experiments, four levels of image grids are used.

Next, we evaluate the performances of the averaging kernel and the product kernel as well as each single kernel. Kernel-based SVM[16] is used as the classifier. We use Chars74k as the benchmark dataset. 15 images are randomly chosen as training images per class. The results are shown in Table 1. Each score is the average performance over 50 runs. In the table, Kernel-1, Kernel-2, Kernel-3 and Kernel-4 represent the kernel acquired by matching graph with grid size of 5×4, 8×6, 4×3 and 10×8 respectively. Kernel-average and Kernel-product are the averaging and product kernel of the above four kernels while Kernel-average3 are the averaging kernel of Kernel-1, Kernel-2 and Kernel-4. From the results we can see that before kernel combination, Kernel-1, Kernel-2 and Kernel-4 perform better than Kernel-3, while after combination, all the combination kernels, Kernel-average, Kernel-product and Kernel-average3 achieve better performance than each single kernel. Moreover, the product kernel does not work as well as averaging kernel. Among all the kernels, Kernel-average which combines all the four kernels achieves the highest recognition result. Surprisingly, however, Kernel-average3, which combines three better kernels and excludes the worst one, does not perform as well as Kernel-average which combines all the four kernels equally, suggesting that to some extent, the averaging kernel could acquire complementary information from the weaker kernels.

2.4 Comparison results with other methods

We test the proposed multi-scale graph-matching based

Table 1 Results of different kernels

Kernels	Kernel-1	Kernel-2	Kernel-3	Kernel-4	Kernel-average	Kernel-average3	Kernel-product
Chars74k-15	63.8 ± 1.1	64.9 ± 0.9	62.7 ± 1.1	64.1 ± 0.6	$\mathbf{69.6 \pm 0.8}$	68.5 ± 0.8	66.1 ± 0.8

kernel for SVM (MGMK-SVM) on two challenging datasets, Chars74k and ICDAR03-CH. The results along with previously published results, including the shape context[5], the OCR software ABBYY[9] and HOG columns[11] are given in Table 2. As the evaluation method proposed by Newell and Griffin[11], each score is the average performance over 50 runs for the Chars74k and 10 runs for the ICDAR03-CH-5. The results show that the proposed MGMK-SVM performs better than all the previously published methods. Specifically, the averaging kernel achieves an increase of 12.5 % than multiple kernel learning of BOW based features on Chars74k-15, suggesting that the structure information ignored by BOW is better represented by the graph-matching based kernel. The proposed MGMK-SVM also outperforms the HOG[9], which is used as basic region descriptor for the graph, showing the superiority of the graph matching to the traditional feature matching. Moreover, it also outperforms the best published result acquired by multiscale HOG, proving the effectiveness of the complementary information of the multi-scale graph constructions as well as the structure information captured by the graph-matching based kernel. The full results of MGMK-SVM, HOG and HOGC[11] for Chars74k dataset, with training sets varying from 5 and 15 images per class are shown in Fig. 5 (a). The results show that MGMK-SVM performs constantly better than multiscale

Table 2 Comparison results of different methods

Methods	Chars74k-5	Chars74k-15	ICDAR03-CH-5
Shape context[5]	26.1 ± 1.7	34.4	18.3
Geometric blur[5]	36.9 ± 1.0	47.1	27.8
Multiple kernel learning[5]	–	55.3	–
ABBYY[9]	18.7	18.7	21.2
HOG features[9]	45.3 ± 1.0	57.5	51.5
HOG multiscale[11]	49.1 ± 1.3	58.8 ± 1.2	48.3 ± 1.2
HOG columns[11]	57.7 ± 1.1	66.5 ± 1.2	57.1 ± 0.9
MGMK-SVM	$\mathbf{58.3 \pm 1.2}$	$\mathbf{69.6 \pm 0.8}$	$\mathbf{60.7 \pm 0.9}$
MGMK-Exp-SVM	$\mathbf{58.5 \pm 1.1}$	$\mathbf{69.9 \pm 0.8}$	$\mathbf{61.3 \pm 1.0}$

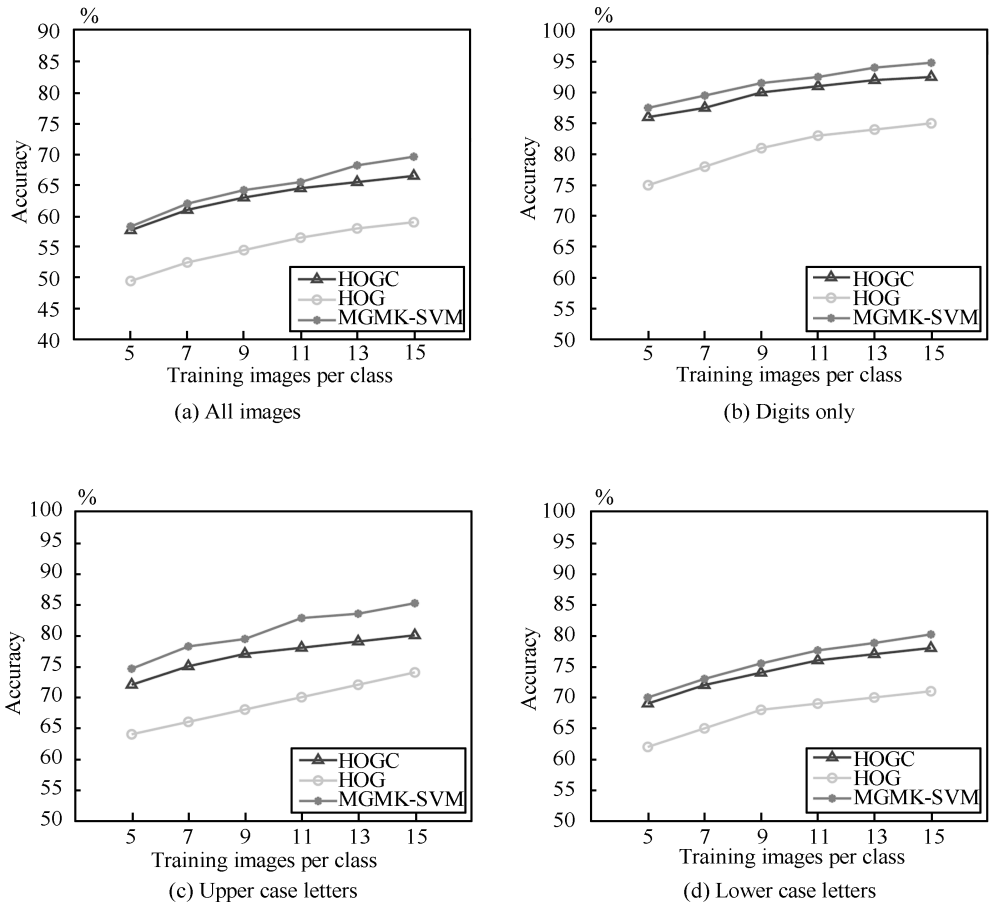


Fig. 5 Performance of Chars74k dataset with increasing training images

HOG with varied training images. We also test the recognition rate using the exponential kernel (MGMK-Exp-SVM) and the results show that the performance could be further improved, suggesting the effectiveness of using graph matching energy between characters as the similarity measure.

2.5 Results of separate dataset

Although the proposed method achieves better result than preciously published methods, the recognition rate is still far from satisfactory. If we look into the dataset, we find some letters are intrinsically difficult to recognize, due to the similarity between the upper and lower case letters, such as letter "c", "l", "o" and so on. Thus, in order to evaluate the performance without this influence, we also test the methods using only digits, only upper or lower case letters. The performance alongside the results reported by Newell and Griffin[11] are shown in Fig. 5 (b) to 5 (d). As we can see, with the increasing of training images, there is a constant growth of the recognition scores of all the three methods and the proposed MGMK-SVM constantly achieves better result. For digits, the proposed method achieves a recognition score of about 94 %. However, for the upper case letters, the recognition score is around 85 %, while for the lower case letters, the score is only around 80 %, indicating that there is still room for improvement.

3 Conclusions and discussions

In this paper, we propose a multi-scale graph-matching based kernel for scene character recognition. Structure information is captured by representing each image with a graph whose nodes correspond to a set of image regions and the edges reflect the geometric constraint. Thus, the similarity between two images is defined as the maximum energy for matching the two graphs so that not only correspondences are found for the nodes in the graph, the spatial consistency across adjacent nodes is also well preserved. Furthermore, complementary information are acquired by combining multi-scale graph-matching based kernels so that the result could be further improved. Experimental results on Chars74k and ICDAR03-CH datasets show that the proposed method outperforms previously published methods.

Although our approach has achieved promising performance, there is still much room for improvement. By graph matching, certain degree of structure information could be considered. However, since we represent each image by a graph with the same image grid, if two characters are not in the same position of the image which is quite common in the dataset, the representation might not be good enough. Moreover, as different characters have different structures, it might be more suitable to represent each character by a graph with unique structure. In the future, we will try to represent each image by more flexible graph so that more valuable structure information could be acquired to further improve the recognition accuracy.

References

1 Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA: IEEE, 2010. 2963−2970

2 Pan Y F, Hou X W, Liu C L. A hybrid approach to detect and localize texts in natural scene images. *IEEE Transactions on Image Processing*, 2011, **20**(3): 800−813

3 Shivakumara P, Phan T, Tan C L. A Laplacian approach to multi-oriented text detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(2): 412−419

4 Shahab A, Shafait F, Dengel A. International Conference on Document Analysis and Recognition (ICDAR) 2011 robust reading competition challenge 2. Reading text in scene images. In: Proceedings of the 2011 IEEE Conference on Document Analysis and Recognition. Beijing, China: IEEE, 2011. 1491 −1496

5 de Campos T E, Babu B R, Varma M. Character recognition in natural images. In: Proceedings of the 2009 IEEE Conference on Computer Vision Theory and Applications (VISAPP). Lisbon, Portugal: IEEE, 2009. 273−280

6 Chen X R, Yuille A L. Detecting and reading text in natural scenes. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE, 2004. 366−373

7 Berg A C, Berg T L, Malik J. Shape matching and object recognition using low distortion correspondences. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 26−33

8 Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**(4): 509−522

9 Wang K, Belongie S. Word spotting in the wild. In: Proceedings of the 11th European Conference on Computer Vision. Berlin, Heidelberg: Springer-Verlag, 2010. 591−604

10 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 886−893

11 Newell A J, Griffin L D. Multiscale histogram of oriented gradient descriptors for robust character recognition. In: Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China: IEEE, 2011. 1085 −1089

12 Lucas S M, Panaretos A, Sosa L, Tang A, Wong S, Young R. ICDAR 2003 robust reading competitions. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, UK: IEEE, 2003. 682−687

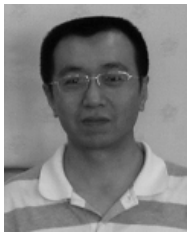
13 Duchenne O, Joulin A, Ponce J. A graph-matching kernel for object categorization. In: Proceedings of the 2011 International Conference on Computer Vision. Barcelona: IEEE, 2011. 1792−1799

14 Gehler P, Nowozin S. On feature combination for multiclass object classification. In: Proceedings of the 12th International Conference on Computer Vision. Kyoto: IEEE, 2009. 221−228

15 Bach F R, Lanckriet G R G, Jordan M I. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning. New York: USA: ACM, 2004. doi: 10.1145/1015330.1015424

16 Chang C C, Lin C J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**(3): 1−27

SHI Cun-Zhao Ph. D. candidate in pattern recognition and artificial intelligence at the Institute of Automation, Chinese Academy of Sciences. She received her bachelor degree in electronic engineering from Wuhan University in 2009. Her research interest covers text detection, text extraction, character recognition, and text recognition in natural images.
E-mail: cunzhao.shi@ia.ac.cn



WANG Chun-Heng Professor at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor and master degrees in electronic engineering from Dalian University of Technology, and the Ph. D. degree in pattern recognition and intelligent control from the Institute of Automation, Chinese Academy of Sciences, in 1993, 1996 and 1999, respectively. From 2004, he has been a professor at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interest covers pattern recognition, image processing, neural networks, and machine learning. Corresponding author of this paper.
E-mail: chunheng.wang@ia.ac.cn

XIAO Bai-Hua Professor at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree in electronic engineering from Northwestern Polytechnical University, and the Ph. D. degree in pattern recognition and intelligent control from the Institute of Automation, Chinese Academy of Sciences, in 1995 and 2000, respectively. From 2005, he has been a professor at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interest covers pattern recognition, image processing, and machine learning.
E-mail: baihua.xiao@ia.ac.cn

ZHANG Yang Research and development staff at Kuyun Interactive Technology Limited Company. He received his bachelor degree from Beijing Normal University in 2006. His research interest covers text detection, text extraction, pattern recognition, image processing, computer vision, and machine learning.
E-mail: zhang.yang@kuyun.com

GAO Song Ph. D. candidate in pattern recognition and artificial intelligence at the Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree from Xi′an Jiaotong University in 2010. His research interest covers text detection, text extraction, character recognition, and text recognition in natural images.
E-mail: song.gao@ia.ac.cn