International Journal of Automation and Computing 11(5), October 2014, 480-488
DOI: 10.1007/s11633-014-0833-2

Robust Text Detection in Natural Scenes Using Text Geometry and Visual Appearance

Sheng-Ye Yan[1] Xin-Xing Xu[2] Qing-Shan Liu[1]

[1]School of Information and Control, Nanjing University of Information Science and Technology, Nanjing 210044, China

[2]School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore

Abstract: This paper proposes a new two-phase approach to robust text detection by integrating the visual appearance and the geometric reasoning rules. In the first phase, geometric rules are used to achieve a higher recall rate. Specifically, a robust stroke width transform (RSWT) feature is proposed to better recover the stroke width by additionally considering the cross of two strokes and the continuousness of the letter border. In the second phase, a classification scheme based on visual appearance features is used to reject the false alarms while keeping the recall rate. To learn a better classifier from multiple visual appearance features, a novel classification method called double soft multiple kernel learning (DS-MKL) is proposed. DS-MKL is motivated by a novel kernel margin perspective for multiple kernel learning and can effectively suppress the influence of noisy base kernels. Comprehensive experiments on the benchmark ICDAR2005 competition dataset demonstrate the effectiveness of the proposed two-phase text detection approach over the state-of-the-art approaches by a performance gain up to 4.4% in terms of F-measure.

Keywords: Text detection, geometric rule, stroke width transform (SWT), support vector machine (SVM), multiple kernel learning (MKL).

1 Introduction

Detecting texts in natural scenes is the first step for understanding the texts or sentences in natural images. It is always a key module for the consumer electronic products, such as car plate recognition based garage door automatic opening system, TV′s subtitle detection, shot boundary detection system aided with detected texts or aiding-device of bind people navigation system. The major challenges for detecting texts mainly come from two aspects: the diversity of the texts and the cluttered backgrounds. Especially, the texts in natural images may be written in different languages, fonts, colors, scales, orientations, etc.

In the past several decades, a large number of methods have been proposed to address the challenging text detection problem[1−8]. Among them, the most popular stream focuses on how to build a text model which can robustly take advantage of the text geometries, such as the letter size, the aspect ratio of letter, the distance/size/layout relationship in the letters when forming a word, etc. In this kind of works, a representative one is the recent work of [2] which takes advantage of these geometric properties based on the extracted stroke width. It has been shown that the text detection can achieve the state-of-the-art classification result by this way. In spite of the achievement made by this kind of methods based on text geometries, we argue in this paper that only utilizing some kinds of geometric properties with simple statistics is not enough, it may fail to capture the very delicate details due to the largely varying visual text appearance in natural scenes. Fig. 1 shows some typical false alarms from the text detection system based on the work of [2]. One can observe that the falsely detected texts and the true texts differ considerably in terms of visual appearance.

We propose a new two-phase approach to integrate both the text geometry and the visual appearance in this paper. In the first phase of the proposed two-phase text detection approach, geometric rules based on stroke width extraction are used as in [2] to detect the possible text regions. Specifically, an enhanced stroke width transform (SWT) called robust SWT is proposed. The robust SWT takes extra considerations on the cross of two strokes and the continuousness of the letter border.

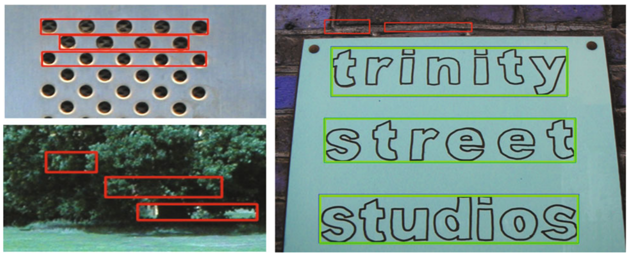


Fig. 1 Detection results from a text detection system based on the work of [2]

In the second phase, classification is performed on the regions passing the first phase′s justification based on geometric rules to further verify the text image region. And the classifier used in this phase is based on the visual appearance features. To learn a more robust classifier, a novel learning method called double soft multiple kernel learning (DS-MKL) is proposed to learn from multiple visual ap-

Regular paper
Special Issue on Massive Visual Computing
Manuscript received January 15, 2014; accepted June 20, 2014
This work was supported by National Natural Science Foundation of China (Nos. 61300163, 61125106 and 61300162) and Jiangsu Key Laboratory of Big Data Analysis Technology.

pearance features. The proposed DS-MKL can model the influence of the noisy base kernels and thus learn a more robust classifier. A block-wise coordinate descent algorithm with an analytical solution is designed to obtain the kernel combination coefficients.

Extensive experiments are conducted on the broadly recognized benchmark ICDAR2005 competition dataset[9]. The experimental results demonstrate the effectiveness of the proposed method, which outperforms the state-of-the-art in all evaluation criteria.

2 Related works

There have been a large number of methods dealing with text detection in natural images and videos[1−8]. Two comprehensive surveys can be found in [10, 11]. The existing approaches to text detection can be roughly divided into two categories: rule based methods and machine learning methods. The rule based method[2−3,12] first takes advantage of the letter-specific distribution rules, such as color uniformity, gradient heap, stroke width, to extract the candidate letter areas. Then the properties of candidate letter size, aspect ratio, variance and so on are used to remove the false candidate letter areas. Finally, the rules based on spatial layout of multiple candidate letters, size differences between letters, and similarities between letters are utilized to remove the false text areas. The machine learning method[7,13−15] treats text as one category and non-text as another category. Based on the classifier learned from text sample images and non-text sample images with some features, such as local intensities, filter responses and wavelet coefficients, classification is performed on the input image to detect the text areas. This kind of methods needs to scan all the windows of different locations and scales to find the text areas. Our method is different from all these methods. In the proposed two-phase text detection method, we take the rule based method as a pre-filter, and take the machine learning method for further verification. More importantly, we provide a robust stroke width transform operator which can enhance the previous stroke width transform in [2]. And we also propose a novel learning method of DS-MKL.

The proposed DS-MKL is most related to multiple kernel learning. The pioneering work for kernel learning was proposed by [16] to train the support vector machine (SVM) classifier and learn the kernel matrix simultaneously, which is known as multiple kernel learning (MKL). Since the objective function proposed in [16] has a simplex constraint for the kernel coefficients, it is also known as ℓ_1-MKL. While the development of efficient algorithms for ℓ_1-MKL is a major research topic in the literature[16−19], Cortes et al.[20, 21] recently pointed out that ℓ_1-MKL cannot even achieve better prediction performance compared with simple baselines for some real world applications. To address this problem, a non-sparse MKL[20, 21] was proposed. In [20], the ℓ_2–norm constraint was proposed to replace the simplex constraint, and it was further extended to the ℓ_p-norm constraint[21]. A soft margin regularization framework has been introduced to incorporate and explain the different types of regularization terms for MKL[22]. In our work, starting from a novel kernel margin perspective to ℓ_p-MKL and motivated by the soft margin MKL framework[22], we propose a novel DS-MKL formulation by considering the regularization from kernel slack variables. In this way, we can tackle the noisy base kernels and learn a more robust model than the existing ℓ_p-MKL for fusing multiple visual appearance features. The kernel slack variables make the key difference of the proposed formulation from others.

3 Overview of the two-phase text detection approach

The flowchart of the proposed two-phase text detection approach is shown in Fig. 2. The upper part shows the four main steps in the first phase. The lower part shows the three main steps in the second phase. In the following, the steps are described sequentially as in Fig. 1.

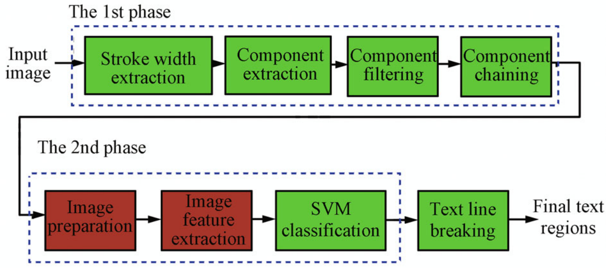


Fig. 2 The flowchart of the proposed two-phase text detection approach

3.1 The first phase

The first phase contains four steps. In the first step, an operator called robust stroke width transform (robust SWT) is conducted on the input image. And this procedure outputs the stroke width image for the input image. For the stroke width image, each pixel value is set to the width of the stroke passing it. The stroke width of the pixel without a stroke on it is set to be infinite. The detailed description of the robust SWT is described in Section 4.

In the second step, components (letter candidates) are found by grouping together the neighboring pixels which have similar stroke widths. A component is composed of a group of pixels that forms a connected area. Two neighboring pixels are grouped together if they have similar stroke widths (In this work, the stroke width ratio of two pixels is restricted to be within the range of $[\frac{1}{3}, 3]$).

The third step utilizes some geometric reasoning rules designed based on letter geometry and some simple statistics to filter out the illegal components. The aspect ratio of each component′s bounding box should be in a reasonable range, which is [0.1, 10] in our experiment. The height of the connected component should be greater than 10 pixels and less than 300 pixels. By this way, we remove those very large and very small letter candidates which normally do not appear in natural scenes. The stroke width variance of each component is also restricted, which is restricted to be less than 5.9 in this work.

The final step is to agglomerate the components into legal chains (a line of text) based on text geometry and some simple statistics. To group the components, legal letter pairs are generated following some rules from all possible pairs. Two components in a legal pair should have similar stoke widths and color variances. The distance between two components in a legal pair should not be greater than 3 times the maximum of the bounding box width of the two components. The legal pairs are then aggregated into chains in a recursive fashion. At first, all legal pairs form the original set of chains. Then two chains sharing the same component are considered to be combined. Finally, the components in the final chain are required to have a near liner form. A simple method as in [9] is used to realize these two targets.

The four steps in the first phase are performed two times to handle both the bright text on dark background and dark text on bright background, one along the gradient direction and the other along the inverse direction. The results of the two passes are fused to form the first-phase detection results.

3.2 The second phase

In the second phase, images are divided into two categories: positive sample images with texts and negative sample images without texts. A classifier is learned to classify these sample images. Then binary classification is performed on the false alarm images survived from the first phase to make the final decision. The second phase can be divided into three steps.

The first step is to prepare the images for classification. All the images for classification are required to have a uniform height of 70 pixels. Each detected rectangle from the first phase is used to generate a final image for classification. First, the detected rectangle from the first phase is enlarged with a fixed center position to include some background. Then the image patch containing in the enlarged rectangle is extracted and re-scaled to generate the image for classification with a height of 70 pixels. In the experimental section, we will describe the preparation of these samples in detail. By resizing, the classification avoids dealing with multi-scale texts.

The second step in the second phase extracts image features of visual appearance. In this paper, we use appearance features which perform well in scene classification, texture classification and object recognition. To be exact, GIST[23], local binary pattern (LBP) histogram[24], bag of words (BoW) feature of scale-invariant feature transform (SIFT)[25] and bag of words feature of structural similarity (SSIM)[26] are used.

In the last step of the second phase, an SVM classifier learned with DS-MKL is applied with visual appearance features to classify the images. The DS-MKL is described in detail in Section 5.

Finally, the detected text region may be composed of several visual words. To separating these words, a word breaking method is provided. The method is realized as following: First, the distance of two components is calculated. Then the text region is broken into word candidates by the saliency of the distance differences between two adjacent component pairs.

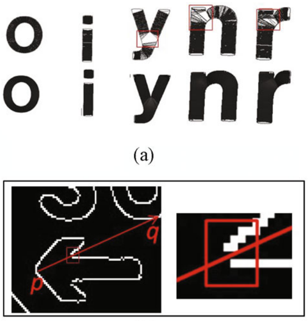
4 Robust stroke width transform

Before introducing the proposed robust SWT, we first give a brief introduction to SWT[2]. SWT takes advantage of the consistent stroke width in the letters to recover regions that are likely to contain letters. To extract stroke width, edges are first extracted from the input image. Then SWT searches for edge pixel pairs which have nearly opposite directions to recover the stroke pixels. To be exact, for each edge pixel p, the SWT operator tries to find an associated stroke by searching along the ray of the gradient direction ($r = p+n\times d_p$, $n > 0$) until another edge pixel q is found. The gradient direction of q is required to be roughly opposite to d_p. In the work of SWT[2], d_p is required to be in the range of $d_q \pm \frac{\pi}{6}$. Then stroke width for each pixel on the ray of $[p, q]$ is assigned the value of $||p-q||$. If the pixel already has a stroke width, the smaller value between the new one and the old one is selected.

This first problem with SWT is: Searching from one pixel of p on one border of the stroke, another pixel of q on another border of the stroke may miss to be found. The account for this is that the opposition constraint of p and q is too strict ($d_q - \frac{\pi}{6} < d_q < d_q + \frac{\pi}{6}$). And this makes the recovering of stroke in some special area fail, such as the sharing area between two crossing strokes. In the upper part of Fig. 3 (a), three areas with badly recovered stroke width by SWT are shown. The three areas are marked with three rectangles. So a reasonably large range of $d_q \pm \frac{\pi}{2}$ is proposed. In the lower part of Fig. 3 (a), the stroke width extraction results by the robust SWT of the same letter are shown. As we can see, the sharing areas are recovered with stroke width successfully with the modified version of SWT. The second problem with SWT is that during the searching along the ray, the search sometimes arrives at a wrong point because of the breaking of the boundary edge. One real example of this case is shown in Fig. 3 (b) with an edge image. The failure happens on the edge pixel of p. p is on the stroke boundary of a printed arrow. The search from p is illustrated with the red arrowed line. One can see that the red arrowed line goes beyond the stroke boundary on another side of the printed arrow (the rectangle area in Fig. 3 (b)), and lands on a wrong edge point of q which is not supposed to land on. So we propose to connect the small local broken edge points. Specifically, a non-edge pixel is set to be an edge pixel if another 2 edge pixels can be found in its neighboring 3×3 area. Considering noisy edge points, the newly generated edge points are only used to end a search and they are not used to start a search. The third problem with SWT is that it only searches one time for each edge pixel p, while the gradient of pixel p is always affected by noise in real application. As a result, the pointed direction of p is disturbed. So multiple times of searches in the neighboring direction of p is tried in the robust SWT. In our experiments, the directions within the range of $d_p \pm \frac{\pi}{2}$ are all used for searching q.

5 Double soft multiple kernel learning

In the following, we denote $||\boldsymbol{d}||_p = (\sum_{m=1}^M d_m^p)^{\frac{1}{p}}$ as the ℓ_p–norm of the M dimensional vector $\boldsymbol{d}$. We also use the superscript "$'$" to indicate the transpose of a vector, and denote the element-wise product of two vectors $\boldsymbol{\alpha}$ and $\boldsymbol{y}$ as $\boldsymbol{\alpha} \odot \boldsymbol{y} = [\alpha_1 y_1, \cdots, \alpha_l y_l]'$. Moreover, $\mathbf{1} \in \mathbf{R}^l$ denotes an l dimensional vector with all elements of 1, and inequality such as $\boldsymbol{d} = [d_1, \cdots, d_M]' \geqslant 0$ signifies that $d_m \geqslant 0$ for $m = 1, \cdots, M$. To simplify notation, we use $\forall\, i$ and $\forall\, m$ to mean the value of i from 1 to l and the value of m from 1 to M, respectively.



(a)

(b)

Fig. 3 The upper row and the lower row of (a) show the stoke width extracted by SWT and the proposed robust SWT, respectively; (b) Left: One failure case of SWT in which the search (the arrowed red line) goes beyond the correct border edge point due to broken edge line (red rectangle); (b) Right: The close-up view of the image around the red rectangle in (b) left.

5.1 A hard kernel margin perspective to multiple kernel learning

Let us denote the training samples as $\{\boldsymbol{x}_i|_{i=1}^l\}$ and the corresponding labels as $\{\boldsymbol{y}_i|_{i=1}^l\}$ with $\boldsymbol{y}_i \in \{-1, +1\}$. The multiple kernel learning[16] was proposed to learn the kernel matrix and the SVM classifier from a set of M pre-defined base kernels $\{\boldsymbol{K}_1, \cdots, \boldsymbol{K}_M\}$, $\boldsymbol{K}_m(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi_m(\boldsymbol{x}_i)'\varphi_m(\boldsymbol{x}_j)$ is a kernel constructed by using mapping $\varphi_m(\cdot)$ from the extracted features. Given the input sample $\boldsymbol{x}$, the decision function $f(\boldsymbol{x})$ of the classifier can be defined as $f(\boldsymbol{x}) = \sum_{m=1}^M \boldsymbol{w}'_m \varphi_m(\boldsymbol{x}) + b$, where $\boldsymbol{w}_m$ is the hyperplane and b is the bias term. The primal objective function for ℓ_p-MKL[21] has been proposed as the structural risk minimization problem as

$$\min_{\boldsymbol{d}\in M, \boldsymbol{w}_m, \xi_i, b} \frac{1}{2}\sum_{m=1}^M \frac{\|\boldsymbol{w}_m\|^2}{d_m} + C\sum_{i=1}^l \xi_i$$

$$\text{s.t.}\ \ y_i\left(\sum_{m=1}^M \boldsymbol{w}'_m \varphi_m(\boldsymbol{x}_i) + b\right) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ \forall i \qquad (1)$$

where $\boldsymbol{M} = \{\boldsymbol{d}|\boldsymbol{d} \geqslant 0, (\sum_{m=1}^M d_m^p)^{\frac{1}{p}} \leqslant 1\}$ is the domain for the kernel combination coefficients $\boldsymbol{d} = [d_1, \cdots, d_M]'$, ξ_i is the slack variable for each sample and C is the SVM regularization parameter.

This primal objective function for ℓ_p-MKL has been commonly discussed in the [20, 21]. However, the Lagrangian dual has not been studied yet. In this part, we first give its Lagrangian dual form as

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\lambda}, \gamma} -\sum_{i=1}^n \alpha_i + \gamma$$

$$\text{s.t.}\ \ \frac{1}{2}(\boldsymbol{\alpha} \odot \boldsymbol{y})'\boldsymbol{K}_m(\boldsymbol{\alpha} \odot \boldsymbol{y}) = \lambda_m, \quad \forall m,$$

$$0 \leqslant \boldsymbol{\alpha} \leqslant C, \quad \boldsymbol{y}'\boldsymbol{\alpha} = 0,$$

$$\left(\sum_{m=1}^M \boldsymbol{\lambda}_m^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} = \gamma \qquad (2)$$

where $\boldsymbol{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n]'$ is the label vector, $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_n]'$ is the SVM dual variable vector, and $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_M]'$.

Different from the primal form, the Lagrangian dual formulation in (2) can be easily interpreted from a kernel margin perspective for the essential property of the multiple kernel learning. If we regard the quadratic term $\frac{1}{2}(\boldsymbol{\alpha} \odot \boldsymbol{y})'\boldsymbol{K}_m(\boldsymbol{\alpha} \odot \boldsymbol{y})$ as the "kernel margin", we can observe that each kernel margin term is associated with a "kernel margin variable" λ_m, which further forms the global kernel margin γ in an ℓ_q–norm manner with $q = \frac{p}{p-1}$. We can observe that the quadratic term strictly equals to λ_m, and there is no error allowance from each of the base kernels, thus we regard the formulation in (2) as a hard kernel margin perspective for multiple kernel learning. In this way, we conjecture that this formulation may be sensitive to noisy base kernels.

5.2 Double soft multiple kernel learning

The slack variable has been successfully introduced for each sample in soft margin SVM[27] to tackle the noisy data which is not considered in hard margin SVM[28]. Similarly, to overcome the hard kernel margin defect, we propose a new objective function called double soft multiple kernel learning to learn a robust classifier by introducing the so-called kernel slack variables for the base kernels. Specifically, we can introduce one slack variable ς_m which models the kernel margin error for each of the base kernels. And with the hinge loss for these kernel slack variables, we propose the new double soft MKL as

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\lambda}, \varsigma, \gamma} -\sum_{i=1}^n \alpha_i + \gamma + \theta\sum_{m=1}^M \varsigma_m$$

$$\text{s.t.}\ \ \frac{1}{2}(\boldsymbol{\alpha} \odot \boldsymbol{y})'\boldsymbol{K}_m(\boldsymbol{\alpha} \odot \boldsymbol{y}) \leqslant \lambda_m + \varsigma_m, \quad \varsigma_m \geqslant 0, \quad \forall m,$$

$$0 \leqslant \boldsymbol{\alpha} \leqslant C, \quad \boldsymbol{y}'\boldsymbol{\alpha} = 0,$$

$$\left(\sum_{m=1}^M \lambda_m^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} = \gamma \qquad (3)$$

where γ is the global margin, $\varsigma = [\varsigma_1, \cdots, \varsigma_M]'$ is the kernel slack variable vector, and θ is the regularization parameter for the loss from each of the base kernels.

This kind of improvement is analogous to the change from the hard margin SVM[28] to hinge loss soft margin SVM[27] in terms of introducing slack variables. The soft margin SVM introduces one slack variable ξ_i for each training instance, while our proposed DS-MKL introduces one slack variable ς_m for each of the base kernels. Thus, our new model will be more robust to tackle the noisy base kernels and learn a classifier with better generalization ability when compared with the ℓ_p-MKL. And it will be demonstrated in the experimental part.

5.3 Solution to DS-MKL

5.3.1 An equivalent form of DS-MKL

The problem in (3) is difficult to optimize due to its quadratic constraints. Fortunately, we can have its equivalent form as shown in the following proposition.

Proposition 1. The problem in (3) is equivalent to the optimization problem as

$$\begin{aligned}
&\min_{d_m,\boldsymbol{w}_m,\xi_i,b} \frac{1}{2}\sum_{m=1}^{M}\frac{||\boldsymbol{w}_m||^2}{d_m}+C\sum_{i=1}^{l}\xi_i\\
&\text{s.t.}\ \ y_i\left(\sum_{m=1}^{M}\boldsymbol{w}_m'\varphi_m(x_i)+b\right)\geqslant 1-\xi_i,\quad \xi_i\geqslant 0,\quad \forall i,\\
&\qquad\left(\sum_{m=1}^{M}d_m^p\right)^{\frac{1}{p}}\leqslant 1,\\
&\qquad 0\leqslant d_m\leqslant\theta,\forall m. \qquad (4)
\end{aligned}$$

A global solution for (4) is guaranteed due to the convex objective function as well as the convex constraints. To solve this problem, we follow the block-wise coordinate descent procedure for ℓ_p-MKL[21, 29], composite kernel learning (CKL)[30] and soft margin MKL[22], and optimize two subproblems with respective to the two sets of variables $\{\boldsymbol{w}_m,\xi_i,b\}$ and $\{\boldsymbol{d}\}$ alternately. Note that, due to the additional box constraints introduced from soft margin regularization for the base kernels, the subproblem for updating $\boldsymbol{d}$ becomes much more difficult than the ones in [21, 29, 30].

5.3.2 Updating SVM variables with fixed $\boldsymbol{d}$

With a fixed $\boldsymbol{d}$, we write the dual of (4) by introducing the non-negative Lagrangian multipliers $\alpha_i(1<i<l)$ as

$$\max_{\boldsymbol{\alpha}\in\mathbf{A}}\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{m=1}^{M}d_m(\boldsymbol{\alpha}\odot\boldsymbol{y})'\boldsymbol{K}_m(\boldsymbol{\alpha}\odot\boldsymbol{y}) \qquad (5)$$

which is a quadratic programming (QP) problem with $A=\{\boldsymbol{\alpha}|0\leqslant\boldsymbol{\alpha}\leqslant C,y'\boldsymbol{\alpha}=0\}$, and can be efficiently solved by any prevailing QP solver. Then, the primal variables $\{\boldsymbol{w}_m,\xi_i,b\}$ can be recovered accordingly. For instance, the ℓ_2–norm for $\boldsymbol{w}_m$ can be expressed as

$$||\boldsymbol{w}_m||=d_m\sqrt{(\boldsymbol{\alpha}\odot\boldsymbol{y})'\boldsymbol{K}_m(\boldsymbol{\alpha}\odot\boldsymbol{y})}. \qquad (6)$$

5.3.3 Updating $\boldsymbol{d}$ with fixed SVM variables

For updating $\boldsymbol{d}$ with fixed SVM variables, the subproblem can be formulated as

$$\begin{aligned}
&\min_{\boldsymbol{d}}\ \frac{1}{2}\sum_{m=1}^{M}\frac{||\boldsymbol{w}_m||^2}{d_m}\\
&\quad(\sum_{m=1}^{M}d_m^p)^{\frac{1}{p}}\leqslant 1,\quad 0\leqslant d_m\leqslant\theta,\quad\forall m. \qquad (7)
\end{aligned}$$

Due to the additional upper bound θ, the existing optimization techniques[21, 29, 30] cannot be directly utilized. Inspired by [31] for simplex projection, the problem in (7) can be solved analytically. Before introducing our solution, let us denote ω as the number of elements, whose value strictly equals to θ in the optimal solution for $\boldsymbol{d}$. The closed-form solution for (7) is obtained as in the following proposition.

Proposition 2. If $\boldsymbol{w}_m$ are sorted such that $||\boldsymbol{w}_1||\geqslant||\boldsymbol{w}_2||\geqslant\cdots\geqslant||\boldsymbol{w}_M||$, then the optimal solution for subproblem (7) is given as

$$d_m=\begin{cases}\theta, & \text{if}\quad m\leqslant\omega\\ \dfrac{(1-\omega\theta^p)||\boldsymbol{w}_m||^{\frac{2}{p+1}}}{\left(\displaystyle\sum_{s=\omega+1}^{M}||\boldsymbol{w}_s||^{\frac{2p}{p+1}}\right)^{\frac{1}{p}}}, & \text{if}\quad m>\omega.\end{cases} \qquad (8)$$

The proof can be done by using the Lagrangian method and is omitted due to the space limitation. The number of elements whose values strictly equal to θ in the optimal solution for $\boldsymbol{d}$ can be obtained by the following lemma.

Lemma 3. Let $\boldsymbol{d}^*$ be the optimal solution to problem (7), and suppose that $||\boldsymbol{w}_1||\geqslant||\boldsymbol{w}_2||\geqslant\cdots\geqslant||\boldsymbol{w}_M||$. Then, ω, the number of elements whose value strictly equals to θ in $\boldsymbol{d}^*$ is

$$\min\left\{s\in\{0,1,\cdots,M-1\}\middle|\frac{||w_{s+1}||^{\frac{2}{p+1}}(1-s\theta^p)}{\left(\sum_{m=s+1}^{M}||w_s||^{\frac{2p}{p+1}}\right)^{\frac{1}{p}}}<\theta\right\}.$$

The proof can be done by using contradiction and is omitted here. This lemma shows that ω can be obtained from a sorting algorithm, and then the optimum solution for $\boldsymbol{d}$ can be obtained analytically according to (8).

5.3.4 The whole optimization procedure

According to the above derivations, we can easily develop the optimization procedure for the DS-MKL as shown in Algorithm 1. With the optimized $\boldsymbol{d}$, $\boldsymbol{\alpha}$ and b, the final decision function is obtained as

$$f(\chi)=\sum_{m=1}^{M}d_m\sum_{i=1}^{l}\alpha_iy_i\boldsymbol{K}_m(\boldsymbol{x},\boldsymbol{x}_i)+b. \qquad (9)$$

Algorithm 1. Procedure of the block-wise coordinate descent algorithm for solving DS-MKL.

```
1) Initialize $\boldsymbol{d}^1$.
2) $t = 1$.
3) While stop criteria are not reached do
4)     Get $\boldsymbol{\alpha}^t$ by solving subproblem (5) using standard
       QP solver with $\boldsymbol{d}^t$.
5)     Calculate $||\boldsymbol{w}_m||$ by using (6) and update $\boldsymbol{d}^{t+1}$ by
```

```
        using (8).
6)  $t = t + 1$.
7) **End while**.
```

6 Experiments

6.1 Experimental setup

The proposed method is evaluated on the broadly recognized text detection ICDAR2005 robust reading competition dataset[32]. The dataset has been used in two text localization competitions: ICDAR2003[9] and ICDAR 2005[32]. It is still the most widely used benchmark for text detection/localization in natural scene. Overall, the ICDAR2005 robust reading competition dataset contains two subsets. One includes 250 images with 1156 annotated texts for training. The other includes 249 images with 1107 annotated texts for testing. All the images from both subsets are full-color images with a minimal size of 307 × 93 and maximal size of 1280 × 960. The texts are annotated with their corresponding bounding boxes. In our experiment, all the 249 images for testing are used for evaluation as the way they were used in [2, 32]. The evaluation protocol is the same as the one described in [2, 32].

The positive training set for DS-MKL is totally generated from the ICDAR2005 training set. The annotated bounding boxes are used for generating the samples. Original bounding boxes with different lengths of sentences or texts are all included to generate the training set. Besides, sub-areas from the overall bounding boxes are extracted to generate more training samples because they are still text images. The sub-areas are required to be with a width larger than 5 times the height. In total, 1312 text image regions are collected.

For the negative training set, the false alarms survived from the first phase are collected. To obtain more negative training samples, extra negative samples are collected from the University of Illinois at Urbana-Champaign (UIUC) sports dataset[32], which contains 1586 images with various scenes. Note that the UIUC sports dataset does not contain common images with the ICDAR2005 test set. In total, 6660 non-text images are collected.

The bounding boxes of the annotated texts and the collected non-text image regions along with a certain margin are resized to generate the final training samples. Some background around the texts is included in the training samples by the margins to capture the difference between the text and the background. The height of the final training sample has 70 pixels. The bounding box is fitted in the center position of the training sample with a height of 50 pixels. And the margins for four sides, i.e., the top side, the bottom side, the left side, and the right side have all 10 pixels. Fig. 4 shows some typical positive and negative training samples.

For the visual appearance features, GIST[23], LBP histogram[24], bag of words features based on SIFT[25] and bag of words features based on SSIM[26] are extracted to capture the visual appearances of the texts. In BoW feature extraction, K-means is employed for building dictionaries. The dictionary size is set to 1024 empirically. Localized soft assignment[33] is used for quantization. Max pooling is applied on a two level spatial pyramid[22] of 1×1 and 2×2.

For learning, ℓ_p-MKL and DS-MKL are implemented using the libsvm package[34]. A total number of four linear kernels are generated from the four types of features as the base kernels. The SVM regularization parameter C is set to 10 throughout the experiments. For both ℓ_p-MKL and DS–MKL, p is fixed to be 1.25 empirically.

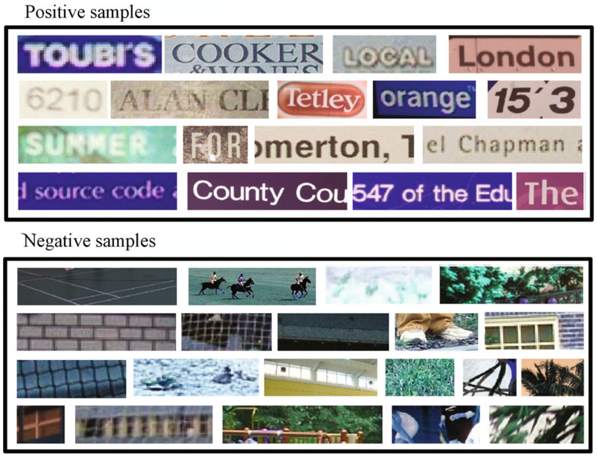


Fig. 4 Some typical positive and negative training samples

6.2 Investigation of the two-phase text detection

Firstly, we evaluate the first phase text detection in its ability to detect the texts. Because the recall rate accounts for the ability to detect the text areas, we report the recall rate to demonstrate the effectiveness of the first phase text detection. Compared with the baseline method in [2] which achieved a recall rate of 63%, the proposed first phase text detection achieves a quite high recall rate of 70%, which is 7% higher.

Secondly, we evaluate the proposed DS-MKL and the overall two-phase text detection approach. As shown from the constraint for the kernel coefficients in (4), one can observe that θ should be in the range of $\theta \geqslant (\frac{1}{M})^{\frac{1}{p}}$. On one hand, if $\theta = (\frac{1}{M})^{\frac{1}{p}}$, the kernel combination coefficients are enforced to be uniform, this corresponds to assigning equal weights to the base kernels. On the other hand, if $\theta \geqslant 1$, DS-MKL reduces to ℓ_p-MKL, which does not consider the kernel margin error for learning the kernel matrix. Thus, to investigate the effectiveness of our proposed DS-MKL, we can show the results of DS-MKL by varying the new regularization parameter θ. The precisions at the same recall for different DS-MKL regularization parameters θ are provided. The recall rate is set to 69% by adjusting the threshold of the classifier. The differences between precisions of DS-MKL with different θ and the precision of ℓ_p-MKL are shown in Fig. 5. We can see that DS-MKL performs better than both the ℓ_p-MKL and the SVM with uniform kernel weights. The maximal improvement over ℓ_p-MKL is 1.6%. This demonstrates that by introducing the regularization from the kernel slack variables, our proposed DS-MKL can learn a more robust classifier with better generalization ability. One may notice that the highest improvement is achieved at $\theta = 0.64$. In the following, the proposed method is compared with the state-of-the-art algorithms based on this parameter setting.

The precision-recall curve by adjusting the threshold of the classifier in the second phase is showed in Fig. 6. To compare with the state-of-the-art results, Table 1 shows two sets of our results in the precision-recall curve along with the results from the previous methods. Note that F-measure which is a combination of the recall and precision is calculated as

$$\text{F\_measure} = \frac{1}{\frac{\alpha}{\text{Recall}} + \frac{\alpha}{\text{Precision}}} \tag{10}$$

where α is set to be 0.5 as in [2, 32].

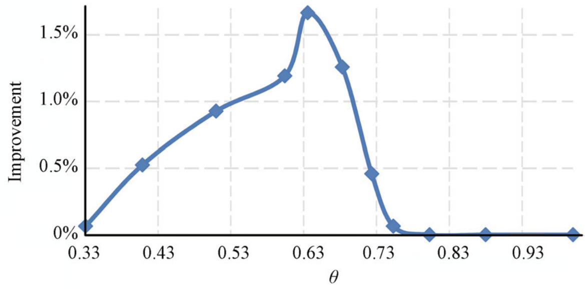


Fig. 5 The precision improvements at the same recall rate with different values for the regularization parameter θ

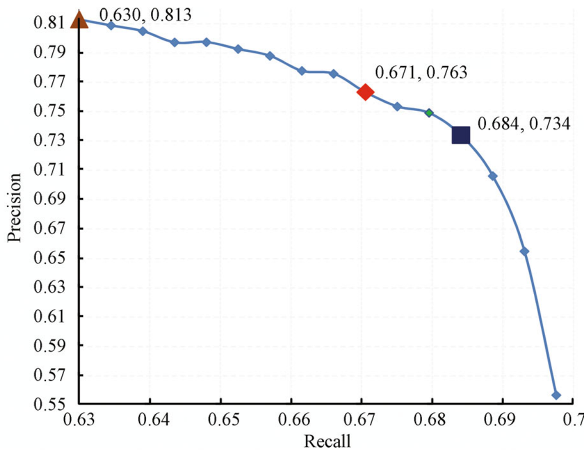


Fig. 6 The precision-recall curve for parameter $\theta = 0.64$

In Table 1, the best recall (67%), precision (73%), and F-measure (67%) from the previous methods are underlined. The two sets of our results are shown in the bottom of Table 1 and termed with "our result 1" and "our result 2". "our result 1" is characterized in that its recall is the same as the best one from the previous methods. "our result 2" is characterized in that its precision is the same as the best one from the previous methods. By this way, the separate gains of Table 1 and the precision can be seen more clearly. From Table 1 ("our result 1"), one can see that at the same recall as the best one from previous methods (0.67 from Becker et al.[32]), the proposed method achieves a much better precision of 76.3%. The precision exceeds the previous best one by 14.3%. And similarly ("our result 2"), at the same precision as the best one from the previous methods (Epshtein et al.[2]), the proposed method achieves a better recall than the best method with an improvement of 8.4%. For the F-measure, the best F-measure from the previous methods is 67% while the best F-measure of the proposed method is 71.4% ("our result 1"), which outperforms the best previous one by 4.4%.

Table 1 Comparison with the state-of-the-art results on the ICDAR2005 dataset

Algorithm	Precision	Recall	F-measure
Yao et al.[4]	0.69	0.66	**0.67**
Epshtein et al.[2]	**0.73**	0.60	0.66
Yi and Tian[35]	0.71	0.62	0.62
Becker et al.[32]	0.62	**0.67**	0.62
Chen and Yuille[13]	0.60	0.60	0.58
Zhu et al.[32]	0.33	0.40	0.33
Kim et al.[32]	0.22	0.28	0.22
Ezaki et al.[32]	0.18	0.36	0.22
Our result 1	**0.763**	0.670	**0.714**
Our result 2	0.734	**0.684**	0.708

To get a direct sense on the text detection results, some of the detected texts are shown in Fig. 7. In Fig. 7, the detected texts regions are boxed with blue rectangles.

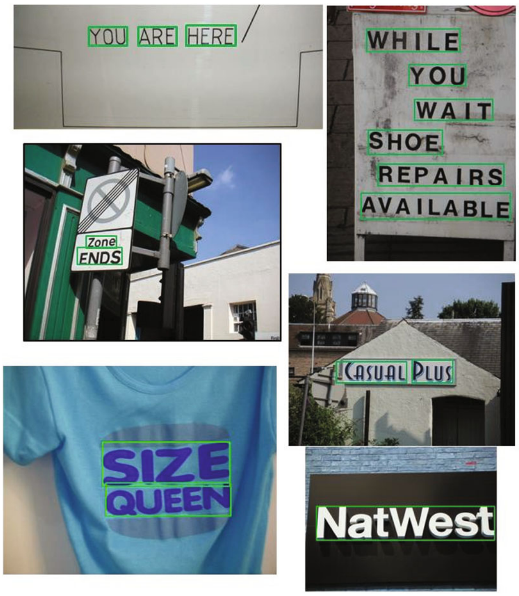


Fig. 7 Some examples of the detected texts

7 Conclusions

The paper proposes to incorporate geometric rules with visual appearance for robust text detection in natural scene. A two-phase approach is designed to take advantage of two kinds of information wisely. Two novel effective techniques are proposed for stroke width extraction classifier learning from multiple types of visual features. Extensive experiments are conducted on broadly accepted benchmark. The experimental results demonstrate the effectiveness of our proposed method. Specifically, the proposed method outperforms the state-of-the-art counterparts by an improvement of 4.4% in terms of F-measure.

References

[1] G. Sahoo, T. Kumar, B. L. Raina, C. M. Bhatia. Text extraction and enhancement of binary images using cellular automata. *International Journal of Automation and Computing*, vol. 6, no. 3, pp. 254–260, 2009.

[2] B. Epshtein, E. Ofek, Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, San Francisco, USA, pp. 2963–2970, 2010.

[3] L. Neumann, J. Matas. A method for text localization and recognition in real-world images. In *Proceedings of the 10th Asian Conference on Computer Vision*, Lecture Notes in Corputer Science, vol. 6494, Springer, Queenstown, New Zealand, pp. 770–783, 2010.

[4] C. Yao, X. Bai, W. Liu, Y. Ma, Z. Tu. Detecting texts of arbitrary orientations in natural images. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Providence, USA, pp. 1083–1090, 2012.

[5] Y. C. Wei, C. H. Lin. A robust video text detection approach using SVM. *Expert Systems with Applications*, vol. 39, no. 12, pp. 10832–10840, 2012.

[6] Y. Y. Qu, W. M. Liao, S. Lu, S. J. Wu. Hierarchical text detection: From word level to character level. In *Proceedings of the 19th International Conference on Advances in Multimedia Modeling, Lecture Notes in Computer Science*, Springer, Huangshan, China, vol. 7733 pp. 24–35, 2013.

[7] V. N. M. Aradhya, M. S. Pavithra. An application of K-means clustering for improving video text detection. In *Proceedings of International Symposium on Intelligent Informatics, Advances in Intelligent Systems and Computer*, Springer, Channai, India, vol. 182, pp. 41–47, 2013.

[8] C. Z. Shi, C. H. Wang, B. H. Xiao, Y. Zhang, S. Gao. Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recognition Letters*, vol. 34, no. 2, pp. 107–116, 2013.

[9] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young. ICDAR 2003 robust reading competitions. In *Proceedings of the 7th International Conference on Document Analysis and Recognition*, IEEE, Edinburgh, Scotland, pp. 682–687, 2003.

[10] J. Liang, D. Doermann, H. P. Li. Camera-based analysis of text and documents: A survey. *International Journal of Document Analysis and Recognition*, vol. 7, no. 2–3, pp. 83–104, 2005.

[11] H. G. Zhang, K. Zhao, Y. Z. Song, J. Guo. Text extraction from natural scene image: A survey. *Neurocomputing*, vol. 122, pp. 310–323, 2013.

[12] A. K. Jain, B. Yu. Automatic text location in images and video frames. *Pattern Recognition*, vol. 31, no. 12, pp. 2055–2076, 1998.

[13] X. R. Chen, A. L. Yuille. Detecting and reading text in natural scenes. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Washington DC, USA, pp. 366–373, 2004.

[14] L. Neumann, R. Ewerth, B. Freisleben. Text detection in images based on unsupervised classification of high frequency wavelet coefficients. In *Proceedings of International Conference on Pattern Recognition*, IEEE, Cambridge, England, pp. 425–428, 2004.

[15] L. Neumann, J. Matas. Real-time scene text localization and recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Providence, USA, pp. 3538–3545, 2012.

[16] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[17] F. R. Bach, G. R. G. Lanckriet, M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, ACM, Banff, Alberta, Canada, 2004.

[18] S. Sonnenburg, G. Rätsch, C. Schäfer, B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[19] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[20] C. Cortes, M. Mohri, A. Rostamizadeh. L_2 regularization for learning kernels. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, AUAI Press, Arlington, Virginia, USA, pp. 109–116, 2009.

[21] M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien. L_p-norm multiple kernel learning. *Journal of Machine Learning Research*, vol. 12, pp. 953–997, 2011.

[22] X. Xu, I. W. Tsang, D. Xu. Soft margin multiple kernel learning. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 749–761, 2013.

[23] J. X. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, San Francisco, USA, pp. 3485–3492, 2010.

[24] T. Ojala, M. Pietikainen, T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[26] E. Shechtman, M. Irani. Matching local self-similarities across images and videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Minneapolis, USA, pp. 1–8, 2007.

[27] C. Cortes, V. Vapnik. Support-vector networks. *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[28] B. E. Boser, I. M. Guyon, V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, ACM, Pittsburgh, PA, USA, pp. 144–152, 1992.

[29] Z. L. Xu, R. Jin, H. Q. Yang, I. King, M. R. Lyu. Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th International Conference on Machine Learning*, Omnipress, Haifa, Israel, pp. 1175–1182, 2010.

[30] M. Szafranski, Y. Grandvalet, A. Rakotomamonjy. Composite kernel learning. *Machine Learning*, vol. 79, no. 1–2, pp. 73–103, 2010.

[31] S. Shalev-Shwartz, Y. Singer. Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, vol. 7, pp. 1567–1599, 2006.

[32] S. M. Lucas. Text locating competition results. In *Proceedings of the 8th International Conference on Document Analysis and Recognition*, IEEE, Seoul, Korea, pp. 80–85, 2005.

[33] S. Y. Yan, X. X. Xu, D. Xu, S. Lin, X. L. Li. Beyond spatial pyramids: A new feature extraction framework with dense spatial sampling for image classification. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 464–478, 2012.

[34] C. C. Chang, C. J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, Article 27, 2011.

[35] C. Yi, Y. L. Tian. Text string detection from natural scenes by structure-based partition and grouping. *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2594–2605, 2011.

Sheng-Ye Yan received his Ph. D. degree from the Institute of Computing Technology of the Chinese Academy of Sciences, China in 2009. He is currently a professor with the School of Automation and Control, Nanjing University of Information Science and Technology, China.

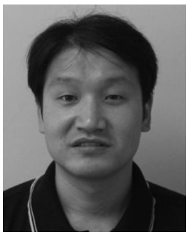
His research interests include computer vision, pattern recognition, machine learning, and their applications in imaging.

E-mail: shengye.yan@gmail.com (Corresponding author)

Xin-Xing Xu received his B. Eng. degree from the University of Science and Technology of China, China in 2009. He is currently a Ph. D. candidate with the School of Computer Engineering, Nanyang Technological University, Singapore.

His research interests include multiple kernel learning as well as image and video understanding.

E-mail: xuxi0006@ntu.edu.sg



Qing-Shan Liu received the M. Sc. degree from South East University, China in 2000, and received the Ph. D. degree from National Laboratory of Pattern Recognition, Chinese Academy of Sciences, China in 2003. He received the president scholarship of Chinese Academy of Sciences in 2003. From 2004 to 2005, he worked as an associate researcher at the Multimedia Laboratory in Chinese University of Hong Kong, China. He used to work as an assistant research professor at Rutgers University. Now, he is a professor at Nanjing University of Information Science and Technology, China. He has published more than 80 papers in journals and conferences including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE International Conference on Computer Vision*, and *IEEE Conference on Computer Vision and Pattern Recognition*. He is an editorial board member of *Neurocomputing* and *Journal of Advance in Multimedia*, and he is a guest editor of *IEEE Transaction on Multimedia*, *Computer Vision and Image Understanding*, and *Pattern Recognition Letters*. He is a senior member of the IEEE.

His research interests include face image analysis, graph and hyper-graph based image and video understanding, medical image analysis, and event-based video analysis.

E-mail: qsliu@nuist.edu.cn