Scene Text Deblurring Using Text-Specific Multiscale Dictionaries

Xiaochun Cao, *Senior Member, IEEE*, Wenqi Ren, Wangmeng Zuo, *Member, IEEE*, Xiaojie Guo, *Member, IEEE*, and Hassan Foroosh, *Senior Member, IEEE*

***Abstract*— Texts in natural scenes carry critical semantic clues for understanding images. When capturing natural scene images, especially by handheld cameras, a common artifact, i.e., blur, frequently happens. To improve the visual quality of such images, deblurring techniques are desired, which also play an important role in character recognition and image understanding. In this paper, we study the problem of recovering the clear scene text by exploiting the text field characteristics. A series of text-specific multiscale dictionaries (TMD) and a natural scene dictionary is learned for separately modeling the priors on the text and nontext fields. The TMD-based text field reconstruction helps to deal with the different scales of strings in a blurry image effectively. Furthermore, an adaptive version of nonuniform deblurring method is proposed to efficiently solve the real-world spatially varying problem. Dictionary learning allows more flexible modeling with respect to the text field property, and the combination with the nonuniform method is more appropriate in real situations where blur kernel sizes are depth dependent. Experimental results show that the proposed method achieves the deblurring results with better visual quality than the state-of-the-art methods.**

***Index Terms*— Scene text, multi-scale dictionaries, text localization, non-unifrom deblurring.**

I. Introduction

TEXT is ubiquitous in natural scenes, *e.g.* billboards, signboard, house numbers and movie posters. Characters and strings in natural scene images provide important information for a wide spectrum of applications, such as context retrieval, assistant navigation, aid reading, and scene understanding [1]. In pervasive image acquisition, image blur caused by camera shake frequently happens, which leads to the abrupt degradation of image quality, and thus makes character recognition and image understanding more difficult. To improve the visual quality for such images, the scene text deblurring has received considerable research interests.

Mathematically, by assuming the motion blur is shift invariant, the blurry image can be modeled as:

$$\boldsymbol{B} = \boldsymbol{K} \otimes \boldsymbol{L} + \boldsymbol{N}, \tag{1}$$

where $\boldsymbol{B}$ stands for the blurry observation, $\boldsymbol{L}$ is the latent clear image and $\boldsymbol{N}$ represents the additive white Gaussian noise. In addition, $\boldsymbol{K}$ denotes the blur kernel, and $\otimes$ is the convolution operator. Within this context, the ultimate goal of this paper is to recover the sharp and clean text from the blurry observation $\boldsymbol{B}$. However, the problem of deblurring is heavily ill-posed, since there are infinite solutions to Eq. (1). The challenge of blind deblurring has attracted much attention recently, and various image blind deblurring methods have been proposed [2]–[10] to conquer the task. Although these methods can provide promising results for natural landscape images, they are still unsuitable for text images. One reason is that most of them employ the natural image statistics instead of the characteristics of text fields as the regularizer. Alternatively, [11] proposes a content-aware prior for document image deblurring, but it is not very robust for general blurry natural scene text images. Recently, Cho *et al.* [12] propose an effective text image deblurring method based on three text-specific properties. The method relies on the properties of text regions detected by the stroke width transform (SWT) [13]. However, SWT is designed to detect text in clear images, the accuracy of which may decrease or even fail when applied to blurry images.

In this work, we propose a robust scene text image deblurring method using Text-specific Multi-scale Dictionaries, namely TMD. The overview of our method is shown in Fig. 2. First, by exploring the text and non-text field characteristics, we learn a natural scene dictionary and a series of text-specific multi-scale dictionaries to model the priors on the background scene and text fields respectively. This step is dictionary learning highlighted by the red dashed rectangle. Second, we run the state-of-the-art text localization method [14] to differentiate text fields from non-text ones. Third, based on TMD and the natural dictionary, we construct the dedicated priors for real world text and non-text fields. As a result, we optimize the cost function to estimate the blur kernel and the latent image. The last step will repeat untill the

Manuscript received June 8, 2014; revised October 15, 2014 and December 22, 2014; accepted January 31, 2015. Date of current version February 19, 2015. This work was supported in part by the National Basic Research Program of China under Grant 2013CB329305, in part by the National Natural Science Foundation of China under Grant 61332012, Grant 61402467, and Grant 61422213, and in part by the 100 Talents Programme through the Chinese Academy of Sciences. Wangmeng Zuo was supported in part by the National Science Foundation of China under Contract 61271093. The work of X. Guo was supported in part by the National Natural Science Foundation of China under Grant 61402467 and in part by the Excellent Young Talent Programme through the Institute Information Engineering, Chinese Academy of Sciences. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Javier Mateos.

X. Cao and W. Ren are with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: caoxiaochun@iie.ac.cn; rwq.renwenqi@gmail.com).

W. Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: cswmzuo@gmail.com).

X. Guo is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: xj.max.guo@gmail.com).

H. Foroosh is with the Computational Imaging Laboratory, School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: foroosh@cs.ucf.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2400217

Fig. 1. An example of the scene text deblurring. Top: A real blurry scene image from [2]. Bottom: Close-ups of deblurring results of different text fields of the image using spatially varying method. Note that these two deblurred text fields have different blur kernels as shown in bottom and the result demonstrates that it is improper to estimate a uniform kernel for the entire image. More details can be found in Fig. 13.

blur kernel converge. Note that the text fields after converges in the uniform deblurring step need to be non-uniformly deblurred as shown in Section II-D, where the non-text field will not because we focus on text fields deblurring in this paper.

Our contribution is two-fold. First, due to the variety of text sizes in different text fields of an image, we propose a novel TMD based text field reconstruction method to deal with different scales of strings in a blurry image effectively. The proposed model can automatically learned multi-scale dictionaries from the training dataset of text fields, which is more flexible and do not require complex filtering strategies for fitting the text properties as in [12]. Second, we use a piece-wise scheme as in [10] to estimate the multiple kernels on different text fields selected by a text localization method automatically. However, the main difference between our method and [10] is that [10] needs to partition whole images into multiple regions, while our method only considers the text fields because we focus on scene text deblurring. Such implementation helps to recover a more clear text in real cases, as illustrated in Fig. 1.

A. Nature Image Deblurring

Blind deblurring has received considerable interests from the communities of image processing, computer vision and graphics. Most of the existing methods achieve good results by designing various priors for optimizing:

$$\arg\min_{\boldsymbol{L},\boldsymbol{K}} \|\boldsymbol{B}-\boldsymbol{K}\otimes\boldsymbol{L}\|^2+\rho(\boldsymbol{K})+\rho(\boldsymbol{L}), \tag{2}$$

where the first term is to suppress the reconstruction error, *i.e.* the restored image should be consistent with the observation with respect to the estimated blur kernel $\boldsymbol{K}$. $\rho(\boldsymbol{K})$ is a regularization term, typically a ℓ_1-norm [15] or ℓ_2-norm [8] penalty. The deblurred result depends largely on the (specifically) designed prior knowledge $\rho(\boldsymbol{L})$ of the latent image.

There are several classic methods to design the latent image priors [4], [5], [9], [15]–[18]. Yuan *et al.* [18] use two images for motion deblurring, one of which is noisy but has sharp edges, and the other is motion blurred. However, the assumption of having such image pairs is not always satisfied. Jia *et al.* [16] recover the latent image from the perspective of transparency by assuming that the transparency map of a clear foreground object should be two-tone. This method is limited as it requires to find regions that produce high-quality matting results. Shan *et al.* [5] propose a combination of global and local priors to fit the gradient distribution and eliminate the local ringing effect. Levin *et al.* [4] use a sparse derivative prior to avoid ringing artifacts for the task of deconvolution. [15] adopts an l_1/l_2 regularization scheme to adapt l_1 norm regularization by treating the l_2 norm of image gradients as a weight in iterations. Cai *et al.* [17] propose to remove motion blurring from the image by regularizing the sparsity of both the original image and the motion-blur kernel under tight wavelet frame systems. This work focuses on the deblurring for natural scene text, which utilizes the priors about the text properties to boost the performance.

Another line of researches try to make use of sparsity property of the latent image $\boldsymbol{L}$. Sparse representation has been extensively applied to many ill-posed problems in image processing, such as denoising [19] and restoration [20]. The main idea of sparse representation is that a patch $\boldsymbol{x}\in\Re^n$ extracted from an image $\boldsymbol{X}$ can be described by a linear combination of a few atoms from a dictionary $\boldsymbol{D}\in\Re^{n\times K}$ $(n\ll K)$ learned from the training data. Thanks to the representation power, sparse representation has also been successfully applied to deblur natural images [21], [22]. Therefore, we design the prior of scene text based on sparse representation in this paper for scene text image deblurring. We extend the ordinary dictionary to a series of text-specific multi-scale dictionaries and a natural dictionary. As a result, based on which we design dedicated priors to be more applicable to scene text handling.

B. Text Image Deblurring

Although numerous methods have been proposed for nature image deblurring, there are few models specifically dealing with scene text cases. Qi *et al.* [23] use cepstral domain techniques for identifying the blur parameters, but it can only deal with images with 1D kernels, i.e. the camera is assumed to move along a straight line with a constant acceleration. Su *et al.* [24] estimate blur parameters by the constructed alpha channel map based on specific image characteristics. Nevertheless, this method is designed for document images only. Li *et al.* [25] propose a statistical method to estimate blur kernels from two-tone images. Chen *et al.* [11] advocate a content-aware prior for image deblurring to handle document images.

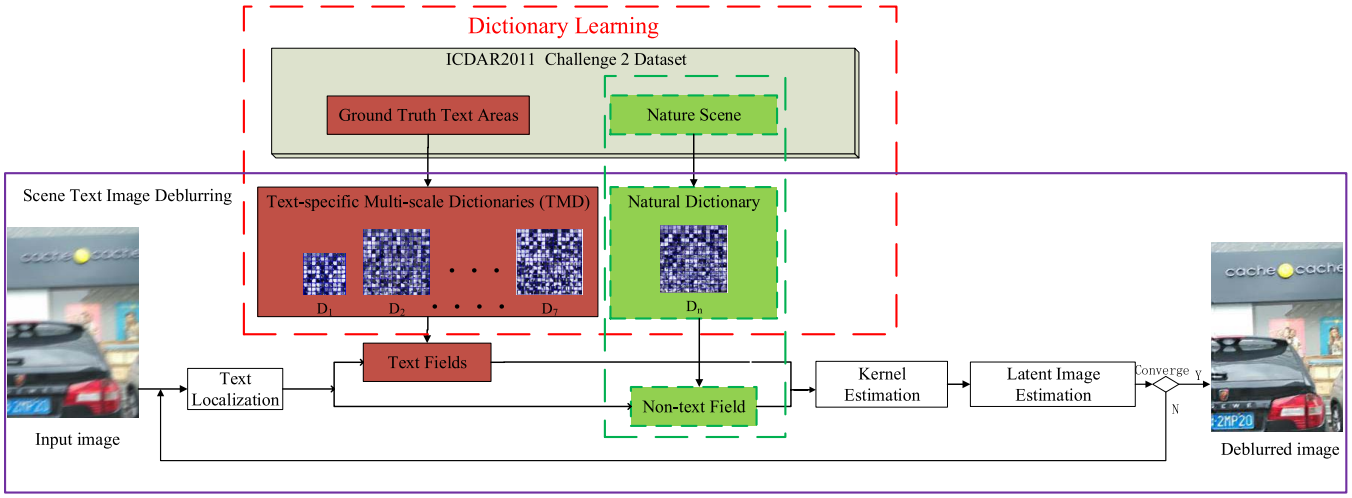


Fig. 2. Overview of our model. Given an input blurry image. The proposed method iteratively estimates the latent scene text image with TMD and a nature scene dictionary until it converges to a stable solution. The three green dashed rectangles are only executed in the uniform stage.

This method uses a text segmentation technique [26] based on thresholding to detect text and then estimates the latent image by a learned intensity relationship between the blurry and the clear document images. However, these two methods are only applicable to two-tone images and are less effective for complicated text images. In other words, they are hard to be directly applied to scene text images. To overcome this challenge, Cho *et al.* [12] propose three text-specific properties based on SWT [13] as the text image priors: 1) text characters have high contrasts against background regions; 2) each character has a near-uniform color; and 3) background gradient values obey natural image statistics. Nevertheless, this method is very sensitive to the accuracy of SWT, which is not effective when the blurry characters are connected and noisy. Pan *et al.* [27] propose an effective intensity and gradient based L0 regularized prior for text image contain black pixels. However, the intensity-based L0 regularized prior would loss effect when text image without black pixel. That is, literature [27] would degenerate to method [8] when scene text in natural image is not black. In this paper, we introduce a rough text localization method, and iteratively refine the localization results to overcome the limitation of SWT's sensitivity to blurry images. In addition, we use a series of text-specific multi-scale dictionaries to model the priors of the localized text fields. Moreover, our method is different from [12] in that non-uniform blur kernels due to depth variation are considered.

C. Non-Uniform Blur

Early works concentrate on removing spatially invariant blur, the performance of which, however, often degenerates or even fails to handle real images since the captured real blur kernels are often spatially varying because of depth variation, camera shake, etc. For example, in Fig. 1, two similar but different blur kernels and their associated clearer text fields are estimated. To model the spatial-varying blur, early relevant work including [10], [28], and [29] model images as some piecewise uniform blur regions. Their performance depends on accurate segmentation results on all regions in the image.

TABLE I

DEFINITION OF MAIN SYMBOLS

Symbol	Description
$\boldsymbol{B}$	Input blurry image
$\boldsymbol{L}$	Latent clear image
$\boldsymbol{K}$	Blur kernel
$\boldsymbol{N}$	Additive White Gaussian noise
$\mathcal{S}$	The training set for dictionary learing
$\mathcal{S}_N$	The natural image in training set $\mathcal{S}$
$\mathcal{S}_T^p$	The ground truth p^{th} scale text fields in training set $\mathcal{S}$
$\boldsymbol{D}_n$	The natural dictionary learned by training set
$\boldsymbol{D}_t^p$	The p^{th} scale Text-specific Dictionary learned by training set
$\mathcal{T}^p$	The p^{th} scale text field detected in blurry image
$\mathcal{N}$	The non-text field in blurry image
∇	The gradient operator

Our method only requires to localize text fields because our method focuses on scene text deblurring. In addition, our kernel estimation method is more appropriate to scene texts thanks to the TMD. Some different approaches [30], [31] to model non-uniform blur have been proposed recently as a linear combination of different blurry intermediate images captured by the camera along the motion trajectory. However, these methods concentrate on handling 3D camera shakes on the cost of assuming a constant scene depth. More importantly, almost all the text deblurring methods focus on estimating a single motion blur kernel for an entire scene text image. In contrast, we aim to restore text images blurred by spatially varying blur kernels in different text fields.

II. The Proposed Approach Based on TMD

Our goal is to restore blurry scene text images which contain spatially-varying blur kernels in different text fields as shown in Fig. 2. Table I lists the main symbols in this paper and their definitions. The main reason that the traditional deblurring methods perform poorly for scene text is that the natural image priors cannot well characterize the text properties. Thus, two major issues arise for natural scene text deblurring: the text detection and the modeling of text properties. To accomplish this, we first need using some

(a) (b)

Fig. 3. Example of text fields and non-text field. (a) Original scene text image from ICDAR2011 dataset. (b) Ground truth of text fields. Each text field is marked in a green rectangle bounding box.

text localization methods [14], [32], [33] to detect text areas. In this section, the text field is detected by [14], based on which we learn the TMD for the text fields, and a natural dictionary for the non-text one. As a result, our problem of interest turns out to be as:

$$\arg\min_{\boldsymbol{L},\boldsymbol{K}} \|\boldsymbol{B}-\boldsymbol{K}\otimes\boldsymbol{L}\|^2+\rho(\boldsymbol{K})+\rho_t(\boldsymbol{L})+\rho_n(\boldsymbol{L}), \quad (3)$$

where the subscript t and n stand for ***text field*** and ***non-text field***, respectively. The text fields correspond to the part of bounding boxes of detected text strings (Fig. 3(b)), while the non-text field is defined analogously.

A. Dictionary Learning for Non-Text and Text Fields

Suppose we have the well-processed training data for text and non-text fields. Consequently, we train TMD $\boldsymbol{D}_t^p$ (where p denotes the scale index) and the natural dictionary $\boldsymbol{D}_n$, respectively, with respect to the text fields and non-text fields. The ICDAR 2011 Robust Reading Competition Challenge 2 dataset is widely used in scene text localization and word recognition in natural images, which contains 229 training images and 255 testing images including ground truth of text bounding boxes. The pixel number of the scene images in the dataset ranges from tens of thousands to ten millions, which has universality in all sorts of scene text to train dictionaries. Given the training set $\mathcal{S}=\{(\sum_{j=1}^{n}\boldsymbol{N}_j,\sum_{k=1}^{m}\boldsymbol{T}_k)\}$, where $\boldsymbol{N}_j$ is a nature scene image and $\boldsymbol{T}_k$ is a bounding box of the text field. In our training set, there are $n=229$ scene images and $m=848$ text fields.

1) Non-Text Dictionary Learning: We use $\mathcal{S}_N=\{\sum_{j=1}^{n}\boldsymbol{N}_j\}$ to train the natural dictionary in a way similar to [34], [35]:

$$\boldsymbol{D}_n=\arg\min_{\boldsymbol{D}_n,\mathcal{Z}_n}\|\mathcal{S}_N-\boldsymbol{D}_n\mathcal{Z}_n\|^2+\lambda\|\mathcal{Z}_n\|_1, \quad (4)$$

where $\mathcal{Z}_n$ is the set of sparse coefficients $\{\alpha_i\}$ of natural patches, λ is the parameter controlling the weight of the sparse term. The natural dictionary used in this paper is of size 64×512, designed to handle image patches with size of 8×8 pixels. The learned non-text image dictionary is partially shown in Fig. 5(a).

2) Text-Specific Multi-Scale Dictionary Learning: As the text in scene image is very subtle, the process of text-specific dictionary learning is different with non-text dictionary learning. We train the text dictionaries on the set of the training text fields $\mathcal{S}_T=\{\sum_{k=1}^{m}\boldsymbol{T}_k\}$. However, we observe that the stroke width of scene text ranges from one pixel to more than one hundred pixels. Fig. 4(a) illustrates the statistics of stroke width density of the scene text in the 848 text bounding boxes in training set. In order to make it more convincing, we also make a statistics for the 716 text bounding boxes in the testing set in ICDAR 2011 Robust Reading Competition Challenge 2 as show in Fig. 4(b). We calculate the average stroke width of every text field based on SWT [13]. Therefore, only training one dictionary on all text using 8×8 patches is insufficient for a variety of sizes of characters in the scene texts. Statistically text scale allocated in our paper is more denser in thin stroke width text because most stroke width concentrate on the range from one to twenty pixels. We extract training text patches from different range of stroke width from the text fields dataset based on the stroke width density in Fig. 4 and Table II. The set of text scales $\mathcal{S}_T^p$ are divided into eight groups according to stroke width: [1, 4], [5, 6], [7, 10], [11, 15], [16, 30], [31, 50], and [51, ∞).

3) Learning Multi-Scale Dictionaries: Now that we have generated the set of training data, the problem of learning the TMD can be formulated as:

$$\boldsymbol{D}_t^p=\arg\min_{\boldsymbol{D}_t^p,\mathcal{Z}_t}\|\mathcal{S}_T^p-\boldsymbol{D}_t^p\mathcal{Z}_t\|^2+\lambda\|\mathcal{Z}_t\|_1, \quad (5)$$

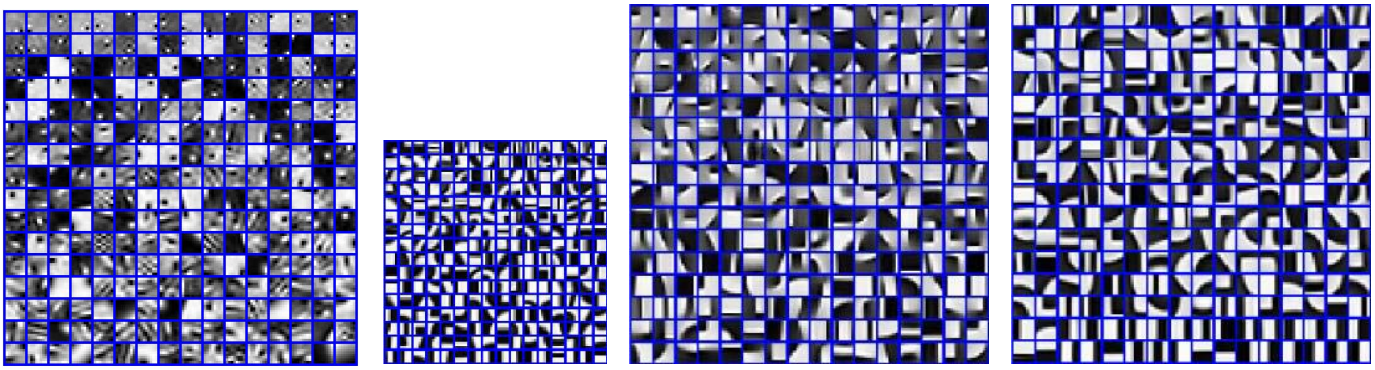
where superscript p denotes the p^{th} scale and $\mathcal{Z}_t$ is the set of sparse coefficients $\{\alpha_i\}$ of text patches. Obviously, when the stroke width of a text is as thin as a couple of pixels and there is only one character contained in the text field, a 8×8 patch size is inappropriate. Therefore, the patch sizes at different scales are assigned as $n_1=5\times5$, and $n_p=8\times8$ where $p=2,3,\ldots,7$. Correspondingly, the text-specific dictionaries at different scales used in this paper are of sizes 25×512 and 64×512, respectively.

The natural image dictionary and text-specific dictionaries have very different characteristics in their atoms. Fig. 5 (b)-(d) depict the learned text-specific dictionaries for scales of $p=1$, $p=4$ and $p=7$. As can be seen, text-specific dictionaries are able to express the various shapes and directions of strokes. Moreover, the text-specific dictionaries are more sparse than the natural dictionary. So scene text can be recover more clear using text-specific dictionaries than natural dictionary, the details can be found in Section IV-B.

B. Text-Specific Regularization Terms

Given an observed blurry scene text image or an intermediate estimated latent image, we use the structure-based partition and grouping (SPG) method in [14] to locate the text field. This method obtains state-of-the-art results of text detection using image partition to find text character candidates, and then imposing character candidate grouping to detect text strings. Next, the TMD are used in text fields according to stroke widths in each text field. We formulate the text-specific multi-scale text priors as:

$$\rho_t(\boldsymbol{L})=\sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{T}^p}\eta\|\boldsymbol{R}_i\boldsymbol{L}-\boldsymbol{D}_t^p\boldsymbol{\alpha}_i\|^2+\sum\lambda\|\boldsymbol{\alpha}_i\|_1, \quad (6)$$

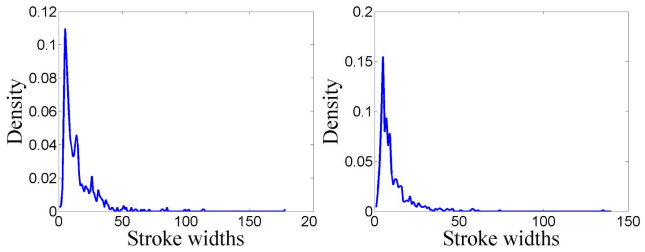


(a) (b) (c) (d)

Fig. 4. Natural dictionary and text-specific dictionaries. (a) Dictionary of natural images. (b), (c) and (d) are the 1^{st}, 4^{th} and 7^{th} scales text-specific dictionaries, respectively. The learned dictionaries demonstrate basic patterns of the image patches, such as orientated edges. Unlike the dictionary of natural image, the text-specific dictionaries are specialized to text data, such as short and curved strokes rather than stochastic nature textures in natural dictionary.

TABLE II
PROPORTION OF THE STROKE WIDTH CORRESPONDING TO FIG. 4 (a)

Stroke width	[1, 4]	[5, 6]	[7, 10]	[11, 15]	[16, 30]
proportion	0.0814	0.2028	0.2158	0.1934	0.2028
Stroke width	[31, 50]	[51, ∞)			
proportion	0.0719	0.0259			



(a) (b)

Fig. 5. The statistics of stroke width density of scene texts. (a) The density of stroke width of the training set in ICDAR 2011 Robust Reading Competition Challenge 2, and the density of stroke width in the testing set is shown in (b) to enhance the persuasiveness.

where $\boldsymbol{R}_i$ is a matrix that extracts the i^{th} patch $\boldsymbol{R}_i\boldsymbol{L}$ from the image $\boldsymbol{L}$. $\mathcal{T}^p$ denotes the set of p^{th} scale text fields. $\boldsymbol{\alpha}_i$ are the sparse coefficients for the patches. η and λ are the parameters controlling the weight of the sparse term. The non-text field priors based on the natural dictionary $\boldsymbol{D}_n$ is similarly defined as:

$$\rho_n(\boldsymbol{L}) = \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{N}} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_n\boldsymbol{\alpha}_i\|^2 + \sum \lambda\|\boldsymbol{\alpha}_i\|_1, \quad (7)$$

where $\mathcal{N}$ denotes the set of non-text fields. Note that in Eq. (6), we first detect the stroke width in every text field and then decide which scale p is used in the corresponding text field. Note also that we use the same η and λ for text and non-text fields.

C. The Uniform Deblurring Model

Apart from the dictionary-based text-specific regularization terms, we also use the sparse hyper-Laplacian prior of the entire gradient image, which is used in many deblurring methods [21], [36]. Finally, our scene text image deblurring model can be formulated as:

$$\begin{aligned}\{\hat{\boldsymbol{L}}, \hat{\boldsymbol{K}}, \hat{\boldsymbol{\alpha}}\} &= \arg\min_{\boldsymbol{L},\boldsymbol{K},\boldsymbol{\alpha}} \|\boldsymbol{B} - \boldsymbol{K}\otimes\boldsymbol{L}\|^2 + \varphi\|\boldsymbol{K}\|^2 \\ &+ \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{T}^p} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_t^p\boldsymbol{\alpha}_i\|^2 + \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{N}} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_n\boldsymbol{\alpha}_i\|^2 \\ &+ \sum \lambda\|\boldsymbol{\alpha}_i\|_1 + \beta\|\nabla\boldsymbol{L}\|^\gamma, \end{aligned} \quad (8)$$

where ∇ denotes the gradient operator. The gradient distributions are assumed to be hyper-Laplacian ($0.5 < \gamma < 0.8$) in nature scene images as in [21] and [36]–[38] to further stabilize the solution. The regularization term of the blur kernel uses l_2-norm, which reduces the noise and enables fast kernel estimation using Fast Fourier Transform (FFT) with the quadratic form as shown in Section III-C.

D. Non-Uniform Deblurring Model for Text Fields

In a captured scene text image, multiple text fields and spatially-varying blur kernels may exist. Therefore, it is in general impossible to correctly deblur the whole image if both the camera motion and the scene geometry are neither unknown. For example, the text fields enclosed by the red and green boxes in Fig. 1 have arguably different blur kernels.

In our implementation, we first use the model in Section II-C to iteratively estimate the uniform blur kernel for the entire image. Once the uniform kernel is estimated and iteratively refined, it is treated as the initial kernel to deblur every text field. The estimated uniform blur is further refined in the following non-uniform restoration stage, where we perform the refinement and image restoration for text fields.

We also introduce a segmentation mask matrix representation $\boldsymbol{M}_i$ to indicate the i^{th} text field. $\boldsymbol{M}_i$ is decided by the results of text localization method performed on the refined latent clear image output in the uniform deblurring step. Then, the blurry text fields $\boldsymbol{B}_t^i = \boldsymbol{M}_i\boldsymbol{B}$. we focus on text field

deblurring rather than background natural scene in this stage. We estimate dedicated blur kernels $\boldsymbol{K}_i$ for each $\boldsymbol{B}_t^i$ to obtain the clear latent text fields $\boldsymbol{L}_t^i$. In this stage, only text-specific dictionaries are used for text field reconstruction. Since the proposed model employs multiple blur kernels, it could handle real scene text images containing various blur kernels due to depth variations.

III. Optimization

The proposed model in Eq. (8) can be optimized efficiently with the alternating minimization scheme, which is widely adopted for dealing with multiple optimization variables. We initialize $\boldsymbol{K}$ as a delta function and $\boldsymbol{L}$ as the input blurry scene image.

A. Sparse Coefficient Computing

With the fixed latent image $\boldsymbol{L}$ and blur kernel $\boldsymbol{K}$, we use the SPG method [14] to locate the text and non-text fields. The optimization problem can be simplified to:

$$\{\hat{\boldsymbol{\alpha}}_i\} = \arg\min_{\boldsymbol{\alpha}} \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{T}^p} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_t^p\boldsymbol{\alpha}_i\|^2 + \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{N}} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_n\boldsymbol{\alpha}_i\|^2 + \sum \lambda\|\boldsymbol{\alpha}_i\|_1. \quad (9)$$

This subproblem actually is decomposable over each patch $\boldsymbol{R}_i\boldsymbol{L}$ from the latent image. Therefore, Eq. (9) is equal to solving the problem of sparse representation for each image patch:

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}_i} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_t^p\boldsymbol{\alpha}_i\|^2 + \sum \lambda\|\boldsymbol{\alpha}_i\|_1, \quad (10)$$

or

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}_i} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_n\boldsymbol{\alpha}_i\|^2 + \sum \lambda\|\boldsymbol{\alpha}_i\|_1. \quad (11)$$

Text dictionary $\boldsymbol{D}_t^p$ is used if the patch belongs to text fields detected by SPG method, and otherwise $\boldsymbol{D}_n$ is used. In addition, after finishing the text localization process during each iteration, we need to detect the average stroke width in each text field based on SWT [13] to select the appropriate scale p.

B. Latent Image Updating

We fix the estimated blur kernel $\hat{\boldsymbol{K}}$ and sparse coefficients $\{\boldsymbol{\alpha}_i\}$ of each patch. The modeling (8) reduces to:

$$\hat{\mathbf{L}} = \arg\min_{\boldsymbol{L}} \|\boldsymbol{B} - \hat{\boldsymbol{K}}\otimes\boldsymbol{L}\|^2 + \beta\|\nabla\boldsymbol{L}\|^\gamma + \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{T}^p} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_t^p\boldsymbol{\alpha}_i\|^2 + \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{N}} \eta\|\boldsymbol{R}_i\boldsymbol{L} - \boldsymbol{D}_n\boldsymbol{\alpha}_i\|^2. \quad (12)$$

We here introduce an auxiliary variable $\boldsymbol{U}$ to stand for the reconstructed image by the updated sparse coefficients. The introduced variable $\boldsymbol{U}$ allows us to eliminate the non-circular matrix $\boldsymbol{R}_i$ in FFT while estimating the latent image. Equation (12) is therefore reformulated as:

$$\{\hat{\boldsymbol{L}},\hat{\boldsymbol{U}}\} = \arg\min_{\boldsymbol{L},\boldsymbol{U}} \|\boldsymbol{B} - \hat{\boldsymbol{K}}\otimes\boldsymbol{L}\|^2 + \beta\|\nabla\boldsymbol{L}\|^\gamma + \sigma\|\boldsymbol{L} - \hat{\boldsymbol{U}}\|^2 + \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{T}^p} \eta\|\boldsymbol{R}_i\hat{\boldsymbol{U}} - \boldsymbol{D}_t^p\hat{\boldsymbol{\alpha}}_i\|^2 + \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{N}} \eta\|\boldsymbol{R}_i\hat{\boldsymbol{U}} - \boldsymbol{D}_n\hat{\alpha}_i\|^2. \quad (13)$$

Thus, this optimization problem can be decomposed into two subproblems of solving the auxiliary variables $\boldsymbol{U}$ and the latent image $\boldsymbol{L}$:

$$\hat{\boldsymbol{U}} = \arg\min_{\boldsymbol{U}} \sigma\|\hat{\boldsymbol{L}} - \boldsymbol{U}\|^2 + \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{T}^p} \eta\|\boldsymbol{R}_i\boldsymbol{U} - \boldsymbol{D}_t^p\hat{\boldsymbol{\alpha}}_i\|^2 + \sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{N}} \eta\|\boldsymbol{R}_i\boldsymbol{U} - \boldsymbol{D}_n\hat{\boldsymbol{\alpha}}_i\|^2. \quad (14)$$

This optimization problem on the auxiliary variable $\boldsymbol{U}$ has a closed-form solution of the form:

$$\hat{\mathbf{U}} = (\sigma + \eta\sum \boldsymbol{R}_i^T\boldsymbol{R}_i)^{-1} \times (\sigma\hat{\boldsymbol{L}} + \eta\sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{T}^p} \boldsymbol{R}_i^T\boldsymbol{D}_t^p\hat{\boldsymbol{\alpha}}_i + \eta\sum_{\boldsymbol{R}_i\boldsymbol{L}\in\mathcal{N}} \boldsymbol{R}_i^T\boldsymbol{D}_n\hat{\boldsymbol{\alpha}}_i). \quad (15)$$

Finally, we can solve the latent image $\boldsymbol{L}$ by optimizing:

$$\hat{\boldsymbol{L}} = \arg\min_{\boldsymbol{L}} \|\boldsymbol{B} - \hat{\boldsymbol{K}}\otimes\boldsymbol{L}\|^2 + \sigma\|\boldsymbol{L} - \hat{\boldsymbol{U}}\|^2 + \beta\|\nabla\boldsymbol{L}\|^\gamma \quad (16)$$

Using the half-quadratic penalty method as in [36], we introduce auxiliary variable w and give a new function:

$$\hat{\boldsymbol{L}} = \arg\min_{\boldsymbol{L}} \|\boldsymbol{B} - \hat{\boldsymbol{K}}\otimes\boldsymbol{L}\|^2 + \sigma\|\boldsymbol{L} - \hat{\boldsymbol{U}}\|^2 + \tau\|w - \nabla\boldsymbol{L}\|^2 + \beta\|\nabla\boldsymbol{L}\|^\gamma \quad (17)$$

This latent image optimization subproblem can be solved efficiently by the method in [36].

C. Blur Kernel Estimation

In this subproblem, we fix latent image $\hat{\boldsymbol{L}}$, and optimize the blur kernel $\hat{\boldsymbol{K}}$. Equation (8) reduces to the following form:

$$\hat{\boldsymbol{K}} = \arg\min_{\boldsymbol{K}} \|\boldsymbol{K}\otimes\hat{\boldsymbol{L}} - \boldsymbol{B}\|^2 + \varphi\|\boldsymbol{K}\|^2. \quad (18)$$

As [39], [40] point out, the kernel computed by (18) is not very accurate. We estimate the kernel by

$$\hat{\boldsymbol{K}} = \arg\min_{\boldsymbol{K}} \|\boldsymbol{K}\otimes\nabla\hat{\boldsymbol{L}} - \nabla\boldsymbol{B}\|^2 + \varphi\|\boldsymbol{K}\|^2. \quad (19)$$

The elements k_i in kernel $\boldsymbol{K}$ subject to the constraints that $k_i \geq 0$ and $\sum_i k_i = 1$. This is a least square problem with Tikhonov regularization, which leads to a closed-form solution for $\boldsymbol{K}$:

$$\hat{\boldsymbol{K}} = \mathcal{F}^{-1}\left(\frac{\overline{\mathcal{F}(\nabla\hat{\boldsymbol{L}})}\circ\mathcal{F}(\nabla\boldsymbol{B})}{\overline{\mathcal{F}(\nabla\hat{\boldsymbol{L}})}\circ\mathcal{F}(\nabla\hat{\boldsymbol{L}}) + \varphi}\right), \quad (20)$$

where $\mathcal{F}$, $\mathcal{F}^{-1}$ and $\overline{\mathcal{F}}$ denote Fast Fourier Transform (FFT), inverse FFT and the complex conjugate of FFT, respectively.

(a) (b) (c)

Fig. 6. Example of text localization and deblurring. (a) Input blurry scene text image. (b) Intermediate latent images and the localized text fields highlighted in green boxes. (c) Final deblurring result after 10 iterations. A zoom in version of the text fields is shown on the bottom.

Algorithm 1 Scene Text Image Blind Deblurring

1. **Input**: blurry text image $\boldsymbol{B}$, text-specific multi-scale Dictionaries $\boldsymbol{D}_t^p$ and natural dictionary $\boldsymbol{D}_n$.
2. **Initialization:**
   Blur kernel K as a delta function.
   Latent image $\hat{\boldsymbol{L}} = \boldsymbol{B}$.
3. **Repeat:**
4. Run text localization algorithm [14] on $\hat{\boldsymbol{L}}$.
5. **Update** sparse vector $\hat{\boldsymbol{\alpha}}_i$ by minimizing Eq. (9) based on $\{\boldsymbol{D}_t\}_p$ and $\boldsymbol{D}_n$.
6. **Update** latent text image $\hat{\boldsymbol{L}}$ via minimizing Eq. (13).
7. **Update** blur kernel $\hat{\boldsymbol{K}}$ by minimizing Eq. (19).
8. **Until** uniform blur kernel $\boldsymbol{K}$ and latent $\boldsymbol{L}$ converge.
9. Run text localization algorithm [14] on $\boldsymbol{L}$.
10. Initialize the kernels $\boldsymbol{K}_i$ in each text field as $\boldsymbol{K}$;
11. Repeat step 5-7 on each text field independently until each kernel $\boldsymbol{K}_i$ and text fields $\boldsymbol{L}_t^i$ converge.

"$\circ$" denotes element-wise multiplication. Equation (20) can be solved efficiently since it only consists of component-wise operators except FFTs. In our experiments, cross-validation is used for parameter determination and we set $\eta = 0.11$, $\sigma = 1.3$, $\lambda = 0.01$, $\beta = 0.005$, $\varphi = 5$, respectively.

After the uniform deblurring stage converges, the text fields localized by SPG method become very accurate (More detailed information can be found in section IV-A). To derive more clearer strings, therefore, we perform the non-uniform deblurring for each text field. The kernels in each text field is initialized as the final uniform kernel computed by Eq. (19). The major algorithmic steps are summarized in Algorithm 1. In the Algorithm 1, we only show the deblurring process and ignore the dictionary learning process for brevity.

IV. Analyses of the Proposed Method

A. An Illustrative Example

We first compare the precision p, recall r and a standard f-measure [14] of text localization accuracy on the clear and synthetic blurry ICDAR2011 dataset. We use the blur kernels in [4] to synthesize blurry ICDAR2011 dataset.

TABLE III
Compare Between Clear and Synthetic Blurry ICDAR 2011 Dataset

	Precision p	Recall r	f-measure
Clear dataset	0.71	0.62	0.62
Blurry dataset	0.57	0.49	0.51

The final quantitative results of p, r and f-measure are listed in Table III. Although the text localization accuracy reduced on synthetic blurry dataset while compared with clear dataset, the accuracy of text area detection goes up as the number of iterations increases in our deblurring process as shown in Fig. 6. Besides, we could use more effective and sophisticated localization methods, such as [32] and [33], to obtain more accurate results.

We illustrate the proposed algorithm with a simple example in Fig. 6. The blurry input is synthetic using the ground truth image in Fig. 3(a), and the kernel from [4]. Given a blurry input image in Fig. 6(a), we show how our method can iteratively deblur it in Fig. 6(b). Without surprise, the text field localized by SPG method become increasingly accurate. Although the final localization results is not exactly the same with the ground truth in Fig. 3, all text fields are detected at last. At the same time, the restored image resembles more and more to the ground truth image in both the text field and non-text field. When images include the non-horizontal text, the performance of our method depends on two issues: (i) the accuracy of text localization and (ii) the adaptability of TMD for non-horizontal text. The SPG method is able to detect the text strings with arbitrary orientations as illustrated in [14]. Our TMD are learned on a large amount of patches in the ICDAR 2011 training set in which some non-horizontal texts are contained although many texts are near-horizontal. That is to say, the trained dictionary is overcomplete to handle non-horizontal text. These properties make our method could also handle non-horizontal texts in blurry images, e.g., the fourth and the sixth images in Fig. 10.

It is worth noting that the text localization method [14] sometimes return false text fields. As shown in Fig. 7, false text areas generally have no harm to the final deblurring results, partially because that the detected false text fields actually seem like strokes and usually possess some text property, such as the table legs in Fig. 7(a), the arc trademark in Fig. 7(b),

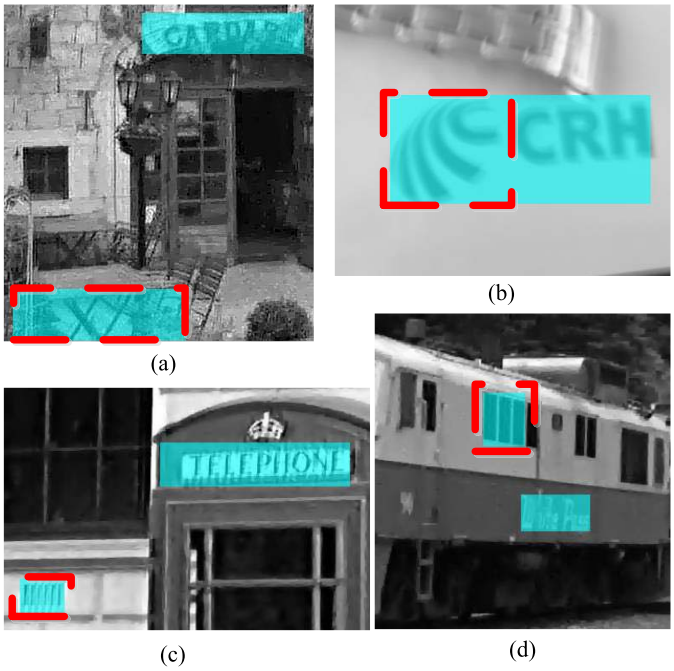


Fig. 7. Some results containing false text fields which are marked in red dashed rectangles. Such as the table legs in (a), the arc trademark in (b), the scratches on the wall in (c) and the black window in (d).

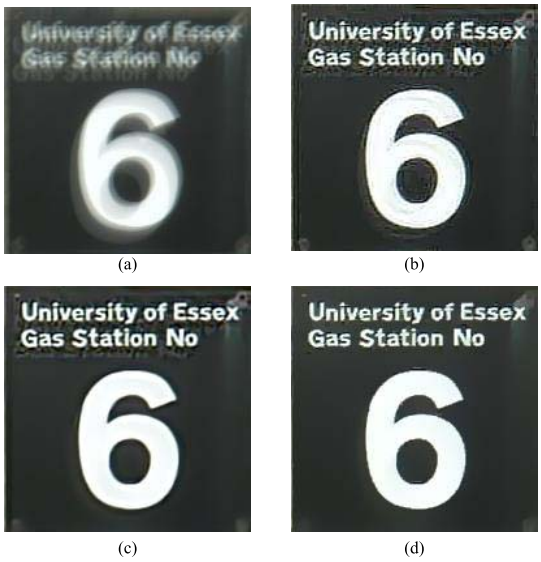


Fig. 8. Role of the TMD. (a) Input blurry image. (b) The deblurring results without text dictionaries. (c) The deblurring result with single-scale text dictionary. (d) Our result.

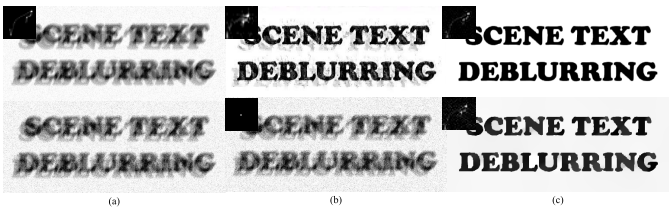


Fig. 9. Comparison with Cho *et al* [12] on blurry text images with noises. The input blurry and noisy examples are examples are shown in (a). These two images have the same blur kernel of different noisy level, *i.e.* σ = %5 (upper) and σ = 10% (blow). (b) are the estimated latent images by [12] and our results are shown in (c).

the tiny scratches on the wall in Fig. 7(c) and the black window in Fig. 7(d). Thanks to this, false text areas will not affect the process of estimating blur kernel because the detected non-text areas usually have some text features and can be well modeled by TMD. The final deblurred results corresponding to Fig. 7 can be found in Figs. 10 and 12.

Our proposed scene text deblurring algorithm is mainly based on the alternating minimization method and has the fast convergence property. In practice we found the process converges usually within 5-10 iterations. Over the iterations, the image details are enhanced. The intermediate (2^{nd}, 4^{th}, 6^{th}, 8^{th}) text localization and deblurring results are shown in Fig. 6(b).

B. Role of TMD

In the task of scene text deblurring, text field deblurring effect is more important than non-text field because text characters and strings contain valuable information and are exploited in many content-based image and video applications. As the text fields are detected more and more stable and accurate along with the iterations, the text field can be represented by TMD effectively because the text dictionaries prefer neat and sharp strokes.

In addition, texts of different sizes may exist in a scene image, so we need to learn different text dictionaries for various sizes of texts. If only one text dictionary is learned for all the sizes of texts, it is highly unlikely to be suitable for different text sizes because texts are very subtle in a scene image and very sensitive to different scales of dictionaries in the reconstruction stage. To verify this, an experiment is conducted and shown in Fig. 8. Fig. 8(a) shows the input blurry scene text images with two sizes of text. The average stroke width of the number "6" is 21 pixels, while others are around 3 pixels width respectively in Fig. 8. Fig. 8(b) displays the deburring result with only the natural dictionary in Fig. 5(a). There are unpleasing artifacts around all the strings in Fig. 8(b). Figure 8(c) shows the result with the natural dictionary and a single scale text dictionary. Although the number "6" is clear, the ring effects in the top strings still exist.

C. Robustness Against Noises

Our approach behaves well with a level of Gaussian noise up to σ = 10% because the reconstruction stage is based sparse representation which has been demonstrated successes in denoising [19], [42], [43]. The state-of-the-arts method Cho *et al.* [12] is the most related work to ours since it also seeks to handle natural blurry text images. One of the limitations in Cho *et al.* [12] is that it cannot handle text images with noises. We ran comparisons on the synthetic blurry and noisy image. One example is shown in Fig. 9, where the latent sharp image with added Gaussian noise.

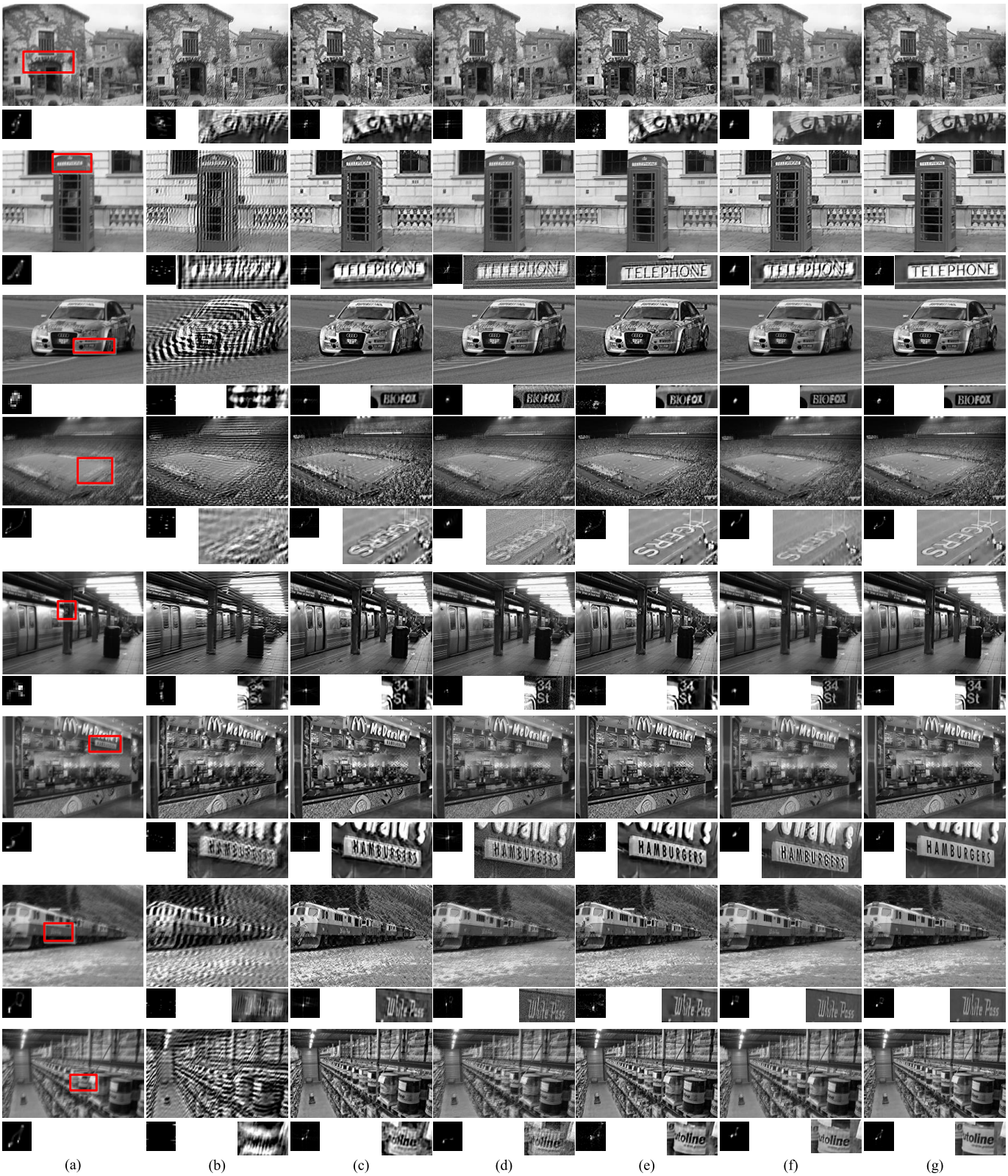


Fig. 10. Comparison with different methods on synthetic images. (a) Input synthetic blurry image. (b) Cho *et al.* [41]. (c) Krishnan *et al.* [15], (d) Cai *et al.* [17], (e) Cho *et al.* [12], (f) Zhong *et al.* [7] and (g) Our uniform deblurring method.

The comparison shows that the method in [12] is sensitive to nontrivial noises. Theoretically, the contributing factor in reducing noises in [12] is using l_0 gradient minimization [44]. Because l_0 gradient minimization cannot handle a severe noise in images, so this damages the deblurring results in [12]. However, our method can improve the result to a certain degree thanks to the property of sparse representation as demonstrated in Fig. 9(c). This property is beneficial to

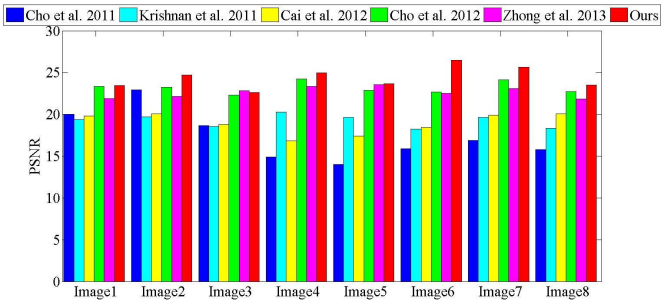


Fig. 11. Summary of the deblurring average PSNR results on 64 (8 images×8 kernels) test images.

handle real blurry images because there are often noises in real images.

V. Experiments

In this section, we conduct the experiments on both synthesized and real data to demonstrate the efficacy of the proposed method, and compare it to the state-of-the-arts including Cho *et al.* [41], Cai *et al.* [17], Krishnan *et al.* [15], Cho *et al.* [12], and Zhong *et al.* [7]. To quantitatively evaluate the performance of the involved methods, we employ the peak-signal-to-noise ratio (PSNR) as the metric.

A. Results on Synthetic Data

We first quantitatively evaluate the scene text image deblurring accuracy on the test set in [45]. The synthetic test set includes 640 high resolution natural images of diverse scenes, which are generated from 80 images and 8 different kernels provided by [4]. We use 64 images in this dataset because this test set is mainly about natural landscape and only few images contain scene texts. In addition, all images in this dataset are added 1% additive Gaussian noises to the blurry images to model sensor noises for better tests on deblurring effects.

We compare our uniform algorithm with the methods [7], [12], [15], [17], [41]. Figure 10 shows the deblurring results on the 64 synthetic data. We show all the 8 images, while 8 different kernels as shown in the left bottom of the input blurry image. Our method outperforms all the other methods, with [12] (in Fig. 10(e)) being the most competitive. The results indicate that the method proposed by Cho *et al.* [41] cannot obtain good results on nearly all images. Although the methods [7], [15], [17] can obtain clear results, their ring effect is serious especially around texts in the result in Fig. 10(c-d). The approach of Cho *et al.* [12] shows reasonable results in both the text and non-text fields, but still could be improved in the text field. Our algorithm generates neater and clearer results compared to competing methods.

The qualitative results agree with the quantitative PSNR metrics. Quantitatively the PSNR results, averaged over all 8 kernels, of each latent clear image output by each method are shown in Fig. 11. On average, our method has the highest PSNR (24.32 dB), while the lowest PSNR is (17.37 dB) deblurred by [41]. The second performance is Cho *et al.* [41], which achieves the average PSNR (23.19 dB), lower than ours. We excel in [7], [12], [15], [17], and [41] out of the 64 input blurry images.

B. Real Data

Although many methods are proposed for image deblurring, different text fields in different locations in a scene image are often contaminated by spatially-varying blur kernels. Therefore, we run our uniform and non-uniform methods in some real images to gain a better insight into the behavior of the proposed deblurring method. We compare our method against the four single image blind deblurring methods, *i.e.*, Krishnan *et al.* [15], Cai *et al.* [17], Cho *et al.* [12] and Zhong *et al.* [7]. For [7], [15], and [17], the results are from the authors implementations and we used the hand-tuned parameters to produce the best possible results. For [12], the results are from our own implementations.

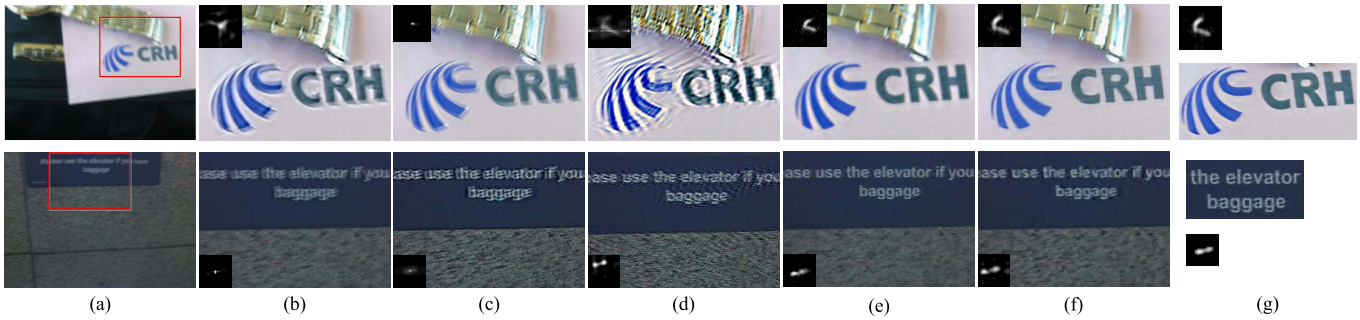
Figs. 12 and 13 show the deblurred results from all methods on four real blurry images from a variety of camera motions. The texts in input blurry images in Figs. 12 are almost at the similar depth. However, the texts in Figs. 13 have depth-variations. The blur kernels estimated by our uniform method are show in Figs. 12-13(f), which is further refined on different text fields by our spatially varying kernel estimation stage, as shown in Figs. 12-13(g). The deblurring results by our uniform and non-uniform methods are much the same in Figs. 12(f-g), because of the trivial depth-variant, that is to say, the blur kernels in these two images are nearly spatially invariant. However, our non-uniform method consistently improve the uniform kernel in Figs. 13. The main reason is that the different text fields have various depth.

At last, Fig. 14 shows the results obtained by our method, Cho *et al.* [12] and other two state-of-the-art L0 norm based methods [8], [27] running on the blurry image from [12]. It can be seen that our proposed algorithm also leads to better performance than any other algorithms in Fig. 14. Figure 14(b) is the highlighted area in blurry input Fig. 14(a). The deblurring result using [12] is shown in Fig. 14 (c). The result using the effective gradient based L_0 sparse regularization [8] is shown in Fig. 14(d) and result using gradient and intensity based L_0 regularization is shown in Fig. 14(e). As it is shown in Fig. 14(d-e), the results deblurred by [8] and [27] are very similar and both better than [12], but less clear and sharp than our result.

We use the free OCR engine (http://www.i2ocr.com/) to recognize the deblurred real images to verify the proposed method. As shown in Table IV, the overall recognition precision on the real images is 0% before deblurring and the highest average precision is 56.8% after deblurring by our non-uniform method.

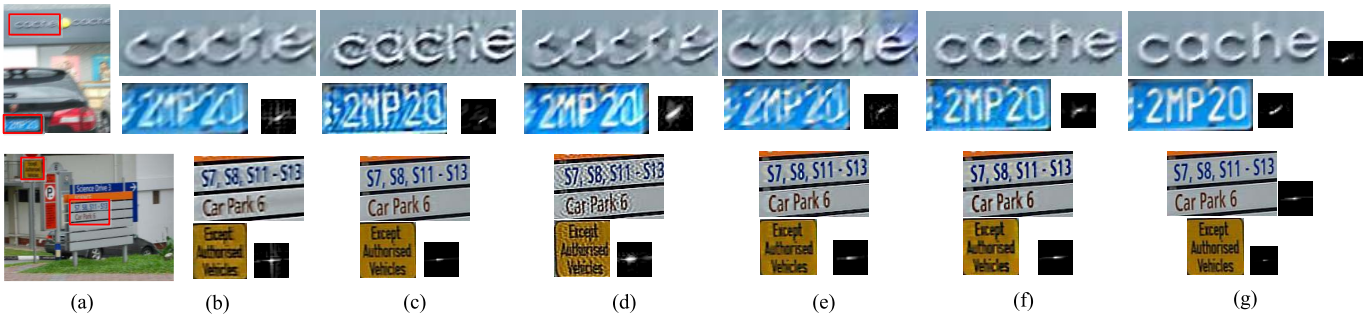
C. Failure Cases

Our deblurring method depends on text localization method [14]. This method may miss some text field



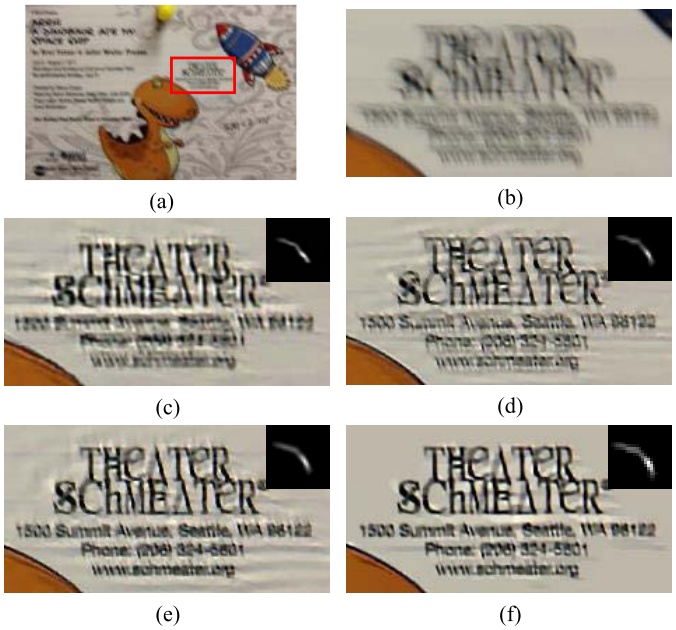
(a) (b) (c) (d) (e) (f) (g)

Fig. 12. The comparison with different methods on real text images without significant depth variations. (a) Input real blurry image and zoomed region. (b) Krishnan *et al.* [15]. (c) Cai *et al.* [17]. (d) Zhong *et al.* [7]. (e) Cho *et al.* [12]. (f) Our uniform result. (g) Our non-uniform result.



(a) (b) (c) (d) (e) (f) (g)

Fig. 13. The comparison with different methods on real text image with varying depth. (a) Input real blurry images. The second image come from [2]. (b) Krishnan *et al.* [15]. (c) Cai *et al.* [17]. (d) Zhong *et al.* [7]. (e) Cho *et al.* [12]. (f) Our uniform result. (g) Our non-uniform result.



(a) (b) (c) (d) (e) (f)

Fig. 14. Compare with the latest state-of-the-art methods on a real blurry scene image (a) from [12]. (b) is the highlighted area in (a). (c) Cho *et al.* [12], (d) Xu *et al.* [8], (e) Pan *et al.* [27] and (f) our result.

in certain cases, such as out-of-focus or poor resolution scene images. Consequently, our text-specific dictionaries lose the power and cannot help in recover clear texts. In other word, our method degenerates to only using natural dictionary to deblur images. One example is shown in Fig. 15.

TABLE IV

OCR RESULTS FOR REAL IMAGES DEBLURRED BY DIFFERENT METHODS

Methods	Input	[15]	[17]	[7]	[12]	Our uni	Our non-uni
precision	0%	2.3%	36.4%	2.3%	40.2%	41.2%	56.8%

(a) (b)

Fig. 15. An example of failure case. Our algorithm fails to recover clear character when the text localization fails. (a) Input image. (b) Our result.

VI. CONCLUSION

We have proposed a novel method for the task of scene text image deblurring by exploiting the characteristics of the text fields. Our method employs the text localization method to locate the text fields, and then take advantages of the TMD and the natural dictionary, trained on the text and non-text patches, to recover the latent image. The dictionary-based method is

more general and flexible in modeling the scene text properties than the ad hoc methods in modeling image priors, and is more robust to noises. In addition, because the blur appearing in real world scenarios is hardly a perfect spatial-invariant motion blurring, our non-uniform method on each text field can better recover clear scene texts. Through these steps, we have conducted extensive experiments on both the synthetic and real data, which demonstrate the certain improvement over the state-of-the-arts. We note that the kernels of text fields in our non-uniform stage are initialized based on uniform deblurring, which makes our model have modest effect when the kernels in different text fields are significantly varying. In future work, we will design more elaborated piece-wise deblurring model to take local evidence and global consistence for robust and accurate estimation of kernels for different text fields.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the three reviewers for their constructive suggestions.

REFERENCES

[1] C. Yi, X. Yang, and Y. Tian, "Feature representations for scene text character recognition: A comparative study," in *Proc. 12th ICDAR*, Aug. 2013, pp. 907–911.

[2] J.-F. Cai, H. Ji, C. Liu, and Z. Shen, "Blind motion deblurring from a single image using sparse approximation," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 104–111.

[3] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 787–794, 2006.

[4] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1964–1971.

[5] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. ID 73.

[6] L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *Proc. 11th ECCV*, 2010, pp. 157–170.

[7] L. Zhong, S. Cho, D. Metaxas, S. Paris, and J. Wang, "Handling noise in single image deblurring using directional filters," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 612–619.

[8] L. Xu, S. Zheng, and J. Jia, "Unnatural L_0 sparse representation for natural image deblurring," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 1107–1114.

[9] H. Madero-Orozco, P. Ruiz, J. Mateos, R. Molina, and A. K. Katsaggelos, "Image deblurring combining poisson singular integral and total variation prior models," in *Proc. 21st EUSIPCO*, Sep. 2013, pp. 1–5.

[10] H. Ji and K. Wang, "A two-stage approach to blind spatially-varying motion deblurring," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 73–80.

[11] X. Chen, X. He, J. Yang, and Q. Wu, "An effective document image deblurring algorithm," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 369–376.

[12] H. Cho, J. Wang, and S. Lee, "Text image deblurring using text-specific properties," in *Proc. 12th ECCV*, 2012, pp. 524–537.

[13] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 2963–2970.

[14] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2594–2605, Sep. 2011.

[15] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 233–240.

[16] J. Jia, "Single image motion deblurring using transparency," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[17] J.-F. Cai, H. Ji, C. Liu, and Z. Shen, "Framelet-based blind motion deblurring from a single image," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 562–572, Feb. 2012.

[18] L. Yuan, J. Sun, L. Quan, and H.-Y. Shum, "Image deblurring with blurred/noisy image pairs," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. ID 1.

[19] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[20] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proc. IEEE*, vol. 98, no. 6, pp. 972–982, Jun. 2010.

[21] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang, "Close the loop: Joint blind image restoration and recognition with sparse representation prior," in *Proc. IEEE ICCV*, Nov. 2011, pp. 770–777.

[22] H. Li, Y. Zhang, H. Zhang, Y. Zhu, and J. Sun, "Blind image deblurring based on sparse prior of dictionary pair," in *Proc. 21st ICPR*, Nov. 2012, pp. 3054–3057.

[23] X. Y. Qi, L. Zhang, and C. L. Tan, "Motion deblurring for optical character recognition," in *Proc. 8th ICDAR*, Aug./Sep. 2005, pp. 389–393.

[24] B. Su, S. Lu, and T. C. Lim, "Restoration of motion blurred document images," in *Proc. 27th Annu. ACM Symp. Appl. Comput.*, 2012, pp. 767–770.

[25] T.-H. Li and K.-S. Lii, "A joint estimation approach for two-tone image deblurring by blind deconvolution," *IEEE Trans. Image Process.*, vol. 11, no. 8, pp. 847–858, Aug. 2002.

[26] T. Kasar, J. Kumar, and A. G. Ramakrishnan, "Font and background color independent text binarization," in *Proc. ICDAR*, 2007, pp. 3–9.

[27] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring text images via L_0-regularized intensity and gradient prior," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 2901–2908.

[28] S. Cho, Y. Matsushita, and S. Lee, "Removing non-uniform motion blur from images," in *Proc. IEEE 11th CVPR*, Oct. 2007, pp. 1–8.

[29] A. Levin, "Blind motion deblurring using image statistics," in *Advances in Neural Information Processing Systems 19*. Cambridge, MA, USA: MIT Press, 2006, pp. 841–848.

[30] A. Gupta, N. Joshi, C. Lawrence Zitnick, M. Cohen, and B. Curless, "Single image deblurring using motion density functions," in *Proc. 11th ECCV*, 2010, pp. 171–184.

[31] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-uniform deblurring for shaken images," *Int. J. Comput. Vis.*, vol. 98, no. 2, pp. 168–186, 2012.

[32] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8027–8048, 2014.

[33] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 1083–1090.

[34] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[35] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, Oct. 2012.

[36] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Advances in Neural Information Processing Systems 22*. Red Hook, NY, USA: Curran Associates, 2009, pp. 1033–1041.

[37] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. ID 70.

[38] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, "A generalized iterated shrinkage algorithm for non-convex sparse coding," in *Proc. IEEE ICCV*, Dec. 2013, pp. 217–224.

[39] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Efficient marginal likelihood optimization in blind deconvolution," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 2657–2664.

[40] S. Cho and S. Lee, "Fast motion deblurring," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. ID 145.

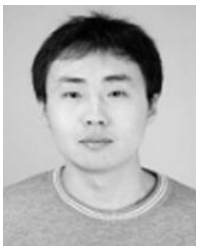
[41] T. S. Cho, S. Paris, B. K. P. Horn, and W. T. Freeman, "Blur kernel estimation using the radon transform," in *Proc. IEEE CVPR*, Jun. 2011, pp. 241–248.

[42] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 27–35, Jan. 2009.

[43] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[44] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via L_0 gradient minimization," *ACM Trans. Graph.*, vol. 30, no. 6, 2011, Art. ID 174.
[45] L. Sun, S. Cho, J. Wang, and J. Hays, "Edge-based blur kernel estimation using patch priors," in *Proc. IEEE ICCP*, Apr. 2013, pp. 1–8.

Xiaochun Cao (SM'14) received the B.E. and M.E. degrees in computer science from Beihang University, China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university level Outstanding Dissertation Award. After graduation, he spent about three years with ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He has authored and co-authored over 80 journal and conference papers. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.



Wenqi Ren received the B.E. degree from the School of Information Engineering, Hebei University of Technology, in 2010, and the M.E. degree from the School of Computer Science and Communication Engineering, Tianjin University of Technology, in 2013. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Tianjin University. His current research interests are image processing and computer vision.

Wangmeng Zuo (M'09) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. In 2004, from 2005 to 2006, and from 2007 to 2008, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. From 2009 to 2010, he was a Visiting Professor with Microsoft Research Asia. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He has authored about 50 papers in those areas. His current research interests include image modeling and low-level vision, discriminative learning, biometrics, and computer vision. He is also an Associate Editor of the *IET Biometrics* and the *Scientific World*.

Xiaojie Guo (M'13) received the B.E. degree in software engineering from the School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China, in 2008, and the M.S. and Ph.D. degrees in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2010 and 2013, respectively. He is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He was a recipient of the Piero Zamperoni Best Student Paper Award with the International Conference on Pattern Recognition (International Association on Pattern Recognition), in 2010.

Hassan Foroosh (M'02–SM'03) is currently a Professor of Electrical Engineering and Computer Science with the University of Central Florida. He has authored and co-authored over 100 peer-reviewed journal and conference papers. In 2004, he was a recipient of the Pierro Zamperoni award from the International Association for Pattern Recognition (IAPR). He has been in the Organizing and the Technical Committees of numerous international conferences. He has been serving as the Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING since 2011. He was also an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2003 to 2008. He received the best scientific paper award in the International Conference on Pattern Recognition of IAPR in 2008.