Journal of Information Hiding and Multimedia Signal Processing ©2016 ISSN 2073-4212
Ubiquitous International Volume **7**, Number **4**, July **2016**

Text Detection Based on Discriminative Dictionary Learning

Chong-Ming Zhao, Zhen-Feng Zhu, and Yao Zhao

Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
zhfzhu@bjtu.edu.cn

Received March, 2016; revised April, 2016

Abstract. *The rapid development of internet information and technology has witnessed the great convenience for the acquisition and transmission of natural scene images. Since the texts embedded in natural scene images can highly contribute to the content analysis and understanding of scene image, text detection in natural scene image has received much attention in recent years. In this paper, we propose a dictionary learning based text detection framework. Different from the previously proposed dictionary learning algorithms for obtaining sparse representation, an atom reduction based discriminative dictionary learning is proposed. Based on the dictionary with reduced atoms, a more powerful discriminative sparse representation classifier, or dSRC, is built for on-line text detection. In addition, some heuristic stratifies in the proposed framework are also adopted for further post-processing to fulfill text block detection and word segmentation. We evaluate the proposed discriminative dictionary learning based text detection method on the open available ICDAR 2003 dataset and the experimental results validate its effectiveness.*
Keywords: Text detection; Sparse representation; Dictionary learning.

1. **Introduction.** There generally exist some embedded texts in a natural image. It has been a consensus that the embedded text can be much beneficial to the content analysis and understanding of scene image. Due to its wide variety of potential applications, scene text detection and localization have gained increasing attention. Although recent progresses in computer vision and machine learning have substantially improved its performance, scene text detection is still an open problem.

To detect text in a natural image, we often confront with some difficulties such as the extreme diversity of text patterns and highly complicated background information. The diversity of text patterns can be variation in size, distortion, occlusion, etc. In addition, there also exist a large amount of noise and text-like outliers in a highly complicated background, which will generally cause high false alarms. Up to now, a great deal of efforts have been put on addressing these problems, and many successful text detection methods have been proposed. Roughly, these methods can be categorized into two groups: texture-based and connected component based methods.

For texture-based methods, text is regarded as a special texture that is distinguishable from the background. By scanning each of sub-windows in multiple scales through all locations of an image, some kind of pre-trained robust classifier is applied to give discrimination on whether the sub-window is a text or not. Typically, some manually designed low-level features, such as *SIFT* and Histogram of Oriented Gradients (*HoG*) [1], are extracted from each of sub-windows. In [3], Zhong et al. apply discrete cosine transform

to extract the horizontal and vertical frequency features to perform text detection. In the work by Ye et al. [4] , the support vector machine classifier is used to classify the wavelet coefficient based texture feature for text line detection. Although some satisfactory text detection performances have been reported, the main limitation of the above methods is the high computational complexity due to the demand for scanning a large amount of windows.

As opposed to texture-based methods, the connected component based, or *CC-based*, approaches extract regions from the image and use some geometric constraints to rule out non-text candidates. For this kind of method, pixels with similar property, for instance, color,etc., are grouped into connected components, and then into text regions. Recently, Epshtein et al.[6] have proposed using the *CCs* in a stroke width transformed image, which is generated by shooting rays from edge pixels along the gradient direction. In [7], the cluster-based templates were explored by Kim et al. for altering out non-text components for multi-segment component.

As an effective signal representation method, the use of sparse representation has recently drawn much attention in diverse classification applications , such as face recognition [9][14]and super-resolution image reconstruction[10]. Pan et al.[11] proposed a new method for text detection with the use of sparse representation. It extracts text-like edges from an image by using *K-SVD* based dictionary learning[12]. However, as the *K-SVD* dictionary is designed for coding, it can be confused by complex background with text-like areas[13].Zhao et al.[13] overcome this deficiency by using a discriminated dictionary.

In this paper, a dictionary learning based text detection framework is proposed. Considering that, for an over-complete dictionary, not all of atoms play the same roles in data reconstruction, thus removing some 'non-representative' atoms would have a negligible impact on the reconstruction of a data from the same class as the training data. Based on this observation, we propose an atom reduction based discriminative dictionary learning, which leads to a more powerful discriminative sparse representation classifier,or *dSRC*. To boost the efficiency of the proposed text detection framework, the maximally stable extremal regions (*MSER*) algorithm is first applied to carry out the preprocessing on the input image. In addition, some heuristic stratifies in the proposed framework are also adopted for further post-processing to fulfill text block detection and word segmentation.

The rest of paper is organized as follow. In Section 2, some preliminaries for notation definition and dictionary learning are presented. The detailed introduction to the proposed discriminative dictionary learning is given in Section 3. Section 4 illustrates the discriminative dictionary learning based text detection framework. Some experimental results and analyses can be found in Section 5.Finally, we give the concluding remarks in Section 6.

2. **Preliminaries.**

2.1. **Notations.** First, let's give some useful mathematical notations. Throughout the paper, we use bold uppercase letter to denote matrix and bold lowercase letter to denote vector. Let $X = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^{M\times N}$ be the input data matrix , where $x_i \in \mathbb{R}^{M\times 1}$ is a data instance. Let $\|A\|_F = \left(\sum_{i=1}^{M}\sum_{j=1}^{N} A[i,j]^2\right)^{1/2}$ denote the Frobenius norm of matrix $A \in \mathbb{R}^{M\times N}$. For a vector $a \in \mathbb{R}^M$,we define its ℓ^2 norm by $\|a\|_2 = \left(\sum_{i=1}^{M} a[i]^2\right)^{1/2}$, ℓ^1 norm by $\|a\|_1 = \sum_{i=1}^{M} |a[i]|$,and ℓ^0 norm by $\|a\|_0 = \#\{j, a[j] \neq 0\}$, which counts the number of nonzero entries in the vector a and is a pseudo-norm in fact due to not satisfying the required axioms. When $\psi \subseteq \{1, 2, \cdots, N\}$ is a finite set of indices, $A_{:,\psi} \in \mathbb{R}^{M\times(N-|\psi|)}$, $A \in \mathbb{R}^{M\times N}$ stands for the sub-matrix of without containing the columns of corresponding to the indices in ψ. Similarly , for $\psi \subseteq \{1, 2, \cdots, N\}$, $A_{..\psi} \in \mathbb{R}^{M\times(N-|\psi|)}$ denotes the

sub-matrix of A without containing the rows of A corresponding to the indices in ψ. In addition, we use $a_{i\cdot}$ and $a_{\cdot j}$ to denote the i^{th} row and the j^{th} column of matrix A, respectively.

2.2. **Dictionary Learning For Sparse Representation.** A common way to represent real-valued data is with a linear combination of a collection of basis functions, which are generally referred to as atoms of a dictionary $D = [d_{\cdot 1}, d_{\cdot 2}, \cdots, d_{\cdot K}] \in \mathbb{R}^{d\times K}$ with each column $d_{\cdot i} \in \mathbb{R}^{M\times 1}$ being an atom. Supposing that the data $x \in \mathbb{R}^{d\times 1}$ admits a sparse approximation over an over-complete dictionary D with K atoms(where $K \gg d$), X can be approximately represented as a linear combination of a few atoms from D. There are a number of algorithms that can be used to learn D. Given the training dataset $X = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^{M\times N}$, one of the most popular algorithms is *K-SVD* [12], by which an over-complete dictionary is obtained by solving the following optimization problem:

$$\langle D, R\rangle = \arg \min_{D,R} \|X - D\cdot R\|_F^2 + \lambda \sum_{t=1}^{N} \|r_{\cdot t}\|_0 \tag{1}$$

where λ is a trading-off parameter to balance the reconstruction error term and sparseness penalty, and $R = [r_{\cdot 1}, r_{\cdot 2}, \cdots, r_{\cdot N}] \in \mathbb{R}^{K\times N}$ denotes the sparse coding matrix with each column r_i being the sparse representation of data instance x_i. Here, for the purpose of promoting the sparseness penalty, it is an intuitive way to adopt the ℓ^0 norm according to its definition. However, it is generally non-trivial to solve Eq.(1) due to its non-convex and non-smooth quality and has indeed been shown to be an *NP-hard* problem. An alternative relaxation way is to replace the ℓ^0 norm in Eq.(1) by using ℓ^1 norm, which yields:

$$\langle D, R\rangle = \arg \min_{D,R} \|X - D\cdot R\|_F^2 + \lambda \sum_{t=1}^{N} \|r_{\cdot t}\|_1 \tag{2}$$

In general, the optimization problem given by Eq.(2) is not jointly convex with respect to variables D and R. But if we fix one of them, either D or R, the objective function with respect to the other variable becomes a convex function. Thus, to solve Eq.(2) will mainly consist of two steps: 1) Sparse coding with the available dictionary D; 2) Dictionary updating with the obtained sparse coding matrix R. The above two steps need to be implemented iteratively until meeting some kind of stoping condition. In fact, when we fix variable D in Eq.(2), the sparse coding procedure can be transformed into a linear programming problem. For the dictionary updating step, the *K-SVD* algorithm [12]is applied due to its stability and efficiency. With the learned over-complete dictionary D, the sparse representation r of the input data instance x can be given by:

$$\langle r\rangle = \arg \min_{r\in\mathbb{R}^{K\times 1}} \left(\|x - D\cdot r\|_2^2 + \lambda\|r\|_1 \right) \tag{3}$$

In order to make the sparse representation take on powerful discrimination ability to adapt to some specific applications, the most popular way is to impose some kind of supervision constraint on the original dictionary learning model, which gives rise to various of supervised dictionary learning algorithms. Although good discrimination ability for the obtained sparse representation have been reported in these works, several limitations for them are also obvious. First, to solve the optimization model involved in these supervised dictionary learning methods becomes much more difficult due to the high complexity of the model. In addition, there are quite a lot of parameters need to be given experimentally in most cases, which inevitably weakens the flexibility of these supervised methods.

Different from the previously proposed supervised dictionary learning algorithms to obtain discriminative sparse representation, we proposed an atom reduction based discriminative dictionary learning method, which will be presented detailedly in the following.

3. **Discriminative Dictionary Learning.**

3.1. **Atom Reduction.** Once we get the sparse representation r for an out-of-sample x, the reconstruction error term of Eq.(3) can be directly used to act as a classifier, which is also known as sparse representation classifier or SRC[9]. Specifically, the output label y given by SRC classifier for the out-of-sample x will be:

$$y = \min_{j\in(1,2,...,c)} (E^j = \left\|x - D^j \cdot r^j\right\|_2^2) \tag{4}$$

where D^j, $j = 1, 2, ..., c$, denotes the j^{th} available dictionary obtained by dictionary learning via Eq.(2). Note that, for SRC classifier, the dictionary D^j , $j = 1, ..., c$, has been learned on the corresponding j^{th} class training dataset X^j.

Given the j^{th} over-complete dictionary $D^j = \left[d^j_{.1}, d^j_{.2}, \cdots, d^j_{.K_j}\right] \in \mathbb{R}^{d_j\times K_j}$ and the corresponding sparse coding matrix $R^j = \left[(r^j_{1.})^T, (r^j_{2.})^T, \cdots, (r^j_{N_j.})^T\right]^T = \begin{bmatrix} r^j_{11} & r^j_{12} & \cdots & r^j_{1N_j} \\ r^j_{21} & r^j_{22} & \cdots & r^j_{2N_j} \\ & & \vdots & \\ r^j_{K_j1} & r^j_{K_j2} & \cdots & r^j_{K_jN_j} \end{bmatrix} \in \mathbb{R}^{K_j\times N_j}$, where K_j and N_j are the sizes of the dictionary D^j and the training set from X^j the j^{th} class, respectively, we define the approximation $\tilde{X^j}$ of X^j by:

$$\tilde{X^j} = D^j \cdot R^j = \sum_{i=1}^{K_j} d^j_{.i} \cdot r^j_{i.} \tag{5}$$

From Eq.(5), it is not hard to find that different atoms $d^j_{.i}$'s play different roles in constructing the approximation $\tilde{X^j}$. That is to say, some of atoms can be seen as 'representative' atoms that play significant roles to form the the approximation $\tilde{X^j}$, and otherwise 'non-representative' ones.

To quantificationally make evaluation whether an atom $d^j_{.i}$ from D^j, $i = 1, ...N_j$, is a representative atom or not, we introduce two kinds of evaluators , information entropy and reconstruction residual, in the following.

Let's now define a row-normalized matrix S^j by :

$$S^j = \begin{bmatrix} s^j_{11} & s^j_{12} & \cdots & s^j_{1N_j} \\ s^j_{21} & s^j_{22} & \cdots & s^j_{2N_j} \\ & & \vdots & \\ s^j_{K_j1} & s^j_{K_j2} & \cdots & s^j_{K_jN_j} \end{bmatrix} \tag{6}$$

where $s^j_{ik} = \frac{|r^j_{ik}| - \min\left\{|\, r^j_{ik}|\right\}_{k=1,...,N_j}}{\max\left\{\, |r^j_{ik}|\right\}_{k=1,...,N_j} - \min\left\{\, |r^j_{ik}|\right\}_{k=1,...,N_j}}$ and $s^j_{ik} \in [0, 1]$. Furthermore, we define the information entropy $H^j(i)$ corresponding to the i^{th} atom $d^j_{.i}$ by:

$$H^j(i) = -\ln(N_j)^{-1} \sum_{k=1}^{N_j} f^j_{ik} \ln f^j_{ik} \tag{7}$$

where $f^j_{ik} = s^j_{ik} / \sum_{k=1}^{N_j} s^j_{ik}$. From the above definition of the information entropy, we can know that bigger information entropy $H^j(i)$ implies that the i^{th} atom will be commonly

shared by most of training samples whose coefficients r_{ik}^j's, $k = 1, ..., N_j$, are non zero;in other words, these non-zero coefficients are closed to even-distributed.

In addition, we define the reconstruction residual evaluator $E^j(i)$ over the training set X^j without involvement of the i^{th} atom $d_{\cdot i}^j$ by:

$$E^j(i) = \left\| X^j - D_{:\,i}^j \cdot R_{\cdot\cdot i}^j \right\|_F^2 \tag{8}$$

For the reconstruction residual evaluator $E^j(i)$, smaller $E^j(i)$ means that the i^{th} atom $d_{\cdot i}^j$ will play an insignificant role in approximating X^j.

Obviously, both the information entropy evaluator $H^j(i)$ and the reconstruction residual evaluator $E^j(i)$ as defined above could be good indicators to reflect the confidence of the i^{th} atom $d_{\cdot i}^j$ to be one of the representative atoms. Thus, it is straightforward to combine them by:

$$C^j(i) = E^j(i) + \beta \cdot H^j(i) \tag{9}$$

where $\beta > 0$ is a trade-off parameter to balance the residual evaluator $E^j(i)$ and the information entropy $H^j(i)$. Based on $C^j(i)$, we can determine which atoms are 'non-representative' atoms.

Intuitively, for a sample $x \in X^j$, instead of D^j we can only use those 'representative' atoms $D_{:\psi^j}^j$ to form the approximation to itself, which will make practically a negligible difference. Here, ψ^j is the index set of 'non-representative' atoms from the dictionary D^j. But, a contrary result will be caused when using $D_{:\psi^k}^k$ to form the approximation, where $k = 1, ..., c$ and $k \neq j$. The above observation directly motivates us in this paper to propose the discriminative dictionary learning based on atom reduction.

3.2. **Algorithm Summarization for Discriminative Dictionary Learning.** The detail implementation for learning a discriminative dictionary based on atom reduction is described in **Algorithm** 1. In on-line text detection stage, for each of sub-window x (sample), we first compute its sparse coding r^j based on the dictionary D^j by Eq.(3), where $j = 1, ..., c$. Then, we eliminate the elements of r^j with the index ψ^j via **Algorithm** 1 to obtain a discriminative sparse representation $r_{\cdot\cdot\psi^j}^j$.

With the set of remained representative atoms $D_{:\psi^j}^j$ from D^j and the corresponding sparse representation $r_{\cdot\cdot\psi^j}^j$, $j = 1, ..., c$, the label output for sample x based on *SRC* classifier as given in Eq.(4) will become:

$$y = \min_{j \in (1,2,...,c)} \left(\left\| x - D_{:\psi^j}^j \cdot r_{\cdot\cdot\psi^j}^j \right\|_2^2 \right) \tag{10}$$

In order to make difference from the traditional *SRC* classifier, we name the proposed classifier for sparse representation given by Eq.(10) by discriminative sparse representation classifier or *dSRC* in abbreviation.

4. Discriminative Dictionary Learning Based Text Detection.

4.1. **Framework.** The proposed text detection framework based on discriminative dictionary learning is illustrated in Figure 1. In off-line learning stage, two atom reduction based discriminative dictionaries, which correspond to foreground (text) and background, respectively, are first trained to form a discriminative sparse representation classifier. Whereas for on-line text detection stage, as we can see, it mainly consists of three parts including pre-processing, online discrimination based on *dSRC*, and post-processing.

In order to improve the efficiency of text detection,the *MSER*(maximally stable extremal regions) method is first applied to preprocess the input image.Then for each *MSERs* component, the *HoG* feature extracted from a sliding window is used to form the sparse

Algorithm 1 Discriminative Dictionary Learning (*DDL*)

Input: c kinds of over-complete dictionaries $D^j = \left[d^j_{\cdot 1}, d^j_{\cdot 2}, \cdots, d^j_{\cdot K_j}\right] \in \mathbb{R}^{d_j \times K_j}$ and the corresponding sparse coding matrices $R^j = \left[(r^j_{1\cdot})^T, (r^j_{2\cdot})^T, \cdots, (r^j_{N_j\cdot})^T\right]^T \in \mathbb{R}^{K_j \times N_j}$, $j = 1, 2, ..., c$; the percentage m of the eliminated 'non-representative' atoms;

Output: The discriminative dictionary $D^j_{:\psi^j}$ and the index set ψ^j of 'non-representative' atoms

1: **for** $j = 1$ to c **do**
2: **for** $i = 1$ to K_j **do**
3: Compute the information entropy $H^j(i)$ according to Eq.(7);
4: Compute the reconstruction residual evaluator $E^j(i)$ according to Eq.(8);
5: Obtain the overall confidence $C^j(i)$ of the i^{th} atom being representative one according to Eq.(9);
6: **end for**
7: Sort $C^j(i)$,$i = 1, ..., K_j$, in ascending order.
8: Choose the leading $m\%$ atoms to be 'non-representative' atoms and let $\psi^j \subseteq \{1, 2, \cdots, K_j\}$ be the index set of them in the original atom set.
9: **end for**

representation.Following the online discrimination based on *dSRC*, the heuristic based post-processing for text-line detection and word segmentation is further implemented.

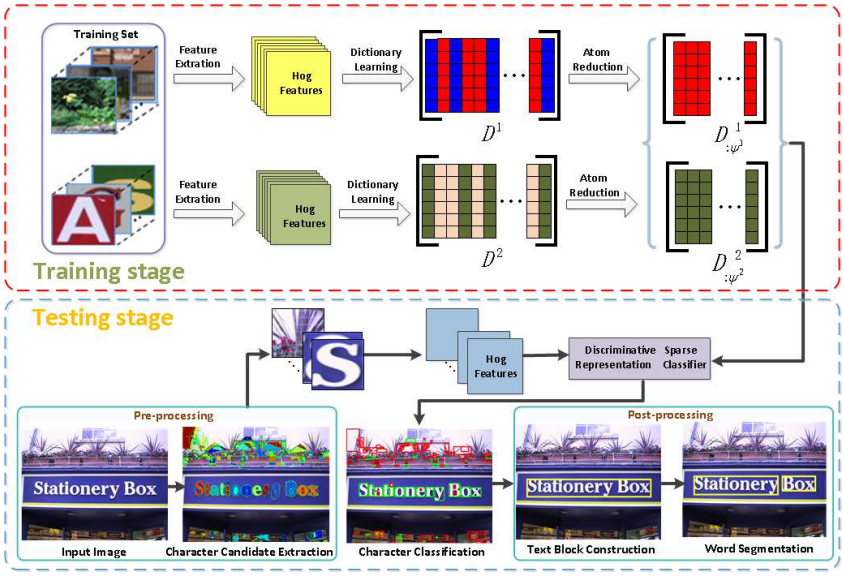

FIGURE 1. The framework of text detection based on discriminated dictionary learning

4.2. **MSER Component Generation.** Maximally stable extremal regions algorithm has originally been proposed for blob detection in image. It extracts from an image a number of co-variant regions, called *MSER*. An *MSER* is a stable connected component of some level sets of the image whose pixels have intensity contrast against its boundary pixels [17].A low contrast value would generate a large number of low-level regions, which are separated by small intensity difference between pixels. When the contrast value increases,a low-level region can be accumulated with current level pixels or merged with other lower level regions to construct a higher level region. Specifically, an extremal region can be constructed when it reaches the largest contrast.Therefore, an *MSER* is defined as a special extremal region whose size remains unchanged over a range of thresholds. An example showing the *MSER* regions in a natural image is given in Figure 2.

As one of the most widely-used region detectors, the *MSER* has been successfully applied for text detection [19]. The success can attribute to its two promising properties. First, the *MSER* detector is computationally fast and can be implemented in linear time of

the number of pixels in an image [17]. Second, it is a powerful detector with high capability for detecting low quality texts, such as low contrast, low resolution and blurring. With this capability, *MSER* is able to detect more scene texts in natural images, leading to high recall on the detection.

However,the capability of *MSER* is penalized by the increasing number of false detections.It would substantially increase the difficulty to identity true texts from a large number of non-text false alarms, which is one of the main challenges for current *MSER* based methods.Thus,it would be more suitable for being a processing step to generate some candidate objects,which can be furthered verified by some other methods. In this paper,as shown in Figure 1,a discrimination method is adopted for the goal of verification.

FIGURE 2. *MSER* regions generated in a natural image

4.3. **Post-processing for Text Block Construction and Word Segmentation.**

4.3.1. *Text Block Construction.* By applying the discrimination method based on discriminative sparse representation classifier as shown above, majority of non-text regions can be filtered out. To further increase the reliability of the algorithm, we continue a step forward to consider grouping multiple letters into a text block. First, we group two neighboring components into a pair if they have similar geometric and heuristic properties, such as similar intensity, color, height, and aspect ratios. Then, the pairs containing a same component and having similar orientations are merged sequentially to construct the final text block. The rules for merging operation are as follows:

$$\begin{aligned} {}^{1}\!/_{T_1} < height(C_i)/height(C_j) < T_1 \quad & |centroid(C_i) - centroid(C_j)| \leq T_2 \\ {}^{1}\!/_{T_3} < \ Area(C_i)/Area(C_j) \ < T_3 \quad & ||Color(C_i) - Color(C_j)||_F^2 \leq T_4 \end{aligned} \tag{11}$$

where $height(C)$, $centroid(C)$, $Area(C)$ and $Color(C)$ are the height , centroid , area, and color feature in RGB space of the detected text component, respectively, and T_1 , T_2 , T_3, and T_4 are corresponding thresholds for merging operation.

4.3.2. *Word Segmentation.* Following text block construction , we finally need to segment the text block into words. Here, an intuitive method based on horizontal and vertical projections for measuring the interval between consecutive characters is utilized for such task. The detail process for performing word segmentation is shown in figure 3.

5. **Experimental Results and Analysis.**

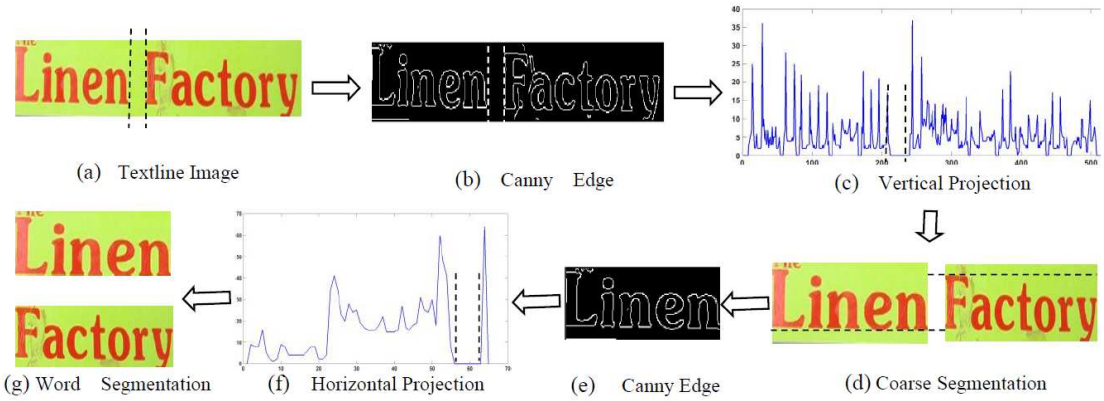

FIGURE 3. Illustration of word segmentation

5.1. **Datasets and Evaluation Methods.** In this section, experiments are performed on *ICDAR 2003* Robust Reading Competition databases[8], which includes 509 color images having sizes varied from 307 × 93 to 1280 × 960. 258 images are included in the training set, while the remaining are used for test. For evaluation, we followed the *ICDAR 2003* competition evaluation protocol, and the evaluation is computed by *Precision*, *Recall*, and *F-measure* which are defined bellow:

$$\begin{aligned} Precision &= \frac{\sum_{R_e \in E} m(R_e, T)}{|E|} \\ Recall &= \frac{\sum_{R_t \in T} m(R_t, E)}{|T|} \\ F &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{aligned} \tag{12}$$

where $\mathrm{m}(r, D) = \max\{m_p(r, r^*) | r^* \in D\}$ and $m_{\mathrm{p}}(r, r^*) = Area(r \cap r^*)/Area(r \cup r^*)$. T and E are the set of ground-truth rectangles and the set of detected rectangles, respectively.

5.2. **Performance of the Proposed Discriminative Sparse Representation Classifier.** In order to evaluate the classification performance of the proposed discriminative sparse representation classifier (*dSRC*) with atom reduction. To train the *dSRC* classifier, we collect 7000 text images and 10000 non-text images from *ICDAR 2003* dataset. Then, we randomly select 2500 text images and 3500 non-text images,respectively, to build the training dataset, and all the rest images are used as the testing dataset. Some positive and negative samples are shown in Figure 4.

(a) (b)

FIGURE 4. Part of samples for training discriminative sparse representation classifier. (a) some positive samples (text) (b) some negative samples (non-text from background)

For each of training sample, we first resize it into 64 × 64, and then extract a 512-dimensional *HoG* feature to form the representation. When using *K-SVD* method to learn the over-complete dictionary, there exists a sparse factor to control the sparseness

of the dictionary. With the sparse factor to be set as $L = \{10, 15, 20, 25, 30\}$, Figure 5 gives the performance comparison of *SRC* and the proposed *dSRC*. Clearly, the performance of *dSRC* greatly outperforms *SRC*, which shows the *dSRC* has more powerful discrimination by atom reduction. In addition, when the sparse factor $L = 20$, *dSRC* achieves the best results. In the following experiments for evaluating text detection performance, the sparsity factor L is constantly set to be 20.

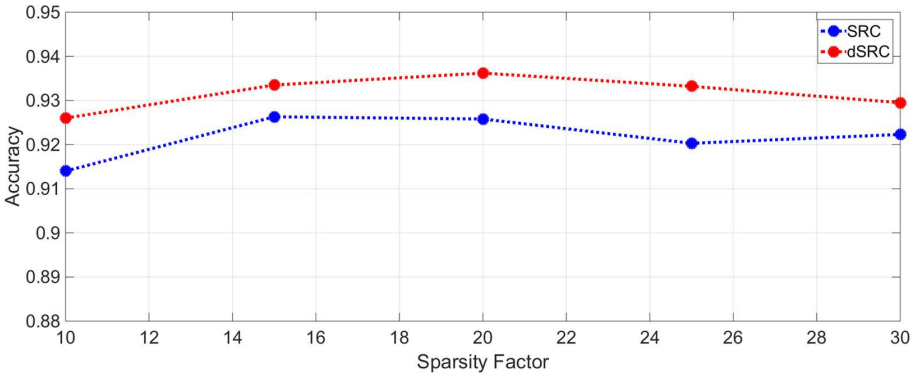

FIGURE 5. Performance comparison of *SRC* and the proposed *dSRC* with different sparse factors

5.3. **Text Detection Performance Comparison.** The text detection performance comparisons of our method with other state of the art algorithms are presented in Table 1. As can be seen, the proposed method shows excellent performance on the *ICDAR 2003* dataset in terms of *Precision*, *Recall*, and F-measure evaluation criterion. In addition, Figure 6 gives text detection results on some natural images. Visually, the proposed method achieves satisfactory detection results.

TABLE 1. Experimental results on the *ICDAR 2003* dataset

Method	Recall	Precision	F-measure
Li [18]	0.59	0.59	0.59
Neumann [5]	0.59	0.55	0.57
Zhang [20]	0.67	0.46	0.55
Liu [2]	0.66	0.46	0.54
Zhou [16]	0.57	0.50	0.53
Ashida [15]	0.55	0.46	0.50
Our Method	0.57	0.64	0.60

6. **Conclusions.** In this paper, we proposed a novel dictionary learning method for text detection. Based on the consideration that not all atoms of a given over-complete dictionary play the same roles in data reconstruction, we proposed an atom reduction based discriminative dictionary learning, which was further used to form a more discriminative sparse representation classifier, i.e., *dSRC*. Due to the powerful discrimination ability of *dSRC*, its application to the on-line text detection achieves satisfactory performance.

7. **Acknowledgements.** This work was supported in part by the National Natural Science Foundation of China (No.61532005 and No.61572068), the Program for Changjiang Scholars and Innovative Research Team in University (No.IRT201206), the Program for New Century Excellent Talents in University (No.13-0661), and the Fundamental Research Funds for the Central Universities (No. 2015JBM039).

REFERENCES

[1] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, *Proceedings of the Conference on Computer Vision and Pattern Recognition*, vol.2, pp.886–893, 2005.

[2] Z. Liu and S. Sarkar, Robust outdoor text detection using text intensity and shape features, *Proceeding of International Conference on Pattern Recognition*, pp.1491–1496, 2008.

[3] Y. Zhong, H. Zhang, and A. K. Jain, Automatic caption localization in compressed video, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.4, pp.385–392, 2000.

[4] Q. Ye, Q. Huang, W. Gao, and D. Zhao, Fast and robust text detection in images and video frames, *Image and Vision Computing*, vol.23, pp.565–576, 2005.

[5] L. Neumann, and J. Matas, A method for text localization and recognition in real-world images, *Proceeding of Asian Conference on Computer Vision*, pp.770–783, 2010.

[6] B. Epshtein, E. Ofek, and Y. Wexler, Detecting text in natural scenes with stroke width transform, *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp.2963–2970, 2010.

[7] E. Y. Kim, K. Jung, K. Y. Jeong, and H. J. Kim, Automatic text region extraction using cluster-based templates, *Proceedings of International Conference on Advances in Pattern Recognition and Digital Techniques*, Calcutta, India, pp.418–421, 1999.

[8] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, ICDAR 2003 robust reading competitions, *International Conference on Document Analysis and Recognition* , 2003.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry,and Y. Ma, Robust face recognition via sparse representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol.31,no.2, pp.201–227,2008.

[10] H. X. Wang, Z. M. Lu, and Y. Zhang, A sparse representation based super-resolution image reconstruction scheme utilizing dual dictionaries, *Journal of Information Hiding and Multimedia Signal Processing*, vol.5, no.4, pp.690–700, 2014.

[11] W. Pan, T. D. Bui, and C. Y. Suen, Text detection from scene images using sparse representation, *Proceeding of International Conference on Pattern Recognition* , pp.1–5, 2008.

[12] M. Aharon, M. Elad, and A. M. Bruckstein, The KSVD: an algorithm for designing of overcomplete dictionaries for sparse representations, *IEEE Transactions on Signal Processing* , vol.54, no.11, pp.4311–4322, 2006.

[13] M. Zhao, S. T. Li, and J. Kwok, Text detection in images using sparse representation with discriminative dictionaries, *Image and Vision Computing* , vol.28, no.12, 2010.

[14] L. J. Zhou, Y. K. Xu, Z. M. Lu, and T. Y. Nie, Face recognition based on multi-wavelet and sparse representation, *Journal of Information Hiding and Multimedia Signal Processing*, vol.5, no.3, pp.399–407, 2014.

[15] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, ICDAR 2003 robust reading competitions, *Proceedings of International Conference on Document Analysis and Recognition* , pp.682–687, 2003.

[16] G. Zhou, Y. Liu, Z. Tian, and Y. Su, A new hybrid method to detect text in natural scene, *Proceeding of International Conference on Image Processing* , pp.2653–2656, 2011.

[17] D. Nister and H. Stewenius, Linear time maximally stable extremal regions, *Proceedings of the European Conference on Computer Vision*, pp.183–196, 2008.

[18] Y. Li and H. Lu, Scene text detection via stroke width, *Proceedings of International Conference on Pattern Recognition* , pp.681-684, 2012.

[19] X. C. Yin, X. Yin, K. Huang, and H. Hao, Robust text detection in natural scene images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.36, no.5, pp.970–983, 2014.

[20] J. Zhang and R. Kasturi, Text detection using edge gradient and graph spectrum, *Proceeding of International Conference on Pattern Recognition*, pp.3979–3982, 2010.

FIGURE 6. Text detection results on some natural images by using the proposed method