Text Detection in Digital Images Captured with Low Resolution Under Nonuniform Illumination Conditions

Julia Diaz-Escobar[1(✉)] and Vitaly Kober[1,2]

[1] Department of Computer Science, CICESE, B.C. 22860 Ensenada, Mexico
jdiaz@cicese.edu.mx, vkober@cicese.mx
[2] Department of Mathematics, Chelyabinsk State University, Chelyabinsk, Russian Federation

Abstract. The text detection task becomes difficult when the image content is complex. Nonuniform illumination, camera perspective, low resolution, complex backgrounds and others, are some of new challenges. Nowadays, most of digital information is obtained using mobile devices. In particular, digital images with textual content bring us useful information which leads to the development of helpful applications such as document classification, augmented reality, language translator, text to voice converter, multimedia retrieval, and so on. However, most of existing text recognition methods are not invariant to illumination, low resolution or geometric distortions. In this work, a method for text detection using adaptive synthetic discriminant functions and a synthetic hit-miss transform is proposed. The suggested method is based on threshold decomposition and a bank of adaptive filters. Finally the performance of the proposed system is tested in terms of miss detections and false alarms with help of computer simulations.

Keywords: Text detection · OCR · Nonuniform illumination

1 Introduction

Nowadays, Optical Character Recognition (well-known as OCR) is considered by many researches as a solved problem when digital images are obtained from scanners [1]. However, in the last years new imaging devices have been developed, including smartphones, digital cameras, web cams, and so on. As a result, digital images are the most important source of information and millions of images are shared every day. In particular, digital images with textual content bring us useful information obtained from everywhere: documents, street signs, books, signboards and so on, which leads to development of helpful applications such as document classification, augmented reality, language translator, text to voice converter, industrial automation, multimedia retrieval and much more.

Unfortunately, traditional OCR engines often fail due to complexity of the imagery, becoming more complicated recognition tasks. Nonuniform illumination, camera perspective, resolution, CCD noise, complex backgrounds and others, are some of new challenges.

J.F. Martínez-Trinidad et al. (Eds.): MCPR 2016, LNCS 9703, pp. 3–12, 2016.
DOI: 10.1007/978-3-319-39393-3_1

Text detection is one of the first stages in character recognition task. OCR techniques consider simple backgrounds without geometric distortions or illumination variations and text detection is usually obtained by only image binarization. However, known binarization and segmentation techniques often fail in nonuniform illumination or low resolution conditions affecting the overall system performance.

Many techniques have been explored to solve the text detection problem. The fundamental goal is to determine whether or not there is text in a given image. Connected Component Analysis (CCA), sliding window classification, Stroke Width Transform (SWT), Maximally Stable Extremal Regions (MSER), and others are some of the state-of-the-art approaches for the extraction of textual information from imagery. For a deep explanation we refer to the following survey [1–3].

The local operator SWT [4] computes the stroke width for each image pixel, then places with similar stroke width can be grouped together into bigger components that are likely to be words. More recently, MSER approach [5] have become one of the basic methods for detection of text in imagery. Newmann and Matas proposed to use all extremal regions whereupon classification is improved using more computationally expensive features [6,7]. More recently, the same authors developed the FASText algorithm based on the well-known FAST corner detector to obtain character strokes as features for AdaBoost classifier [8]. On the other hand, Yin et al. use the MSER method to extract character candidates and then they are grouped in text candidates using single-link clustering [9].

However, most of existing text recognition methods are not invariant to nonuniform illumination, low resolution or geometric distortions. In this work, a method for text detection using adaptive Synthetic Discriminant Functions (SDF) [10] and Synthetic Hit-Miss Transform (SHMT) [11] is proposed. The suggested method is based on threshold decomposition and a bank of adaptive SDF filters. The filters are designed by incorporating information from a set of training images. Finally, the performance of the proposed method is tested in terms of miss and false detections with the help of computer simulation.

The paper is organized as follows. In Sect. 2, threshold decomposition, SDF filters and SHMT are recalled. In Sect. 3, the proposed text detection method is described. In Sect. 4, computer simulation results are presented and discussed. Section 5, summarizes our conclusions.

2 Background

In this section we briefly describe some of techniques used for the proposed text detection method.

2.1 Threshold Decomposition

In accordance with the concept of threshold decomposition, a halftone image $S(x, y)$ with Q quantization levels can be represented as a sum of binary slices $\{S_q(x, y), q = 1, ..., Q - 1\}$ as follows [12]:

$$S(x,y) = \sum_{q=1}^{Q-1} S_q(x,y), \tag{1}$$

with

$$S_q(x,y) = \begin{cases} 1, & \text{if } S(x,y) \geq q \\ 0, & \text{otherwise} \end{cases} . \tag{2}$$

2.2 Synthetic Discriminant Functions

The SDF filter is designed to yield a specific value at the origin of the correlation plane in response to each training image [10]. A SDF filter can be composed as a linear combination of the images of training set $\mathrm{T} = \{t_i(m,n), i = 1, ..., N\}$, where N is the number of available views of the target. Let u_i be the value at the origin of the correlation plane $c_i(m,n)$, produced by the filter $h(m,n)$ in response to a training pattern $t_i(m,n)$, as follows:

$$u_i = c_i = t_i \otimes h \ , \tag{3}$$

with $\otimes$ the correlation operator and

$$h(m,n) = \sum_{i=1}^{N} w_i t_i(m,n), \tag{4}$$

where the coefficients $\{w_i, i = 1, ..., N\}$ are chosen to satisfy the prespecified output u_i for each pattern in T.

Using vector-matrix notation, we denote by $\mathbf{R}$ a matrix with N columns and d rows (number of pixels in each image) where each column is given by the vector version of $t_i(m,n)$. Let $\mathbf{u} = [u_1, ..., u_N]^T$ the desired responses to the training patterns, and $\mathbf{S}$ the matrix whose columns are the elements. Equations (3) and (4) can be rewritten as follows:

$$\mathbf{u} = \mathbf{R}^{+}\mathbf{h}, \tag{5}$$

$$\mathbf{h} = \mathbf{R}\mathbf{a}, \tag{6}$$

with $\mathbf{a} = [w_1, ..., w_N]^{\mathrm{T}}$ a vector of coefficients, where the superscripts $^{\mathrm{T}}$ and $^{+}$ denotes transpose and conjugate transpose, respectively. By substituting (6) into (5) we obtain,

$$\mathbf{u} = (\mathbf{R}^{+}\mathbf{R})\mathbf{a}. \tag{7}$$

The $(i,j)'th$ element of the matrix $\mathbf{S} = \mathbf{R}^{+}\mathbf{R}$ is the value at the origin of cross-correlation between the training patterns $t_i(m,n)$ and $t_j(m,n)$. If the matrix $\mathbf{S}$ is nonsingular, the solution of the equation system is given by:

$$\mathbf{a} = \mathbf{S}^{-1}\mathbf{u} \tag{8}$$

and the filter vector is:

$$\mathbf{h} = \mathbf{R}\mathbf{S}^{-1}\mathbf{u}. \tag{9}$$

The SDF filter with equal output correlation peaks can be used for intraclass distortion-invariant pattern recognition. This can be done by setting all elements of $\mathbf{u}$ to unity.

2.3 Hit-Miss Transform

Consider a composite Structural Element (SE) $B = (B_1, B_2)$ with $B_1 \cap B_2 = \emptyset$. The set of points at which the shifted pair (B_1, B_2) fits inside the image I is the hit-miss transformation $(\odot)$ of X by (B_1, B_2):

$$\mathrm{I} \odot B = (\mathrm{I} \ominus B_1) \cap (\mathrm{I} \ominus B_2), \tag{10}$$

where $\ominus$ is the erosion operator.

Doh et al. [11] proposed a SHMT for the recognition of distorted objects. The algorithm uses SDF filters (see Sect. 2.2) as Structural Elements (SE) for distortion-invariant recognition.

Using the synthetic hit SE, H_{SDF}, as the linear combination of the hit reference images $\{H_i, i = 1, .., k\}$ and the synthetic miss SE, M_{SDF}, as the linear combination of the miss reference images $\{M_i, i = 1, .., k\}$, the proposed synthetic SEs are defined as follows:

$$H_{\mathrm{SDF}} = \sum_{i=1}^{k} a_i H_i \;\; and \;\; M_{\mathrm{SDF}} = \sum_{i=1}^{k} b_i M_i. \tag{11}$$

Let I be a binary image and I^c be the complement of I, using the synthetic hit SE, H_{SDF} and the synthetic miss SE, M_{SDF}, the proposed SHMT is given as follows:

$$X \odot (H_{SDF}, M_{SDF}) \cong (I \otimes H_{SDF})_{T_{\mathrm{H}}} \cap (I^c \otimes M_{SDF})_{T_{\mathrm{M}}}, \tag{12}$$

where T_{H} is the hit threshold, T_{M} is the miss threshold, and $\cap$ is the intersection operator.

3 Proposed Text Detection Method

To solve the text detection problem, we propose to use the threshold decomposition approach and the SHMT to obtain invariance to nonuniform illumination, noise and slight geometric distortions.

3.1 Adaptive SDF Filters

Based on the work of Aguilar-Gonzalez et al. [13], we design a bank of adaptive SDF filters to obtain distortion invariance. Each filter is created using a modification of the adaptive algorithm proposed by Gonzalez-Fraga et al. [14]. In contrast to the Gonzalez-Fraga's algorithm, we want to recognize a set of characters with the help of SDF filters. The adaptive algorithm for the design of SDF filters is presented in Fig. 1. The algorithm steps can be summarized as follows:

1. Compose a basic SDF filter using the training image of prior known views of a character using (9).

2. Correlate the resulting filter with an image containing all the remaining characters in the scene and find the maximum in correlation plane.
3. Synthesize a pattern to be accepted at the location of the highest value in the correlation plane and include it in the training set of true objects.
4. If the number of images is greater or equal to a prescribed value, the algorithm is finished, else go to step 2.

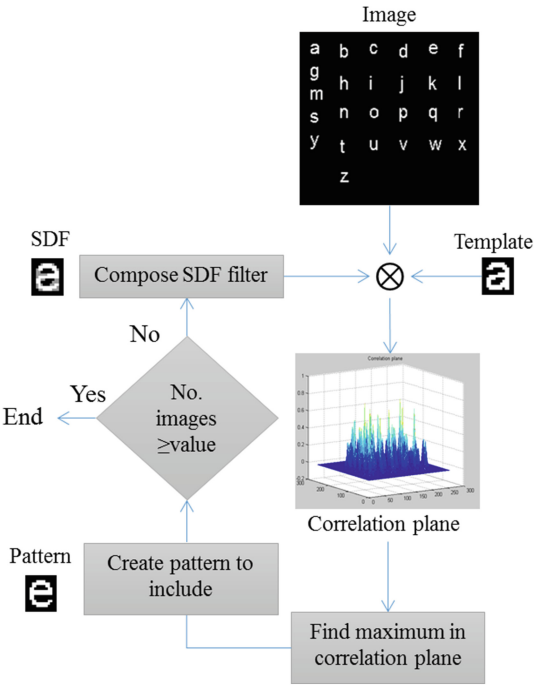


Fig. 1. Block-diagram of the adaptive algorithm for SDF filter design.

As a result we obtain a bank of composite filters. The number of filters depends of the complexity of the geometric distortions. There exists a trade-off between the number of filters and the time consuming.

3.2 Detection Method

In the first stage, using the threshold decomposition concept (described in Sect. 2.1), the image I and it complement I^c are decomposed into binary slices, $\{I_q(x,y),\ q = 1,...,Q-1\}$ and $\{I_q^c(x,y), q = 1,...,Q-1\}$, respectively. Each binary image is correlated with each filter of the bank, as follows:

$$C_q(x,y) = I_q(x,y) \otimes H_{\text{SDF}} \tag{13}$$

and

$$C_q^c(x,y) = I_q^c(x,y) \otimes M_{\mathrm{SDF}}. \tag{14}$$

Then all correlation planes $C_q(x,y)$ are thresholded by a predefined value T_{H} and T_{M},

$$(C_q(x,y))_{T_{\mathrm{H}}} = \begin{cases} 1, & \text{if } C_q(x,y) \geq T_{\mathrm{H}}, \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

and

$$(C_q^c(x,y))_{T_{\mathrm{M}}} = \begin{cases} 1, & \text{if } C_q^c(x,y) \geq T_{\mathrm{M}} \\ 0, & \text{otherwise} \end{cases}, \tag{16}$$

respectively. Then $SHMT_q$ is obtained by intersection of each pair of binary slices,

$$SHMT_q(x,y) = (C_q(x,y))_{T_{\mathrm{H}}} \cap (C_q^c(x,y))_{T_{\mathrm{M}}}, \tag{17}$$

and, finally, the detection is carried out as union of all $SHMT_q$ results,

$$SHMT_{\mathrm{u}} = \bigcup_{q=1}^{Q-1} SHMT_q. \tag{18}$$

Fig. 2 shows the block-diagram of the proposed method.

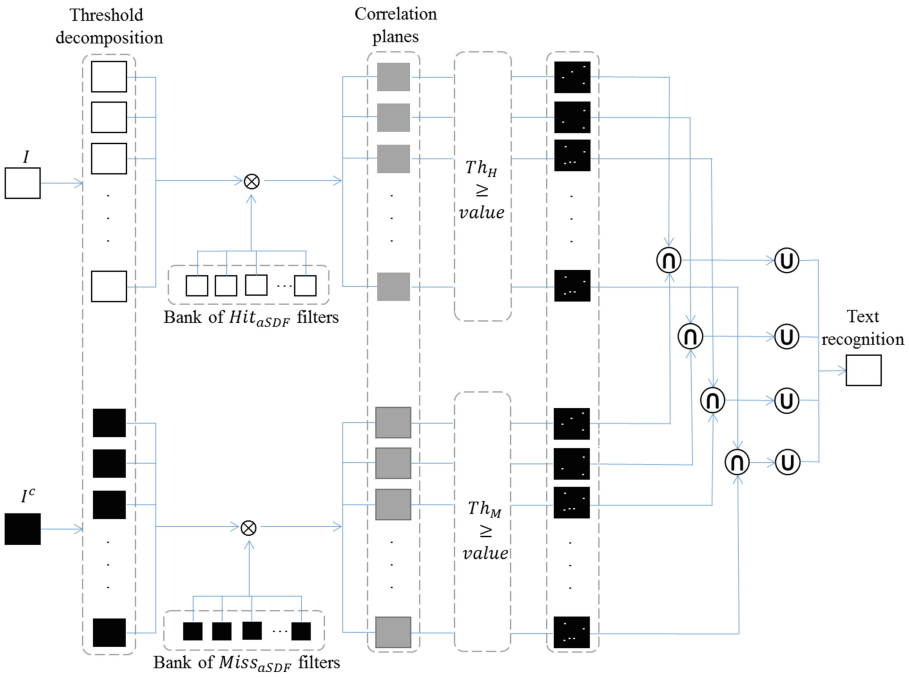


Fig. 2. Diagram of the proposed text detection method.

4 Computer Simulations

In this section we present the results of computer simulations. The performance of the proposed filters is evaluated in terms of false alarms and miss detections.

The size of all synthetic grayscale images used in experiments is 256×256 pixels and the size of character templates is 15×14, using Arial font with size of 16.

In order to analyze the tolerance of the proposed method to geometric distortions (rotation, scaling and shearing) and degradations (noise and nonuniform illumination) we perform experiments using synthetic images, Fig. 3 shows some examples.

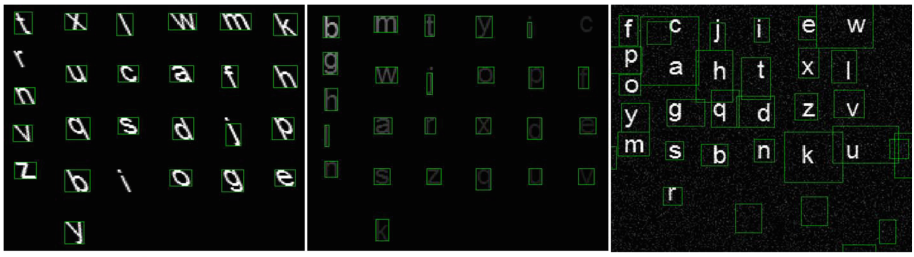
We perform 30 experiments for each geometric distortion and degradation changing the character position randomly. The simulation results yield detection errors below than 2 % except for additive noise degradation, where false detections occur. Since false detections could be eliminated in the recognition stage, we do not worry about them. Tables 1 and 2 show the results of geometric distortions and degradations, respectively, in terms of False Positives (FP) and False Negatives (FN).

Inhomogeneous illumination is simulated using a Lambertian model [15],

$$d(x,y) = \cos\{\frac{\pi}{2} - \operatorname{atan}[\frac{\rho}{\cos(\phi)}[(\rho\tan(\phi)\cos(\varphi) - x)^2 + (\rho\tan(\phi)\sin(\varphi) - y)^2]^{-1/2}]\}, \tag{19}$$

Table 1. Tolerance of the proposed method to geometric distortions.

Geometric distortions	Interval	Step size	FP (%)	FN(%)
Rotation	[-30,30]	3	0.00	0.00
Scaling	[0.8,1.5]	0.1	0.00	1.15
Shearing	[-0.5,0.5]	0.1	0.00	1.67



a) b) c)

Fig. 3. Example of synthetic images: (a) shearing by factor of 0.5, (b) nonuniform illumination with $\rho = 20$, (c) additive noise with $\sigma = 10$.

where $d(x, y)$ is a multiplicative function which depends on the parameters ρ that is the distance between a point in the surface and the light source, and ϕ, φ that are tilt and slang angles, respectively. For our experiments we use the following parameters: $\phi = 65$, $\varphi = 60$ and varying the parameter ρ in the range of $[5, 50]$ (see Table 2).

Table 2. Tolerance of the proposed method to degradations.

Degradations	Interval	Step size	FP (%)	FN(%)
Illumination	[5,50]	5	0.00	0.64
Additive noise	[0,10]	1	11.41	1.03
Impulsive noise	[-0,0.5]	0.05	0.13	0.90

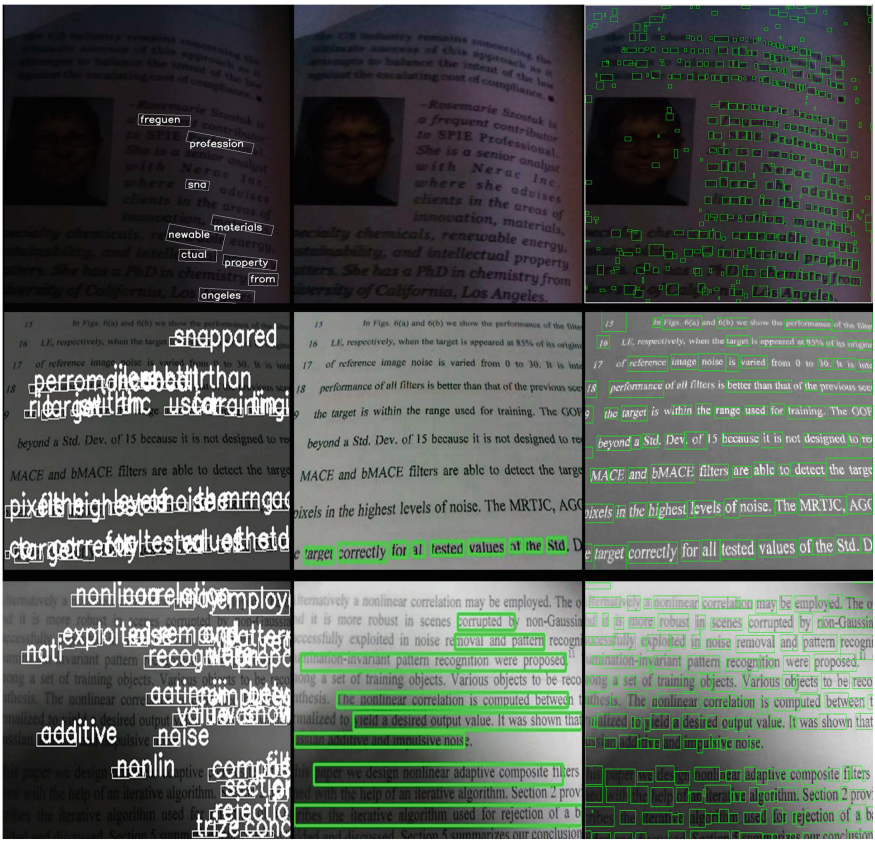


Fig. 4. First column: Neumann's TextSpotter detector, second column: Yin's TextDetector, third column: proposed text detector.

4.1 Real Images

Finally, some preliminary experiments were performed with real images to compare the proposed method with TextSpotter[1] by Neumann et al. [8] and TextDetector[2] by Yin et al. [9] (both described in Sect. 1), using real images. Figure 3 shows the results.

The same image for the three detectors was used. The resulting images look a little different due to the processing of each detector, the best results are obtained with the proposed method.

5 Conclusion

In this work we proposed a new method for text detection in degraded images using the threshold decomposition and adaptive synthetic hit-miss transform. The suggested text detector is robust to slight geometric distortions and degradations such as nonuniform illumination, noise and low resolution. In future we continue to improve the text detection and recognition algorithms to create a real-time OCR system, which is able to reliably recognize characters in low quality images.

Acknowledgments. This work was supported by the Ministry of Education and Science of Russian Federation (grant 2.1766.2014K).

References

1. Qixiang, Y., Doermann, D.: Text detection and recognition in imagery: a survey. Pattern Anal. Mach. Intell. **37**(7), 1480–1500 (2015)
2. Zhu, Y., Yao, C., Bai, X.: Scene text detection and recognition: recent advances and future trends. Front. Comput. Sci. **10**(1), 19–36 (2015)
3. Zhang, H., Zhao, K., Song, Y., Guo, J.: Text extraction from natural scene image: a survey. Neurocomputing **122**, 310–323 (2013)
4. Epshtein, B., Eyal, O., Yonatan, W.: Detecting text in natural scenes with stroke width transform. In: Computer Vision and Pattern Recognition, pp. 2963–2970 (2010)
5. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vis. Comput. **22**(10), 761–767 (2004)
6. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6494, pp. 770–783. Springer, Heidelberg (2011)
7. Neumann, L., Matas, J.: Real-time scene text localization and recognition. Computer Vision and Pattern Recognition, pp. 3538–3545 (2012)
8. Busta, M., Neumann, L., Matas, J.: FASText: Efficient Unconstrained Scene Text Detector. Computer Vision, pp. 1206–1214 (2015)

[1] http://www.textspotter.org/.
[2] http://kems.ustb.edu.cn/learning/yin/dtext/.

9. Yin, X.C., Yin, X., Huang, K., Hao, H.W.: Robust text detection in natural scene images. Pattern Anal. Mach. Intell. **36**(5), 970–983 (2014)
10. Casasent, D.: Unified synthetic discriminant function computational formulation. Appl. Opt. **23**(10), 1620–1627 (1984)
11. Doh, Y., Kim, J., Kim, J., Choi, K., Kim, S., Alam, M.: Distortion-invariant pattern recognition based on a synthetic hit-miss transform. Opt. Eng. **43**(8), 1798–1803 (2004)
12. Fitch, J., Coyle, E., Gallagher Jr., N.: Median filtering by threshold decomposition. Acoust. Speech Sig. Proc. **32**(6), 1183–1188 (1984)
13. Aguilar-Gonzalez, P., Kober, V., Diaz-Ramirez, V.: Adaptive composite filters for pattern recognition in nonoverlapping scenes using noisy training images. Pattern Recogn. Lett. **41**, 83–92 (2014)
14. Gonzalez-Fraga, J., Kober, V., Alvarez-Borrego, J.: Adaptive synthetic discriminant function filters for pattern recognition. Opt. Eng. **45**(5), 057005 (2006)
15. Diaz-Ramirez, V., Picos, K., Kober, V.: Target tracking in nonuniform illumination conditions using locally adaptive correlation filters. Opt. Comm. **323**(1), 32–43 (2014)