Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

Text Detection in Multi-feature Fusion Natural Scenes Based on Convolution Deep Belief Network

Yuan Zhong*, Jian'an Fang

College of Information Science and Technology, Donghua University, Shanghai 201600, China
**Corresponding author e-mail: zy849359963@163.com*

ABSTRACT. *For the traditional MSER algorithm in the background of complex scenes of text detection will appear in the case of false detection. In this paper, a method of multi-feature fusion of natural scene image pseudo-character filtering based on convolution depth confidence network is proposed. The candidate character text regions obtained from MSER algorithm were extracted and fused with LBP feature, HOG feature and CDBN feature, and finally the characters and pseudo-characters were classified by SVM classifier. And merge the resulting characters to produce the final line of text. Experimental results show that this algorithm can filter out more false character areas and improve the accuracy of text location.*

KEYWORDS: *scene text detection, lbp, hog, cdbn, svm*

1. Introduction

Extracting information from images and video is a hot area in computer science. The character text information in the image provides important information such as helping the visually impaired, robot navigation, and wearable devices. Scene text refers to the text in the current scene, including billboards and players' uniforms [1]. There are many methods for text detection, which can be roughly divided into texture-based methods, component-based methods, mixed methods and deep learning-based methods. Generally speaking, the text area should be located first, and then the text block should be extracted from the location area to prepare for further recognition. In the hybrid method, the suspected text character region is extracted by the maximum stable extremum region (MSER) algorithm, and then the real text region is distinguished from the fake text region by filtering algorithm and classifier. Finally, the text line region is merged to achieve the purpose of precise positioning. Therefore, designing a filtering algorithm that can extract more, richer and deeper character features is the foundation and core of text detection. HOG and

ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

LBP feature express the texture feature and local feature operator of character area, which can well extract the appearance and shape of local target of character area. The text background of the scene is complex, so to distinguish text and non-text areas more accurately, it needs to distinguish deeper features. At present, there are two classical models of deep learning: convolutional neural network and deep confidence network. The convolutional neural network is composed of input, hidden and output layers, in which the hidden layer is alternately composed of the convolution kernel and the lower sampling layer for feature extraction. Every time the image goes through one layer of convolution operation, more deep features will be extracted. The convolutional neural network has good adaptability to other features such as image size, displacement and rotation change, but it is not enough to extract and learn high-order statistical features. The deep confidence network [2] is composed of multiple limited Boltzmann machines stacked on top of each other. It does not rely on manual selection. It actively learns the input data and automatically mines the rich information hidden in the known data. It can extract higher-order statistical features of the image, and the lack of convolution layer operation also makes the deep confidence network more sensitive to external changes of the image. In 2009, Lee [3] et al. introduced convolution operation into deep confidence network and proposed the deep confidence network of convolution. The problems of the convolutional neural network , which is not enough to extract high-order statistical features from images, and the depth confidence network, which is sensitive to the size, displacement and rotation of images, are solved. To sum up, this paper considers to combine HOG and LBP features and convolution depth confidence network to solve the problem of text detection in complex background with low resolution, so as to improve the accuracy rate and recall rate of text detection in natural scenes.

2. Methodology

2.1 Basic content of Multi-feature Text DetectionFeature processing method

The natural scene text detection process proposed in this paper is shown in Figure 1.

ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

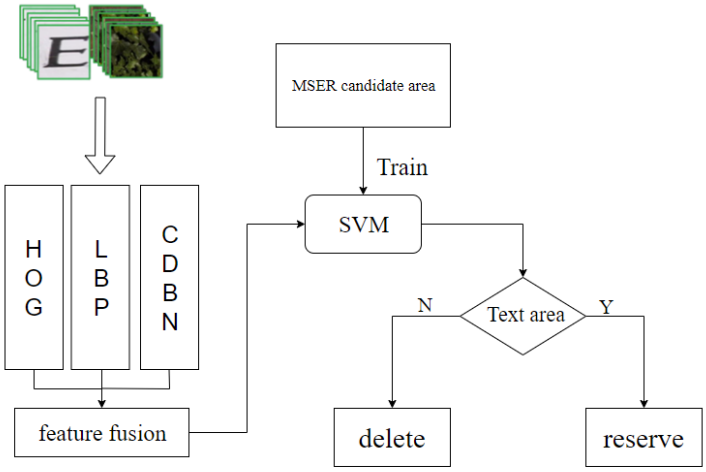

Figure. 1 natural scene text detection process

The main steps of text detection proposed in this paper are as follows:

1) Character candidate regions are obtained by the maximum extremum stable region (MSER) algorithm.

2) Image preprocessing: The character candidate areas obtained are roughly classified. Firstly, the geometric features of the areas are used to filter out some areas with aspect ratios lower than 0.1 and higher than 10, and the remaining areas are uniformly organized into 32×32 size.

3) Feature extraction: HOG, LBP and CDBN feature extraction and fusion are performed on the image obtained in the second step, so as to obtain more hidden features.

4) Classifier training and classification: the classifier output test results of the experiment for the rectangular box of character level, so we will data set of the above characteristics of the positive and negative samples are extracted and the input to the SVM to train the SVM model, and then the third step is the characteristics of the input into the trained SVM model, the character of the candidate for validation, in order to realize accurate classification filter out false regional characters.

ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

2.2 Basic technology of moving target tracking in Wireless Sensor Networks

2.1.1 Gradient direction histogram (HOG) feature

The Histogram of Oriented Gradient (HOG) was originally proposed by French researchers Dala and Triggs [4]. By calculating the gradient of pixel and making statistics, the gradient direction histogram of local area reflects the edge texture feature of character image. Character image edge gradient changes can also display the character image outline. The feature extraction steps are as follows: image preprocessing: In order to reduce the impact of light and shadow on text positioning, the image should first be γ corrected and the image color normalized. Calculation of gradient: the gradient information of character images is mainly counted, and the texture and contour information of each pixel is obtained from the gradient. Constitute the gradient histogram: The gradient obtained in the second step is taken as the HOG feature. The dimension is too high, so the method of gradient histogram is needed to count the HOG feature. Obtain the HOG feature vector: The HOG feature of the image can be obtained by concatenating all features and conducting normalization processing.1.2 Features of local binary Mode (LBP)Local Binary Pattem (LBP) is mainly an operator used to reflect the Local features of an image [5]. It was proposed by T.Ojala,M. Etikainen and D. Hirwood [6]. The original LBP operator is defined as a 3*3 window. The center pixel of the window is used as the threshold value, and the gray value of the corresponding 8 pixels is compared with it. If the value of surrounding pixels is greater than or equal to the value of the center pixel, the position of the pixel point is marked as 1, and if less than, it is 0. In this way, 8 points in the 3*3 domain window can be generated into 8-bit binary number, that is, the LBP value of the center pixel, which is used to display the texture information in the region. The calculation method is:

$$\mathrm{LBP} = \sum_{\mathrm{p}=0}^{\mathrm{P}-1} \mathrm{s}(i_p - i_c)2^p, s(x) = \begin{cases} 1, x \geq 0 \\ 0, x \leq 0 \end{cases}$$

2.1.2 Convolution depth confidence network

Convolutional Deep Belief Networks (CDBN) is a hierarchical probabilistic generation model, which can extract data characteristics automatically. The Convolutional depth confidence network is constructed by the Convolutional Restricted Boltzmann Machine (CRBM) on the basis of which the Convolutional Restricted Boltzmann Machine combines the Convolutional neural network with the Restricted Boltzmann Machine [7]. Using the limited Boltzmann machine and the convolutional neural network, the deep architecture of the convolutional depth confidence network can extract meaningful features from full-size images by generating convolutional filters, which can reduce considerable connection weights and learn spatial information from adjacent image blocks more effectively. As

ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

shown in Figure 1, the convolutional CRBM model is composed of three layers, including visual layer V, hidden layer H and pooling layer P.

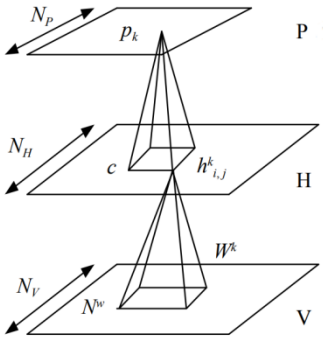

Figure. 2 Convolution of CRBM model

In this paper, the subject structure of the convolutional depth confidence network is composed of two convolutional limited Boltzmann machines (CRBM) stacked, each of which is followed by a pooling layer. The input layer of the model was set as 32×32×3(that is, three mapping layers with a size of 32×32). The first CRBM contained 9 feature mappings, with a convolution kernel size of 7×7 and a pooling size of 2×2. The second CRBM contained 16 feature mappings, with a convolution kernel size of 5×5 and a pooling size of 2×2. The learning rate was 0.05, the activation function was Sigmod function, and the hidden layer was 50% Dropout method for random dropping.

2.1.3 Support Vector Machine (SVM)

Support Vector Machines (SVM) is a discriminant classifier defined by the classification hyperplane, in essence, it is a hyperplane that can segment different categories of samples in the sample space, and it has advantages in solving small samples, two-dimensional linearly separable and high-dimensional pattern recognition. As shown in Figure 3, triangles and circles represent training samples. H2 stands for optimal hyperplane. In this chapter, SVM is used to deal with binary linear classification problem, that is, to determine which of the candidate text regions text regions are and which are non-text regions.

ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

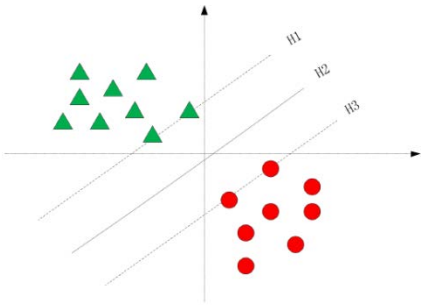

Figure. 3 Support Vector Machine

3. Results and discussion

3.1 The data set

The public data set of the current natural scene detection is used to verify the performance of our algorithm and compare it with other methods. These open data sets include the ICDAR2011 Robust Reading Competition data set and ICDAR2013 Robust Reading Competition data set.The ICDAR2011 data set contains 486 color images, including 231 images from the training set (848 words, 5500 characters) and 255 images from the test set (1189 words, 6393 characters).ICDAR2013 data set contains 462 color images, including 229 images from the training set (900 words, 5,600 characters) and 233 images from the test set (910 words, 5,550 characters).

3.2 Evaluation Methods

Recall rate (R), accuracy rate (P), and harmonic mean were used for natural scene text detection.The recall rate is defined as the ratio between the number of detected real text areas and the number of marked real text areas

$$\mathrm{R} = \frac{NumTG}{NumAG}$$

Accuracy is defined as the ratio between the number of real text areas detected and the total number of areas detected

$$\mathrm{P} = \frac{NumTG}{NumT}$$

Where NumTG represents the number of real text areas detected, NumAG represents the number of marked real text, and NumT represents the total number of detected text areas.In order to prevent the influence of some special cases on the index of the algorithm, the harmonic average is introduced

ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

$$F = 2 \times \frac{R \times P}{R + P}$$

3.3 Development and experimental environment

Hardware environment Intel (R) Core (TM) I5-6300Software environment: Windows10 enterprise edition, Matlab R2016b.In this paper, data preparation was carried out in PyCharm2019 and Opencv2.4.8 environments, and natural scene text detection based on multi-feature fusion was carried out in Matlab R2016b environments.

3.4 Experiment and results

In order to verify the effectiveness of the algorithm in this paper, other better methods on ICDAR2011 and ICDAR2013 are compared. Table 1 shows the comparison effect of text detection on ICDAR2011 data set.

Table 1 Experimental comparison results on ICDAR2011 dataset

Methods	Precision(%)	Recall(%)	F(%)
Our Method	90.53	76.77	83.23
Wang[8]	89.48	73.23	80.54
Huang[9]	86.32	71.05	78.63
Neumann[10]	85.42	67.39	75.34

It can be seen that the accuracy and recall rate of the algorithm in this paper are improved, and the accuracy is 1.05% higher than that of the CDBN feature alone, with the F value reaching 83.23%.It is mainly due to the fusion of multiple features and the addition of characters' edge texture and other low-level features compared with single convolution depth confidence network features. Figure.4 shows the partial detection results of the algorithm in ICDAR2011.

Figure. 4 The algorithm in this paper is an experimental example on ICDAR2011 data set

ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

Table 2 shows the comparison results of text detection on THE ICDAR2013 dataset.

Table 2 Experimental comparison results on ICDAR2013 dataset

Methods	Precision (%)	Recall(%)	F(%)
Our Method	91.56	80.91	85.90
Zhou[11]	89.18	80.31	84.51
Huang	91.98	77.20	84.53
Neumann	78.47	67.98	72.85

It can be seen that the accuracy and recall rate of the algorithm in this paper are improved, and the accuracy is 2.38% higher than that of CNN feature alone, with the F value reaching 85.90%.This is mainly due to the fact that the convolutional deep confidence network can learn more higher-order statistical features of images than the convolutional neural network.Figure.5 shows the partial detection results of the algorithm in ICDAR2013.

Figure. 5 Experimental example of the algorithm in this paper on ICDAR2013 data set

4. Conclusion

It can be seen from the above table that the algorithm in this paper is improved to some extent compared with the algorithm using CNN and CDBN alone, mainly due to the fact that the convolution depth confidence network can learn more higher-order statistical features of the image than CNN, and the fusion of Dot can also enable the classifier to learn the unique texture edge features of the text. In this paper, the proposed deep belief network algorithm based on multi-feature fusion is applied to text detection in natural scenes, which results in the problem of complex background and low resolution, and the accuracy rate and F value are also improved in the two data sets.

ISSN 2616-5775 Vol. 3, Issue 4: 41-49, DOI: 10.25236/AJCIS.030405

References

[1] Zhou G , Liu Y , Meng Q , et al. Detecting multilingual text in natural scene [C] International Symposium on Access Spaces. IEEE, 2011.
[2] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. Proceedings of Computer Vision and Pattern Recognition. San Francisco,CA, USA. 2010. 2963–2970.
[3] Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th Annual International Conference on Machine Learning. Montreal,QC, Canada. 2009. 609–616.
[4] Karatzas D , Shafait F , Uchida S , et al. ICDAR 2013 Robust Reading Competition [C] 2013 12th International Conference on Document Analysis and Recognition. IEEE Computer Society, 2013.
[5] Yin H F, Wu X J. A New Feature Fusion Approach Based on LBP and Sparse Representation and Its Application to Face Recognition [J]. 2013.
[6] Ojala T, Pietikäinen M, Mäenpää T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24 (7): 971-987.
[7] Lee H, Ekanadham C, Ng A Y. Sparse deep belief net model for visual area V2 [C]. International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada Curran Associates Inc, 2007: 873-880.
[8] Wang Lin, Zhang Xiao-Feng. Scene Text Detection in Convolutional Deep Belief Networks [J]. Computer Systems & Applications, 2018, 27 (6): 231–235.
[9] Huang WL, Qiao Y, Tang XO. Robust scene text detection with convolution neural network induced MSER trees.Computer Vision(ECCV 2014). Cham: Springer, 2014.497–511.
[10] Neumann L, Matas J. A Method for Text Localization and Recognition in Real-World Images [J]. 2010.
[11] Pengfei Zhou. Research on text Detection and Recognition in natural Scene images [D]. Xi 'an University of Technology, 2019.