Correction and Recognition of Oriented Scene Text Using Convolutional Neural Networks

Kiptanui Linus

Department of P.G Studies and Research in Computer Science Kuvempu University, Karnataka, India
linusunguli@gmail.com

Prabhakar C J

Department of P.G Studies and Research in Computer Science Kuvempu University, Karnataka, India
psajjan@yahoo.com

Abstract: Recognition of text in scene images plays significant role in computer vision due to its wide range of applications. A good number of existing techniques for recognition of texts in scene images only recognize horizontal or near to horizontal texts and fail to recognize oriented scene text. This proposed work, leverages the powerful concept of Alex Net Convolutional Neural Networks (CNN) to predict the rotational angle of the oriented text in the image and make the necessary correction. After correcting the text, Character segmentation of corrected text is performed using low variation Extremal Regions (ER) and extract the features from the segmented characters using Pyramid Histogram of oriented gradients (PHOG). The extracted features are then categorized into text or non-text classes using support Vector machine (SVM). Recognition is done using optical character recognition (OCR) to recognize the corrected scene text. Lastly, this work was evaluated using three benchmarks database, Street View Text (SVT), IIIT5K and ICDAR 2003 using the character rate recognition. The recognition results surpasses the existing systems by a large margin.

Keywords- Convolution Neural Networks (CNN), Pyramid Histogram of oriented gradients (PHOG).

1. Introduction

In the contemporary world, text performs a vital role in spreading and acquiring information across continuum. The text written on various objects and surfaces provide the useful information about environments, which helps various activities such as finding places, getting products and location description, identifying vehicles and many more. Therefore, natural scene images, which embed the text, play a very important role due to its text information which is an important region of interest and contains useful semantic information regarding the scene. Alternatively, the useful and accurate connotation incorporated in the text could be helpful in comprehending our surroundings. Recently, the researchers have paid much attention for text recognition in natural images on the account of its extensive uses in real-life including: computerized aid for visually impaired, assisting in safe vehicle driving, robotic navigation in urban environments, content-based image/video indexing and so on.

Recognition of text in natural scene images is notably a demanding task in machine/computer vision. This is due to unconstrained outdoor environment that poses various challenges, which includes complex backgrounds, uneven illumination, poor lighting conditions, and occlusions. Apart from the above-mentioned challenges caused by the environmental factors, also there are other factors that make the recognition of these texts even more challenging and these includes, text orientations, font and styles, blur and noise. In real-world applications, it has been observed that, many scene text images especially those found on billboards, T-shirts, signboards and house walls, in most cases are in curved shapes. The progress on different techniques reveals that most of the techniques on scene text recognition focusing only on the recognition of horizontal or near to horizontal

text in scene images and totally ignore the presence or arbitrary oriented texts, which has created a gap that needs to be filled. Thus, the techniques meant for horizontal text regions cannot be widely applied for curved text. Therefore, it is very essential to develop techniques that are capable of recognizing curved scene text, in order to utilize its rich semantic information. In this paper, in order to fill the research gap, proposes a system that is able to recognize both the horizontal and oriented texts present in scene images.

The advancement and applications of deep neural networks have motivated the researchers to develop some efficient techniques [1][2][3] for recognition of text in natural scene images. They utilize CNN to extractor features, in the input patch, recurrent neural network (RNN) for character decoding and language modeling. These techniques have achieved good accuracy with all the challenges and they are developed based on the idea that most input texts are always horizontal. These promising methods failed to detect arbitrary oriented texts which are important challenges in real time applications. The literature survey reveals that many techniques [4][1] use rectification of curved text prior to recognition. Recently, Yang et al. (2019) [5] proposed a Network that utilizes two major modules. The rectification module uses the geometric features to perform rectification of the text in the image while the recognition module translates rectified feature maps into character sequence. In our work, we employed well trained CNN for rectification of curved text prior to recognition where CNN can be trained to learn the features that are very important in predicting the orientation of the text image.

Rectification of oriented scene text images is more challenging because the text image has no vertical or horizontal lines that can be used to make the correction. In this paper, we propose an algorithm that is able to recognize oriented text present in natural scene images. The major idea of our system is to utilize the powerful concept of CNN in predicting the rotational angle of the oriented text and make the necessary correction. After tilt correction has been done, we perform character segmentation using low variation Extremal Regions (ER) [6], from the segmented characters; we extract features using Pyramid Histogram of Oriented Gradients (PHOG)[7]. We then classified these extracted features into two classes that is, text and non-text using Support Vector Machine (SVM) [8]. Lastly, we perform candidate character recognition using the conventional Optical Character Recognition (OCR) [9]. Our major contribution is a novel scene text recognition system, which is able to recognize both horizontal and oriented scene texts by combining the powerful concept of convolution neural networks and conventional methods. This is evident in our results.

2. Related work

Scene text recognition is an ongoing field of study in computer vision communities and a good progress realized. Recognizing text in natural scenes has attracted researchers in recent years due to its importance and challenges. However, most recognition techniques that are available only focus on horizontal scene texts and less addresses the issue of oriented scene texts. In comparison to regular text recognition, recognizing irregular text of arbitrary shape is much more challenging. In this section, we center our attention on research related to recognition of oriented scene texts as proposed by various scientists. The traditional methods use local features to recognize texts in images, some of these local features include: stroke width transformations, connected components and histogram of oriented gradients among others. Nonetheless, these techniques do not achieve the expected results due to their low capacity features. The introduction of new technologies like deep learning has led to development of new methods for text recognition. Deep learning techniques have proved to work extremely well in scene text recognition in terms of accurate and efficiency. Jaderberg et al. (2014) [10] proposed a deep features technique for recognizing texts in scene images where they use sliding window for

character segmentation and deep neural network for character classification. Spatial Transformation Networks (STN) has been applied prior to recognition to align text image into horizontal and regular character widths and heights [11] [Gao, Y et al., 2018). Baoguang et al. (2016)[4] proposed a Robust text recognizer with Automatic Rectification (RARE). This method is designed to recognize irregular texts that include curved and perspective scene texts. The model is a deep neural network made up of STN and a Sequence Recognition Network. The first module (STN) performs the rectification of the images. The second module (SRN) does the recognition of the corrected image produced in the first module (horizontally aligned text). The major weakness of this method is that, the rectification done by STN is not accurate and at the same time, the recognition by RARE fails if the orientations of the images are substantial.

Shi et al. (2016) [1] proposed a technique for recognizing scene texts through flexible rectification (ASTER), for scene text recognition. This model has two major networks for rectification and the other network for recognition. The first network corrects the input image through parameterized Thin Plate Spline (TPS) which is able to handle various text orientations. The second Network performs character predictions from the corrected image. The major drawback of this technique is that, text detection focuses only around the target texts.

Liu et al. (2018) [12] proposed a technique that uses text images that has no impairments to perform supervised learning. The major drawback is that, it does not work with images containing complex backgrounds. Li et al. (2018) proposed a semi-supervised Spatial Transformation Network, which brought extra-supervision to (STN). The system learns through identification of points of control distributed throughout the text edges. They should be symmetrically aligned along the text center. The major drawback of this technique is that, it does not take into consideration the various control points during rectification which causes rectification problems.

Bai et al. [13] used multi-scaled representation of characters that captures the character information. Based on this information, the characters can be recognized. Goel et al. [14] utilizes sub-image features through comparisons with lexicon words. Weinman et al. [15] proposed an algorithm that recognizes characters by relying on character classifier, consolidated in the text image .Wei et al. [16] leverages the interpolation artifacts caused by rotating the images. It fails on images that are not upright. Sun et al. [17] proposed a method that ascertains both the orientation and the slant angles of the image and the text characters respectively, using gradient orientation histogram, based on the fact that the gradient orientation should be perpendicular to the text line. The image is then corrected by rotating based on orientation angle.

3. Methodology

The proposed methodology for recognition of oriented scene text involves two modules where, the first module performs correction or alignment of oriented/curved scene text using pre trained CNN, it is then followed by second module where character segmentation and feature extraction is done using low variation extremal regions technique and pyramid histogram of oriented gradients technique respectively. The extracted features are then categorized into text or non-text classes using SVM. Finally, we perform character recognition using OCR system. The in-depth explanation of each module is presented in the subsequent sections. The proposed methodology focused on recognition rather than detection of oriented scene text. Hence, the proposed methodology assumes that the curved/oriented scene text has been detected and hence it accepts manually cropped curved scene text images where it contains text area without any complex background and sample images of curved scene text are shown in the Figure 1.

Fig.1. Sample images of curved/oriented scene texts of Street View Text (SVT) dataset.

3.1 Correction of Text Orientation

When dealing with orientation correction of curved scene text images, we realize that it is actually extremely hard to detect and correct position of the image automatically. Since most of the scene text images do not contain the vertical or the horizontal lines that helps to perform orientation correction. In this paper, we adopted the procedure used by P. Fischer et al. [18] where, the pre trained CNN considers the problem in three levels. The most challenging task is that an estimation of orientation angle of curved scene text without the prior knowledge of rough orientation angle of the image. Another challenging task is that training the CNN network for estimation of orientation angle of given curved scene text. Since we are using publicly available collection of training images for training, we randomly rotate the readily available training images of datasets by various angles before training of CNN.

The three major considered orientation levels are $\pm 30^{\circ}$,$\pm 45^{\circ}$ and $\pm 360^{\circ}$. We name these networks as N-30, N-45 and N-360 all built on Alex Net Architecture [19]. Alex Net Architecture comprises of five convolution layers, three fully connected layers and a linear unit. The network has two output units for both positive and negative orientations; the outputs are active at a time. When we have an angle α, the expected output vector is as follows; [max (0, α), max (0, −α)] trained with L1 loss. In the case of differentiating between the bottom up, portrait and land scape images, we use a network with 4 outputs that is (0, 90, 180 and 270) degrees. This is useful in determination of 360 degrees prior to evaluation of the orientation angle. Hence, the trained network can successfully predict the various orientations. In overcoming the issue of over fitting, we applied transformations on training sets while training the network as follows; gammaYϵ $[-0.5, 1]$, brightness [-0.2, 0.2], Gaussian noise α [0, 0.02]. The optimum parameters for training the network is based on Adam optimization method [20], β1 = 0.9 and β2 = 0.999.Learning rate λ = 1e−4 then reduce to λ = 2e−7. The beginning of learning was λ = 1e−6 meaning we begin with low learning rate then increase steadily until the required rate is achieved. Once we calculate the orientation angle of the curved scene text, we perform orientation correction based on the estimated orientation angle using CNN based on the technique proposed by P. Fischer et al. [18].The learning rate of this model is based on the technique proposed by Kingma et al. [20]. As follows;

Require: α: step size

Require: $\beta_1, \beta_2 \in [0,1]$Exponential decay rates

Require: $f(\theta)$Stochastic objective function with parameters θ

Require: θ_0Initial parameter vector

$m_0 \leftarrow 0$(Initialize 1st moment vector)

$u_0 \leftarrow 0$(Initialize the exponentially weighted infinity norm)

$t \leftarrow 0$(Initialize time step)

Where $\theta_t$not converged **do**

$t \leftarrow t + 1/$

$gt \leftarrow \nabla_\theta f_t(\theta_t - 1)$(Get gradients w.r.t. stochastic objective at time step t)

$m_t \leftarrow \beta_1 . m_t - 1 + (1 - \beta_1) . gt$(Update biased first moment estimate)

$u_t \leftarrow max(\beta_2 \, . u_{t-1}, |gt|)$(Update the exponentially weighted infinity norm)

$\theta_t \leftarrow \theta_{t-1} - \left(\frac{\alpha}{(1-\beta_1^t)}\right) . m_t / u_t$(Update parameter)

End while

Return θ_t(Resultant parameter).

Fig.2. Result of orientation correction using proposed method. The first row shows sample images of oriented texts and the second row shows orientation correction using proposed method based on CNN.

3.2 Character Segmentation and Features Extraction

After orientation correction of curved scene text images, we perform character segmentation which helpful for recognition of individual characters. To segment the character candidates, we use Low Variation Extremal Regions (ER). The main advantage of using ER for segmentation is that ER has pixels with either low or high values compared to the boundary pixels. These variations in pixel values are used to split the local paths into sub paths that represent character candidates. This gives a more accurate results compared to popular technique such as Maximally Stable Extremal Region (MSER) that remains stable for a range of intensity levels, this may lead to missing important character regions.

Fig.3. Result of character segmentation using proposed method for the orientation correction scene text images shown in the Figure 2.

We extracted the features from the segmented characters using Pyramid Histogram of Oriented Gradients (PHOG) technique. The PHOG extracts local features from the intended characters and passes these extracted features to the SVM, which classifies the characters as either a character or non-character based on the features. The major advantage of PHOG is that it is able to give more spatial information of the image including the representation of its shape also its computational complexity is lower compared to other techniques. The proposed work utilizes the conventional Optical Character Recognition system (OCR) for recognition where the corrected as well segmented characters of scene text is given to the OCR as the input, each character in the text is recognized by the OCR. The Figure 4, Figure 5 and Figure 6 shows recognition results of proposed method on sample images of Street View Text (SVT), IIIT5K and ICDAR2003 dataset respectively.

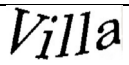
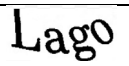
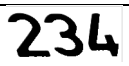
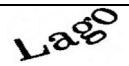
Input image oriented	Segmented image	Ground Truth	Prediction
		India	India
		Villa	Villa
		Lago	Lago
		234	234
		Lago	Failed
		State	Failed

Fig.4. The result of recognition of curved/oriented scene text using proposed method. The first column shows the original curved scene text images of Street View Text

(SVT) dataset. The second column shows the orientation correction and segmented characters. The third column shows the ground truth, and finally last column shows its corresponding predictions by the proposed approach.

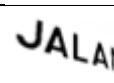
Input image	*Segmented image*	*Ground Truth*	*Predicti on*
		FLY	FAILED
		NOKIA	FAILED
		FUTURE	FAILED
		JALAN	JALAN

Fig.5. The results of recognition of oriented scene text without orientation correction. The first column shows the original oriented text, the second column shows the segmented text, the third and fourth columns shows the ground truth and predictions respectively.

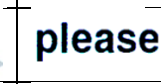
Input image	**Segmented image/corrected**	**Ground truth**	**Prediction**
		FLOOR	Floor
		PLEASE	Please
		WITH	With
		DESIGN	Design

Fig.6. *The result of recognition of oriented scene text using proposed method. The first column shows the original oriented scene text images of IIIT 5K dataset. The second olumn shows the orientation correction and segmented characters. The third column shows the ground truth, and finally, the last column shows its corresponding predictions.*

the ground truth, and finally last column shows its corresponding predictions by our approach.

4. Experimental Results

The proposed method was evaluated using three standard datasets, ICDAR 2003 [21], IIIT 5K [22] and Street View Text (SVT) [23]. Street View Text (SVT) dataset comes from the Google Street View. The dataset has 62,058 high quality images. From the dataset, 647 word images are cropped. It is challenging due to low resolution and visibility issues. ICDAR 2003 dataset contains 251 scene text images with labeled text bounding boxes. From the database, 860 text images are cropped. IIIT5K dataset has 3,000 images. From the dataset, 860 text images are selected. Character based recognition evaluation outlined in [30] was utilized as follows:

$$\text{Recognition Rate} = \frac{\text{CR x } 100}{\text{TT}}$$

where, CR is the amount of correctly recognized characters and TT is the total amount of tested characters. The recognition accuracies on these three benchmark databases, obtained by the developed framework and the recent state-of-the-arts techniques including the approaches based on deep models [24, 25, 26, 27, 28] are shown in Table 1. These results demonstrated that the developed framework extensively improves the performance of oriented text recognition even in the presence of the blurred, curved, or different orientations of text in real scene image. From the results demonstrated in Table 1 on the ICDAR 2003, SVT and IIIT5k database, it can be seen that the developed framework outperforms existing methods with respect to character recognition accuracy.

Table 1: Compare the oriented text recognition results of the proposed method with the existing methods on the ICDAR 2003, SVT and IIIT5K datasets.

Methods	Recognition Rates (%)		
	ICDAR 2003	SVT	IIIT5K
Yao et al [25]	77.58	74.65	75.35
Islam et al [26]	81.39	78.31	81.70
RARE [29]	88.75	86.28	87.56
SSDAN [27]	90.11	89.25	92.36
FACLSTM [28]	91.25	90.13	92.15
Proposed Method	92.48	92.78	93.81

5. Conclusion

In this paper, an oriented scene text recognition framework is proposed for text with contrast varies and font size varying based on orientation correction and recognition. The pre-trained Alex Net network was employed for orientation correction and conventional methods were used for recognition. It presents recognition results of the developed framework for oriented scene text using three public datasets based on character recognition metric. The proposed method corrects the orientations of the scene text images using convolutional neural networks, which depicted outstanding results in orientation correction. On the other hand, the conventional character segmentation, feature extraction and character recognition using OCR performed extremely well on the corrected scene texts. The experimental results demonstrated a great improvement in terms of character recognition and therefore, this proposed method shows an outstanding performance compared to the obtainable methods. The major advantage of using this method is that it is able to perform scene text orientation correction and recognition automatically.

6. References

1. Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence., 39(11), 2298–2304

2. Yang, C. Liang, Z. Zhou, G. Alexander, I. Ororbia, D. Kifer, and C. L. Giles,2017, Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild. InProc. Of IEEE Intl.Conf. on Computer Vision and Pattern Recognition,pp.474–483.
3. Cheng, Bai, Xu, Zheng, Pu, and Zhou, ‘‘Focusing attention Towards accurate text recognition in natural images, Oct. 2017, pp. 5086–5094.

4. Baoguang Shi, Xiang Bai,and CongYao, 2016, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, IEEE transactionson pattern analysis and machine intelligence, 39(11):2298–2304.

5. Ming Kun Yang, Yushuo Guan, Minghui Liao, Xin He, Kaigui Bian, Song Bai, Cong Yao,Xiang Bai, 2019, Symmetry-constrained Rectification Network for Scene Text Recognition, IEEE.
6. Matko Saric. Scene Text Segmentation using Low Variation Extremal Regions and Sorting Based Character Grouping, Neuro Computing, 2017
7. Zhi Rong Tan, Shangxuan Tian, and Chew Lim Tan. Using Pyramid of Histogram of Oriented Gradients on Natural scene text recognition.
8. Anurag Sarkar, Saptarshi Chatterjee, Writayan Das, Debabrata Datta. Text Classification using Support Vector Machine. PP.33-37, 2015.
9. Kai Wang, Boris Babenko and Serge Belongie. End-to-End Scene Text Recognition. November 2011.

10. Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman, 2014, Deep Features for Text Spotting. In *Computer Vision – ECCV 2014*, number 8692 in Lecture

Notes in Computer Science, pages 512–528. Springer International Publishing.

11. Gao, Y. Chen, Y. Wang, J.; Lei, Z. Zhang, X.-Y. and Lu, H., 2018. Recurrent calibration network for irregular text recognition. arXiv:1812.07145.
12. Liu Z., G. Lin, S. Yang, J. Feng, W. Lin, and W.L.Goh, 2018, Learning markov clustering networks for scene text detection," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, pp. 6936–6944.
13. X. Bai, C. Yao, and W. Liu, "Stroke lets: A learned multi-scale mid-level representation for scene text recognition," IEEE Transactions on Image Processing, vol. 25, no. 6, pp. 2789–2802, 2016.
14. Goel, Mishra, Alahari, Jawahar. "Whole is greater than sum of parts: Recognizing scene text words". IEEE International Conference on Document Analysis and Recognition. 398–402, 2013.
15. Weinman, Butler, Z, Knoll, D, &Feild, Towards Integrated Scene Text Reading. Pp. 375–387, 2014.
16. Wei, W., Wang, S., Zhang, X., Tang, Z.: Estimation of Image Rotation Angle using interpolation-related spectral signatures with application to blind detection of image forgery. Pp.507–517 , 2010.
17. Changming Sun, Deyi Si.Skew and Slant Correction for Document Images Using Gradient Direction, pp.142-146, 1997.
18. Philipp Fischer, Alexey Dosovitskiy, Thomas Brox. Image Orientation Estimation with Convolutional Networks, 2015.
19. Krizhevsky, A., Sutskever, I.,Hinton, G.E.: Image classification with deep convolutional neural networks. pp. 1106–1114, 2012.
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. 2015.
21. S.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, ICDAR 2003 robust reading competitions, pp. 682–687.
22. Mishra, A., Alahari, K., Jawahar, C.: Scene text recognition using higher order language priors. In: BMVC 2012-23rd British Machine Vision Conference. BMVA (2012)
23. K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In ICCV, 2011
24. Joanna Isabelle Olszewska. Active Contour Based Optical Character Recognition for Automated Scene Understanding. Neurocomputing, http://dx.doi.org/10.1016/j.neucom.2014.12.089.
25. Yao, C., Bai, X., & Liu, W. (2014). A unified framework for multioriented text detection and recognition. IEEE Transactions on Image Processing., 23(11), 4737–4749.
26. Islam, M.R., Mondal, C., Azam, M.K., Islam, A.S. (2016). Text detection and recognition using enhanced MSER detection and a novel OCR technique. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 15-20). IEEE
27. Zhang, Yaping, et al. Sequence-to-sequence domain adaptation network for robust text image recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019).
28. Wang, Q., Huang, Y., Jia, W., He, X., Blumenstein, M., Lyu, S., & Lu, Y. (2020). FACLSTM: ConvLSTM with focused attention for scene text recognition. Science China Information Sciences., 63(2).
29. Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, Xiang Bai. Robust scene text recognition with Automatic Rectification (RARE). Computer Vision and Pattern Recognition (2016).
30. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X. (2016). Multi-oriented text detection with fully convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4159-4167).