Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Text Extraction For Natural Scene Images Using Convolutional Neural Networks

Rashi Gupta[1] and Javed Wasim[2]

[1,2]Institute of Engineering & Technology, Mangalayatan University Aligarh, India.

Abstract

Various methods have been used for scene text detection based on neural networks and they have shown promising results with respect to different representations of text embedded in an image. The main drawbacks while detecting characters are the lack of individual character level annotations, blurred images etc. To overcome these utilised characters as basic elements which allows optimization of text with a CNN based recognition. These improvements will make the image clearer with better resolution, reliable characters recognition in a distorted. In this paper, we describe the detection of text from natural images which includes images with text having different languages, different orientations (vertical, horizontal, etc.), different styles (stray images, text shape etc. We have tested this technique on standard datasets that are as follows-ICDAR 2015, MSRA-TD 500, SVT, Personalised Dataset. The effective deftness is presumptuous for the precise results of our technique. The accuracy of the result enables us to show that our technique can be used substantially to detect text written in different languages from natural images.

Keywords: Text Detection, Segmentation, residual network, CNN, dense Network.

1. Introduction

With the advancement of interactive media data innovation, the innovation of getting valuable data from enormous information has expansive possibilities. The data contained in the text is immediate and successful. The digitization of text data is of extraordinary importance to work on the capacity of interactive media recovery, modern mechanization and scene understanding. Text recognition is a significant piece of scene understanding. Its goal is to observe a bunch of sub-locales in the picture with text and insignificant non-text content. Text discovery can be sorted in various ways, Text detection by locating text in bounding boxes (best fit or horizontal), Text extraction by binarizing the scene image such that all text pixels are foreground and the rest are background, Text region proposals methods giving multiple possible text bounding boxes. Text detection and

recognition in natural scene images is an important part of the field of computer vision. The flow chart of text detection and recognition in natural scenes is shown in theFig. 1. The first step is to get the image containing the characters to be recognized through image information acquisition and analyze its structure. In the second step, imageprocessing methods such as threshold operation are used to denoise and correct the measuredobject. The third step, because of the particularity of text information, requires row and column segmentation to detect a single character or several characters. In the fourth step, thesegmented character image is imported into the recognition model for processing, and the character information in the original image is obtained.

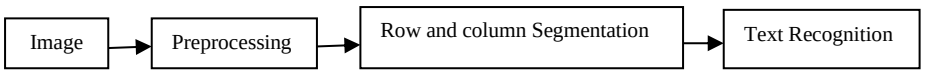

Fig. 1: Flow chart of text detection

The key and trouble of character identification and detection in natural images are character highlight extraction. Albeit a ton of work has been done to characterize a decent arrangement of text highlights, a large portion of the elements utilized in viable applications are not all-inclusive.

In outrageous cases, many elements are practically invalid or even difficult to remove, for example, strokes, shape includes, etc. Then again, characterizing and separating counterfeit elements is a tedious and energy-consuming undertaking. In this manner, text discovery and acknowledgment of complicated natural images have extraordinary strain and difficulties. We attempt to consolidate the deep learning neural network technology with text detection technology and propose a viable text detection method, which gives another down-to-earth technique to tackle the text detection issues in natural scene images. In this paper, we use Res Net network to extricate undeniable level visual highlights from the base pixels in text discovery [1], we want these undeniable level highlights to communicate more valuable substance. The BLSTM layer is utilized to separate the setting elements of character successions [2]. Notwithstanding, the possibility of faster R-CNN vertical anchor is acquainted with observe the limits of distinguished text, which actually works on the impact of text recognition [3]. Furthermore, in-text acknowledgment assignments, Dense Net model is utilized to build text acknowledgment in view of Kares, and the result of Soft max is utilized to order each character. At long last, join with corpus to track down relating characters. The exploratory outcomes show that the strategy accomplishes great outcomes in text detection and recognition of natural scene images. It carries out a productive system for profundity learning [4, 5, 6] which can uphold an assortment of neural organization structures and give a progression of viable preparation methodologies. The system to begin with

approves the adequacy of text discovery and acknowledgment technique in light of profound learning in natural scene images.

2. Related work

Text detection is the method involved with changing over image data into a succession of images that can be addressed and handled by computer. There are many related investigations on text detection in natural scene images [7,8]. Through the investigation of characters, the powerful component data of characters is removed, and the text region in the picture is identified precisely and actually. The errand of text acknowledgment can be viewed as a unique interpretation process: making an interpretation of image signals into regular dialects. This is like discourse acknowledgment and machine interpretation. According to a numerical perspective, they will contain an enormous number of commotion input arrangements, and structure a bunch of given label yield groupings through programmed learning model. FCN (Full convolutional network) is the fundamental network that eliminates the full association (fc) layer [9]. It was initially used to execute semantic division assignments. The upside of FC is that it utilizes up sampling activities like deconvolution and un pooling to reestablish the element grid to the size near the first image, and afterward makes classification expectation for the pixels at every area, in order to perceive the more clear article limit. In the detection network in view of FCN, the article limits are anticipated straightforwardly as per the high goal include map as opposed to getting back to the item limits through candidate regions. FCN is more vigorous in foreseeing unpredictable article limits than Faster-RCNN in light of the fact that it doesn't have to characterize the proportion of candidate box length to width prior to preparing. In light of the great pixel goal of the last layer highlight guide of FCN network, and the need to depend on clear text strokes to recognize various characters (particularly Chinese characters) in the undertaking of text recognition, FCN network is truly appropriate for removing text highlights. At the point when FCN is utilized for text acknowledgment undertakings, every pixel in the last layer of component chart will be isolated into two classes: text line (foreground) and non-text line (background). In EAST (Efficient and Accuracy Scene Text Detection Pipeline) model, the full convolution network (FCN) is utilized to produce multi-scale intertwined highlight guides, and afterward pixel-level text block expectation is performed straightforwardly. In this model, there are two sorts of text region labeling, which are spinning square shape box and self-assertive quadrilateral. The model has better impact in distinguishing English words and less impact in identifying Chinese long text lines. Maybe, as indicated by the attributes of Chinese information designated preparing, the location impact actually has space for development. FTSN (Fused Text Segmentation Networks) model uses division network to help slanted text location [7]. It utilizes Resnet-101 as the fundamental network and utilizations multi-scale fusion feature maps. The

explanation information incorporates the pixel cover and the line of the text occurrence, and the joint preparation of pixel

forecast and line identification is utilized. Fast Oriented Text Spotting is a start to finish learnable network model for coordinated preparation of image text detection and recognition [8]. Detection and recognition undertakings share the convolution feature layer, which saves figuring time, yet in addition learns more image highlights than two-stage training. RoI rotate is acquainted with create arranged text areas from convolutional feature maps, which can uphold the acknowledgment of slanted text. Customary techniques for text detection and recognition and a few strategies for text detection and recognition in light of top to bottom learning are generally multi-stage, which should be enhanced in many stages in the preparation interaction, which will definitely influence the impact of the last model, and is very tedious.

3. Methodology

We initially identify the text and digital domain in the training image, and afterward perceive the recognized text areas. Notwithstanding, the intricacy of preparing information with complex scenes, textual styles in pictures and shooting points might be hard for natural eyes to recognize, aside from computers. This is likewise the trouble of our work. Our detection method depends on deep convolution neural network (CNN) in natural scene images, utilizing Res Net network to get more elevated level elements from the fundamental pixels. We pick Res Net network model to identify text in images, in light of the fact that its location quality is higher than other network models, and it is reasonable for distinguishing harsh and complex image data. The BLSTM layer is utilized to remove the setting highlights of character arrangements, and afterward the possibility of quicker R-CNN vertical anchor is acquainted with recognize text limits, which actually works on the impact of text discovery. Res Net [1] proposed a deep residual network called Res Net, the best commitment of what is to keep away from the issue of angle vanishing or slope blast as the quantity of network layers builds utilizing a residual network. It speeds up the speed of combination, affirms the exactness of the deep neural network, and extends it. An alternate route association is added to each residual unit. According to a utilitarian perspective, the adjustment of personality is expanded. From the forward propagation, the introduction of identity transformation can deliver the change of the network parameters all the more impressive. From the backpropagation, immediate, forward-layer propogation is added to the blunder term to ease the issue of inclination decrease. In this manner, the issue of angle vanishing is addressed. The residual unit recipe can be communicated as follows: $x=R(x)=\sigma(F(x, W))+x$, where x represents the output of the residual unit, x represents its input, F is additionally called the residual function, W is a weight matrix, and σ represents the ReLU activation function. Faster R-CNN as a test

network framework [3,10] it will probably observe the minimal encompassed by recognizing the Bounding Box around the object. It presents RPN (Region Proposal Network) based on Fast RCNN detection framework to rapidly create various candidate Region reference frames that are near the length and width of the objective article. It creates the regional elements of standardized fixed sizes through the Region of Interest Pooling layer for a very long time reference boxes. It utilizes the common CNN to all the while input Feature Maps to the above RPN network and ROI Pooling layer, subsequently decreasing the quantity of convolutional layer boundaries and how much estimation.

3.1 Character Recognition

We use Dense Net network model to build our character recognition model in view of Kares [11]. The fundamental benefit of Dense Net is to improve the dispersal of elements and empower feature reuse. The center thought is to make a cross-layer connection with interface the front and back layers of the network, which is truly reasonable for scene character recognition. We utilize the Soft max classifier as the final output layer [12]. Softmax work depends on Soft max regression. It is a regulated learning calculation.

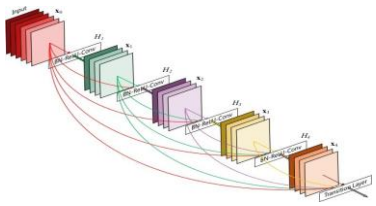

Fig. 2**:** Dense Net model

Dense Net is a convolutional neural network with dense connections. In this network, there is an immediate connection between any two layers, the contribution of each layer of the network is the association of the result of the relative multitude of past layers, and the feature maps learned by this layer will be passed straightforwardly to every one of the layers behind it as input. The construction of a square is as per the following: BN-ReLU-Conv (1*1) BN-ReLU-Conv (3*3), while a Dense Net is made out of a few such squares. The layer between each Dense Net block is called transition layers, which is made out of BN-Conv (1*1) normal Pooling (2*2). One might say that Dense Net has assimilated the most fundamental piece of Res Net, and has accomplished more creative work on it, making the network execution further moved along. Dense connection, mitigating the issue of gradient vanishing, upgrading highlight proliferation, empowering highlight reuse, significantly decreasing how much boundaries.$xl = Hl([x0, x1, ..., xl-1])$,The equation represents the Dense Net module [x0, x1, ..., xl−1]. The output feature map from 0 to l, 1 is concatenation. Concatenation is the consolidation of channels, very

much like Inception. Hl incorporates convolutions of BN, ReLU and kernel.

3.2 Architecture

We first map the detailed area information set through the Res Net network and then insert the closure convolution map into the 3*3 convolution map. These factors might be applied to assume the evaluating class information and region information of K anchors. BLSTM is applied to recursively upload successive windows to every line, wherein the convolution traits of each window (3*3*C) are applied as input for 256-dimensional BLSTM. The BLSTM layer is trailed by the 512 FC layer, which collectively outcomes textual content or non-textual content probabilities, Y-axis coordinates and the parallel refinement offset of the K anchor. We set the threshold of the detected textual content vicinity to 0.8, and anchor values above 0.8 are recognized as textual content, in addition to the opposite manner around. The network shape is Dense Net. At long final, we are able to make use of the organized model to check the untrained information. Through the trained model, textual content information in natural scene images may be prominent and perceived.

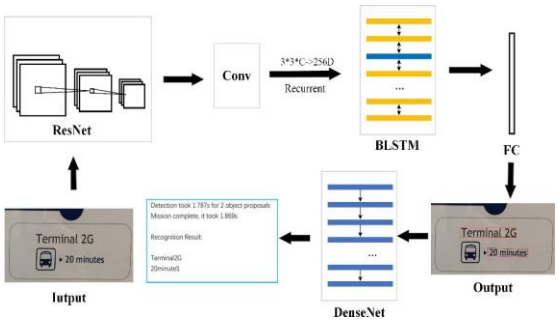

Fig. 3: Structure of text extraction CCN

4. Experiments and Results

We arbitrarily applied all datasets to divide them into five sections. As to follow the 3:1:1 split, we have created a training set, a validation set, and a test set respectively. The training technique is iterated via way of means of 50 epoch, and the training iteration is examined via way of means of 1 epoch. At lengthy last, the accuracy of the training version is tested via way of means of the test set. We will introduce this factor from 3 aspects. First, the results of textual content detection are offered and compared with the existing advanced methods; second, the results of textual content popularity are offered and compared to existing advanced methods; finally, the elements that affect the overall experimental performance are analyzed. To entire this experiment, we used precision, recall, and f-measure as a degree of textual content detection.

4.1 Text Detection

We transformed all the images into a dataset in VOC format, using folder annotations, image sets, and JPEG images. When folder annotation mainly stores XML files, each XML is an image and each XML stores file location and category information of destinations and name marked are usually the same as the corresponding original image, while in image sets we only have to use the main folder, which contains some text files, typically train, test, the contents of this file are the names of the images that need to be trained or tested the JPEG Images folder contains the original images that we have named according to the uniform rules. We put the dataset into the network for training, iteration 50 Epoch, each input 30 images, each iteration is completed to check together and finally save the best model as a detection model. We evaluated our test results on two benchmark data sets ICDAR 2011 [13] and ICDAR 2013[14].

These test data sets are grueling, including different angles, veritably small scales, and low resolution. Table.1 shows the results of our evaluation of two sets of public data. Comparing our work with other existing methods [15,16,17,18,19,20], it is easy to find that our work achieves optimal performance on both data sets. Especially on the ICDAR 2013 data set [14], it has increased from precision of 0.85 to 0.93 over the latest method Text Flow [20] and has made tremendous progress in R/F. In addition, we also test our method in the data set in Fig. 4. The text regions in natural scene images are automatically detected by our model and surrounded by red boxes. It can be found that our approach works perfectly in these challenging situations, some of which are difficult for many previous approaches. It can be seen that our method is very suitable for challenging and complex natural scene image detection. In addition, we also validated the results on the MSRA-TD500 dataset, and the experimental results also showed amazing results. It should be noted that our technique is appropriate for some datasets in natural image text detection. It has a decent speculation capacity.

Table 1: Results on the ICDAR 2011(L) and ICDAR 2013(R)

Method	Precision	Recall	F-score
Huang, Lin, Yang et al. (2013)]	0.82	0.75	0.73
[Yao, Bai and Liu (2014)]	0.82	0.66	0.73
[Yin, Yin and Huang (2013)]	0.86	0.68	0.76
[Zhang, Shen, Yao et al. (2015)]	0.84	0.76	0.80
Text Flow [20]	0.86	0.76	0.81
Our	0.89	0.79	0.83

Method	Precision	Recall	F-score
[Yin, Yin and Huang (2013)]	0.88	0.66	0.76
[Neumann and Matas (2015)]	0.82	0.72	0.77
FAS Text [Buta, Neumann andMatas (2015)]	0.84	0.69	0.77
[Zhang, Shen, Yao et al.(2015)]	0.88	0.74	0.80
Text Flow [20]	0.85	0.76	0.80
Our	0.93	0.81	0.84

4.2 Text Identification

We use Dense Net to train a text identification model. Our training data set is used to iterate over 50 epochs, input 30 pictures (batch size=30) at a time, verify each iteration, and finally save the best model as the training model. We put the image that detected the text box in the model form for the identification and combines it with the corpus to find corresponding characters. Tab. 2 shows our final results. We can see that the identification rate of the Le Net method on the dataset is 0.872 [21)], while that of the Nin Net method is only 0.824 [22]. The accuracy is improved to 0.933 by using the VGG Net method [23].In our method, Dense Net can achieve an accuracy of 0.941, which shows the effectiveness of our method. Figure. 5 shows the result of the MSRA-TD500 dataset. As can be seen from the figure, our method is very effective in text identification of natural images. Not only can English characters be detected, but also multi-characters can be detected. Although there are still some mistakes, there are also a lot of objective factors in it. Such as font

deformations, lighting conditions, and scale and orientation. In addition, there is also a lack of the ability to correctly detect the impact of the text area on the subsequent recognition of character.

Fig. 4: Results on standard datasets

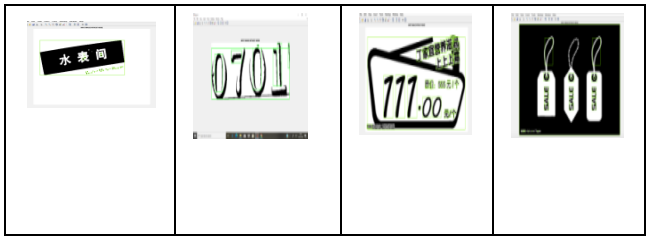

Fig. 5: Results on MSRA-TD500

Table 2: Comparisons of extraction results between existing advanced methods and our proposed methods on standard datasets

Model	Training Sets	Testing Sets	Accuracy
Le Net [21]	254800	10000	0.872
Nin Net [22]	254800	10000	0.824
Vgg Net [23]	254800	10000	0.933
Dense Net (our)	254800	10000	0.941

Discussion

We have proposed text detection and identification for images of natural scenes using deep convolutional neural networks. It expands on CNN by adding the Residual network, and BLSTM layer to our model. At last, utilizing the Dense Network model to construct text identification. Our strategy can beat the leaving

techniques. We investigate the accompanying significant elements which can empower our network to perform better. Res Net showed excellent execution in our proposed strategy. It tackled the issue by preparing the network to become testing with expanding profundity. During the preparation time, the forward spread improved the superior qualities of the results. Back propagation mitigated the issue of gradient reduction, in this manner settling the issue of gradient disappearance. The last test results show that the proposed technique is powerful. BLSTM CNN learns spatial data in the open field. Besides, with the extension of the network, the highlights that CNN learns become increasingly dynamic. For text succession detection, clearly, the theoretical spatial highlights advanced by CNN are required. Furthermore, the arrangement component of text is likewise useful for text detection. For horizontal text lines, every text portion is associated, so a network construction of BLSTM is embraced to make the detection result heartier. We pick Dense Net, whose most noteworthy benefit is to upgrade the dissemination of features and encourage feature reuse. The center thought is to make a cross-layer connection to the interface between the front and back layers of the network. Dense Net connection is dense, which can alleviate the problem of gradient disappearance, enhance feature propagation, encourage feature reuse, greatly reduce the number of parameters, and is truly reasonable for scene character recognition.

5. Conclusion

This paper presents a text extraction method for unstructured images in view of a convolution neural network. The results show that this technique is better than the current strategies inexactness. In spite of the fact that it distinguishes some unacceptable regions in a couple of cases, this is likewise because of genuine reasons, images, and numerous experts are as yet attempting to take care of this issue. Later on, to precisely and really detect text regions from the natural scene and accurately distinguish them, we are focused on working on the Deep Matching Prior Network that can effectively reduce the background interference.

References

[1] He, K., Zhang, X., Ren, S. and Sun, J., 2015. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[2] Tian, Z., Huang, W., He, T., He, P. and Qiao, Y., 2016, October. Detecting text in natural image with connectionist text proposal network. In European conference on computer vision (pp. 56-72). Springer, Cham.

[3] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[4] Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117

[5] Sun, Y., Wang, X. and Tang, X., 2014. Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1891-1898).

[6] Glorot, X., Bordes, A. and Bengio, Y**.,** 2011,January. Domain adaptation for large-scale sentiment classification: a deep learning approach. International Conference on International Conference on Machine Learning, pp. 513-520.

[7] Dai, Y., Huang, Z., Gao, Y., Xu, Y., Chen, K., Guo, J. and Qiu, W., 2018, August. Fused text segmentation networks for multi-oriented scene text detection. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 3604-3609). IEEE.

[8] Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y. and Yan, J., 2018. Fots: Fast oriented text spotting with a unified network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5676-5685).

[9] Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[10] Meng, R., Rice, S.G., Wang, J. and Sun, X., 2018. A fusion steganographic algorithm based on faster R-CNN. Computers, Materials & Continua, 55(1), pp.1-16.

[11] Huang, G., Liu, S., Van der Maaten, L. and Weinberger, K.Q., 2018. Condense net: An efficient densenet using learned group convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2752-2761).

[12] Liu, W., Wen, Y., Yu, Z. and Yang, M., 2016, June. Large-margin softmax loss for convolutional neural networks. In ICML (Vol. 2, No. 3, p. 7).

[13] Minetto, R., Thome, N., Cord, M., Fabrizio, J. and Marcotegui, B., 2010, September. Snooper text: A multiresolution system for text detection in complex visual scenes. In 2010 IEEE international conference on image processing (pp. 3861-3864). IEEE.

[14] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A. and De Las Heras, L.P., 2013, August. ICDAR 2013 robust reading competition. In 2013 12th International Conference on Document Analysis and Recognition (pp. 1484-1493). IEEE.

[15] Huang, W., Lin, Z., Yang, J. and Wang, J., 2013. Text localization in natural images using stroke feature transform and text covariance descriptors. In Proceedings of the IEEE international conference on computer vision (pp. 1241-1248).

[16] Yao, C., Bai, X. and Liu, W., 2014. A unified framework for multi oriented text detection and recognition. IEEE Transactions on Image Processing, 23(11), pp.4737-4749.

[17] Yin, X.C., Yin, X., Huang, K. and Hao, H.W., 2013. Robust text detection in natural scene images. IEEE transactions on pattern analysis and machine intelligence, 36(5), pp.970-983.

[18] Busta, M., Neumann, L. and Matas, J., 2015. Fastext: Efficient unconstrained scene text detector. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1206-1214).

[19] Zhang, Z., Shen, W., Yao, C. and Bai, X., 2015. Symmetry-based text line detection in natural scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2558-2567).

[20] Tian, Z., Huang, W., He, T., He, P. and Qiao, Y., 2016, October. Detecting text in natural image with connectionist text proposal network. In European conference on computer vision (pp. 56-72). Springer, Cham.

[21] Yu, N., Jiao, P. and Zheng, Y., 2015, May. Handwritten digits recognition base on improved LeNet5. In The 27th Chinese Control and Decision Conference (2015 CCDC) (pp. 4871-4875). IEEE.

[22] Lin, M., Chen, Q. and Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.

[23] Wang, L., Guo, S., Huang, W. and Qiao, Y., 2015. Places205-vggnet models for scene recognition. Ar Xiv preprint arXiv:1508.01667.

[24] Giordano, D., Murabito, F., Palazzo, S. and Spampinato, C., 2015. Superpixel-based video object segmentation using perceptual organization and location prior. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4814-4822).

[25] Neumann, L. and Matas, J., 2015, August. Efficient scene text localization and recognition with local character refinement. In 2015 13th international conference on document analysis and recognition (ICDAR) (pp. 746-750). IEEE.

[26] Qin, X., Zhou, Y., He, Z., Wang, Y. and Tang, Z., 2017, November. A faster R-CNN based method for comic characters face detection. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 1074-1080). IEEE.

[27] Sebe, N., Gevers, T., Dijkstra, S. and van de Weije, J., 2006, June. Evaluation of intensity and color corner detectors for affine invariant salient regions. In 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06) (pp. 18-18). IEEE.

[28] Tian, S., Pan, Y., Huang, C., Lu, S., Yu, K. and Tan, C.L., 2015. Text flow: A unified text detection system in natural scene images. In Proceedings of the IEEE international conference on computer vision (pp. 4651-4659).

[29] Vondrick, C., Khosla, A., Malisiewicz, T. and Torralba, A., 2012. Inverting and visualizing features for object detection. arXiv preprint arXiv:1212.2278.

[30] Yang, M.H., 2002, May. Kernal Eigenfaces vs. Kernal Fisher faces: Face Recognition Using Kernal Methods, Automatrix Face and Gesture Recognition, 202. In Proceedings, Fourth IEEE International Conference on (pp. 208-213).