IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, March 2024

Content-Based Text Extraction from Image using Deep Learning

[1]Brijen Rajak and [2] Raghavendra R
PG Student, Department of MSc CS-IT[1]
Assistant Professor, School of CS & IT[2]
Jain (Deemed-to-be University), Bangalore, India
[1]brijenrajak@gmail.com and [2]r.raghavendra@jainuniversity.ac.in

Abstract: *Text extraction proves beneficial in various scenarios, as it allows the conversion of information stored in non-text formats, like images or documents, into machine-readable and searchable text. In contemporary times, this technique serves as a time-efficient tool across different sectors such as real estate, finance, law, food ordering and delivery, and e-commerce. Industries are increasingly adopting text extraction methods. Previously, numerous models centered around text extraction utilized OCR, CNN, and RNN. When it comes to extracting text from images using content-based approaches, CNNs play a crucial role in recognizing and locating text regions within the images. In situations where the identification and transcription of text from images are essential, RNNs prove valuable for content-based text extraction. While CNNs and RNNs independently yield accurate outcomes in content-driven text extraction from photos, the combined utilization of both methods surpasses the individual effectiveness of each. The proposed CRNN system stands out in various aspects compared to existing methodologies. It not only demonstrates heightened accuracy and efficiency but also exhibits superior performance overall. Our investigation's findings highlight that the CRNN methodology, when applied, outperforms previous approaches by recognizing text in images with a reduced latency and more precise recognition*

Keywords: CRNN, CNN, RNN, Content-based, Text Extraction, Images.

I. INTRODUCTION

In recent years, there has been a proliferation of digital photographs containing text, underscoring the necessity for efficient techniques in extracting and comprehending this textual content [10]. The contemporary field of context-based text mining from images, driven by deep learning, has emerged to automatically recognize and transcribe textual material discovered in visual sources [9]. Deep learning, a subfield of AI that considers neural network architectures, has revolutionized text extraction from images, employing neural networks such as RNNs, CNNs, and their variants to significantly enhance accuracy and efficiency in identifying and interpreting text across various visual contexts [11]. RNNs play a crucial role in deciphering the ordered structure of textual information embedded within images, and their combination with feature extraction techniques, like CNNs (Convolutional Neural Networks), in models such as CRNNs (Convolutional Recurrent Neural Networks), substantially amplifies the precision of text recognition in the context of text extraction from images [12]. The extraction method involves multiple phases, including word region identification, text recognition, and, in some instances, determining the layout and structure of the text within the image [13].

Irrespective of text length, orientation, or font variations, CNNs and RNNs exhibit proficiency in detecting and transcribing text sequences [18]. A CNN is crucial for feature extraction from images and text detection, while an RNN is essential for capturing temporal associations in sequential data [14]. Combining both CNN and RNN yields a highly efficient and accurate solution for content-based extraction of words from images, denoted by the acronym CRNN, which signifies the integration of CNN and RNN [15].

The utilization of a Convolutional Recurrent Neural Network (CRNN) for extracting content-based text from photos offers numerous advantages, including heightened accuracy in text acknowledgment, adaptability to varying text

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, March 2024

properties, reduced error rates, and more [16]. This method proves valuable for achieving complete extraction of text from image files, with the model's performance enhanced through training on an extensive collection of photos [17].

II. LITERATURE SURVEY

S. Ghaleb et al. proposed an approach emphasizing the influential role of images in conveying a brand, service, or item, as images possess the ability to add depth, context, and a narrative to the overall experience, surpassing the impact of words alone [1]. Image retrieval, particularly in extensive databases, stands out as the most efficient method of searching. Content-based search proves to be an effective instrument for various sectors, recognizing the complexity of images compared to texts. Each image incorporates crucial details such as color, texture, and borders, leading to the concept of content-based picture retrieval. Leveraging Convolutional Neural Networks (CNNs), the authors highlight the capability to extract characteristics in photographs and classify them with remarkable accuracy. Their proposed intelligent CNN-based model aims for high-performance picture retrieval [1].

Devareddi R.B. et al. contribute to the field of image retrieval by drawing inspiration from recent advancements in multimedia resources and the proliferation of online image archives [2]. The challenge of locating a specific image within a large collection is a common concern for professionals across various disciplines, including journalists, designers, and art historians. Addressing this issue, scholars have devised novel techniques for representing and indexing visual data. One such innovative method is Content-Based Image Retrieval (CBIR), which relies on inherent characteristics within the images [2]. In a "content-based" search, the focus is on examining the actual content of the image. The characteristics of picture content considered in CBIR include color, texture, shape, and spatial relationships [2]. This approach proves to be a significant advancement in effectively retrieving and organizing visual data, catering to the needs of professionals dealing with diverse design requirements.

S. K. J. et al. highlight the substantial growth in the field of Content-Based Image Retrieval over a span of twenty years, where the focus is on finding suitable images through image and video perception [3]. The proliferation of the Internet and the abundance of graphical and multimedia data necessitate effective field matching. In contrast to text search-based approaches for visual retrieval, which may face inconsistencies between visual content and entire texts, the study addresses the need for field matching due to the availability of vast multimedia resources [3]. In response to these challenges, a diverse range of programs and tools has been developed to assist users in creating, implementing, and researching audio visual content. The study incorporates essential elements from various image and framework retrieval techniques, ranging from simple recognition of features in images to advanced deep-learning approaches [3]. This comprehensive approach underscores the evolution of methods in the realm of Content-Based Image Retrieval.

Y. Zhenyu et al. highlight the advancing impact of deep learning on image tagging, recognizing the ongoing development in this domain [4]. However, they point out a limitation in the traditional feature extraction method, where local characteristics and the connections between local and global features are often overlooked in favor of the overall picture. Additionally, the authors note that text explanations generated by this method lack focus and are overly general. To address these issues, they propose a novel approach grounded in mixed picture features. In this innovative method, an enhanced ResNet is employed to extract global characteristics, while a deep RetinaNet is utilized to extract local characteristics. An attention mechanism is introduced to seamlessly combine the global and local properties of the image, translating them into embedding vectors. Furthermore, an LSTM (long short-term memory) is incorporated to establish correspondence among images and descriptions. The syntax generation model, based on an attention process, is employed to generate content about the imagery by integrating semantic and visual elements [4]. This approach aims to overcome the limitations of the traditional feature extraction method and enhance the specificity of text explanations in image tagging.

G. Sairam et al. delve into the approach of generating captions for images through the utilization of deep neural networks [5]. The method involves feeding the model an image, and the resulting output is presented in three distinct formats: an mp3 audio file, an image file, and an expression that combines the image in each of the three languages. This comprehensive approach integrates techniques from both natural language processing and computer vision. The primary goal is to establish a model capable of generating captions by merging the strategies of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). Convolutional neural networks play a pivotal role in comparing the target picture with the initial training images present in a sizable dataset [5]. This multi-modal approach

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, March 2024

aims to enhance the caption generation process through the synergy of natural language processing and computer vision techniques.

K. Wangi et al. address the challenge posed by massive datasets that require more time for searching and retrieving photos using traditional retrieval techniques such as text and query-based image retrieval [6]. In response, the study employs deep learning neural network technology, specifically the Autoencoder, to expedite the search process and enhance results. Autoencoders, a type of neural network, specialize in reconstructing images. The proposed work incorporates the technique known as "Information Base Image Extraction" to retrieve similar visual information, textures, colors, and shapes from images. The autoencoder technique is a focal point of the study, utilized to convert images into a feature-encoded vector format. This format proves instrumental in searching for images within larger datasets, significantly reducing the time required for retrieval [6]. The implementation of autoencoders underscores their effectiveness in accelerating the image search process in the context of massive datasets.

M. Sheppard et al. [7] focus on "Text Extraction through Image Processing" in the Journal of Advanced Research in Computer Science. Optical Character Recognition (OCR) is explored as a computer system designed for converting scanned images of handwritten text into machine-editable content or a universal character encoding scheme. The roots of OCR trace back to the research conducted in computer vision and artificial intelligence. In practical applications, such as post offices, banks, universities, and various formal tasks requiring extensive data input, text recognition becomes crucial. It serves as a means to extract knowledge from images containing handwritten or printed text. This intersection of image processing and text extraction represents a significant area of study with diverse real-world applications [7].

Dalal N. et al. [8] highlight the prevalence of digital electronic devices producing diverse data types, including text, images, and videos, in modern times. Consequently, the need for an efficient system for archiving and collecting data becomes imperative. The specific challenge addressed is the retrieval of images within a substantial image collection, termed as content-based image retrieval (CBIR). CBIR involves employing comparable image identification and feature extraction techniques to overcome the challenge of locating images within a large database. Previous research in this domain has leveraged controlled deep neural networks for feature learning, coupled with a comprehensive search strategy to locate related photos [8]. This approach underscores the significance of integrating advanced technologies to address the complexities associated with managing and retrieving diverse digital data.

III. PROPOSED MODEL

Mathematical Notations:

CNN:

Convolution operation:

Convolutional Neural Networks (CNNs) employ convolution and pooling processes for text extraction from images, with fully connected layers added at the end for classification or regression tasks. The fundamental mathematical equations are outlined as follows:

Convolutional Operation:

For the given input Feature map F and a Filter K, the convolutional operation is;

$C(m,n) = (m * n) + b$ (1)

m– Represents the input **feature map**

n – Represents the **filter** or **kernel**

b – Represents**bias** term

*** -** Denotes **operation** of the convolution.

ReLU Activation:

ReLU stands for Rectified Linear Unit. It is the next step of the convolution operation. It is defined as;

$f(m) = \max(0, m)$ (2)

The function *f(m)* instantiates non-linearity by changing all negative pixel values with zeroes.

Pooling Operation:

It is the process of taking maximum value in a sliding window. Represented as;

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, March 2024

$P(m) = \max(m)$ (3)

P(m) – reduces the spatial dimensions and reduces computational complexity.

Fully Connected Layers:

For classification or regression, the equations involve matrix multiplication with weights and addition of biases. It is represented as;

$FC = XW + b$ (4)

X – flattened input

W – weights

b - bias term

FC - output of fully connected layer

RNN:

Recurrent Operation:

Forget Gate (f_t)

$g_t = \sigma(w_f \cdot [H_{t-1}, m_t] + B_f)$(5)

Input Gate (I_t)

$I_t = \sigma(w_f \cdot [H_{t-1}, m_t] + B_i)$ (6)

Candidate Memory ($\widetilde{c_t}$)

$\tilde{c}_t = \tanh(w_C \cdot [H_{t-1}, m_t] + B_C$ (7)

Cell State Update (c_t)

$c_t = g_t \cdot c_{t-1} + I_t \cdot \tilde{c}_t$ (8)

Output Gate (O_t)

$O_t = \sigma(w_o \cdot [H_{t-1}, m_t] + B_o)$(9)

Hidden State (H_t)

$H_t = O_t \cdot \tanh(c_t)$ (10)

Extracting text:

In a larger sense, CRNN employs a sequence-to-sequence modeling methodology, which is mathematically expressed as:

Sequence-to-Sequence: $Y = f(X)$ (11)

X – Represents the input image

Y - Represents the text sequence extracted from the image using function *f*

Algorithm:

Step 1: Data Preparation:

Compile a dataset of images with corresponding captions and additional information. Ensure consistency through preprocessing steps, including resizing, normalizing, and grayscale conversion.

Step 2: Prepare the Dataset:

Segment historical data into validation and training sets. Create data loaders for efficient handling of images and annotations.

Step 3: Model Construction:

Detail the components of the CRNN architecture, specifically the CNN and RNN parts. Configure the CNN to extract features from images and utilize RNN (such as LSTM) for sequence modeling.

Step 4: Model Training:

Set up the CRNN model, selecting a loss function (e.g., CTC loss) and an optimizer (e.g., Adam). Enable the model to learn:

Utilize CNN to extract features from images.

Employ RNN for predicting character sequences and processing attributes.

Update the model's weights through backpropagation and optimization.

Step 5: Validation and Fine-tuning:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, March 2024

Verify model accuracy using a distinct validation dataset. Adjust model parameters based on validation results.

Step 6: Text Extraction:

Train the model with new images:

Prepare images to align with the model's input format.

Feed preprocessed image data to the trained CRNN.

Extract and decode anticipated character sequences.

Step 7: Post-processing:

Enhance text extraction by removing unnecessary characters or white spaces. Make language-based or spell-checking corrections as needed.

Step 8: Evaluation and Refinement:

Assess accuracy and performance by comparing the extracted text with actual data. Optimize the CRNN model based on evaluation results for enhanced accuracy.

Step 9: Continuous Improvement:

Iterate through steps 4 to 8, adjusting the architecture, regularization algorithms, or settings as necessary to achieve improved performance.

Step 10: Application and Deployment:

Deploy the enhanced CRNN model for real-world applications, such as Optical Character Recognition (OCR) and document comprehension.

Architecture:

The image processing workflow begins with feeding images into the model, progressing through two distinct stages:

Convolutional Neural Networks (CNNs) in the Initial Stage:

CNNs play a pivotal role in extracting high-level visual information from the input images.Convolutional layers within the CNN are employed to recognize various visual aspects, including edges, shapes, and textures, capturing these intrinsic properties.

Recurrent Neural Network (RNN) Activation in the Next Stage:

A specially designed RNN is activated to process the sequence of features obtained from the CNN.The RNN's role is to decode the sequence of visual data into a sequence of characters or tokens, representing the textual information.Typically, the RNN comprises layers with Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) layers, facilitating the effective processing of sequential information.

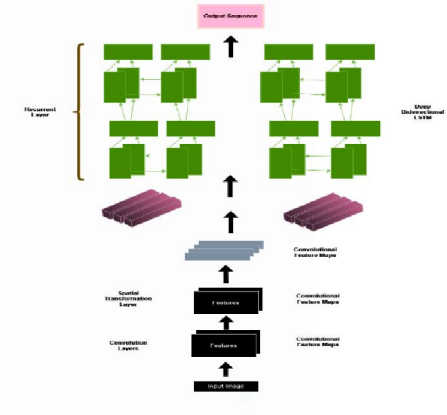

Fig-1: Proposed Architecture

IV. EXPERIMENTAL ANALYSIS

Dataset

This project relies on a substantial dataset generously provided by the Visual Geometry Group. The dataset is extensive, totaling 10 gigabytes of photos. To train the model, a training set comprising 135,000 photos and a validation dataset

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, March 2024

consisting of 15,000 images were utilized. Two approaches are available for acquiring this dataset: either use the provided link containing the dataset or employ built-in tools in the terminal for a seamless download.

Results and Discussions:

Two pre-existing models were taken into consideration and juxtaposed with the proposed system. The comparison of results is illustrated below through graphs, showcasing an analysis of the outcomes. Various evaluation metrics, including Accuracy, Precision, Loss, and Recall, were employed to assess and compare the performance between the proposed system and the existing systems.

Accuracy:

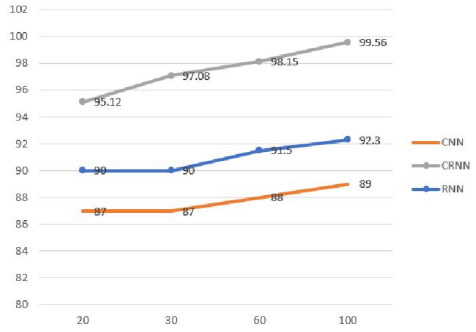

Fig-2: Epochs vs Accuracy

As the graph above illustrates, CRNN out performs the other two current systems, CNN and RNN, with a 99.56 % accuracy rate. As the graph above illustrates, CRNN outperforms the other two current systems, CNN and RNN, with a 99.56 % accuracy rate.

Loss Score:

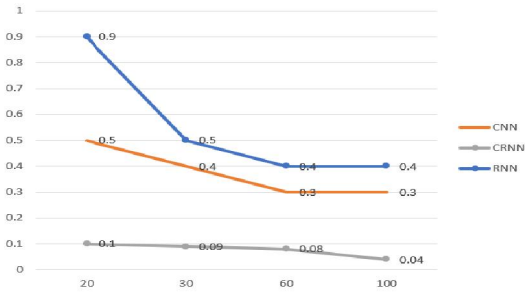

Fig-3: Epochs vs Loss score

It is evident from the results of the above graph that the suggested model outperforms the two models, CNN and RNN, in general.

Precision:

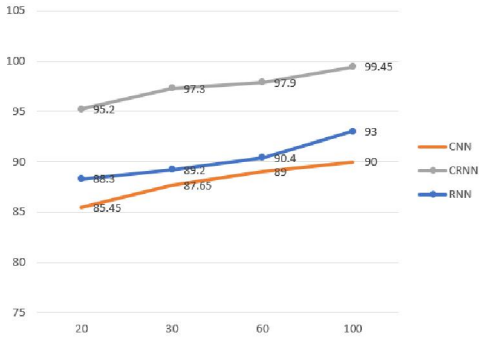

Fig-4: Epochs vs Precision

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, March 2024

The findings of the preceding graph show that, when precision between the proposed model and current models is compared, the suggested model performs better overall than both of the models, CNN and RNN.

Recall:

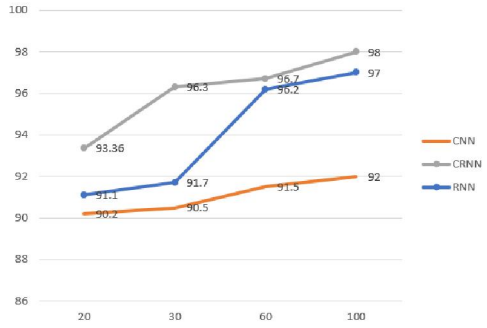

Fig-5: Epochs vs Recall

Recall metrics for the current and Existing systems are compared, the proposed system, CRNN, has obtained 98 recall, while CNN and RNN gained 92 and 97 recall, respectively.

F1-score:

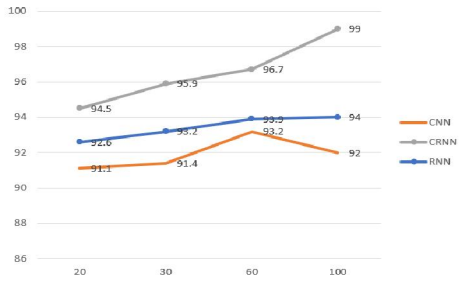

Fig-6: Epochs vs F1-Score

Comparing the F1-score values of the proposed and existing systems, proposed system CRNNhas acquired 99F1-scorewhereas CNN and RNN acquired 92 and 94.

V. CONCLUSION

In this study, the CRNN model was employed as the state-of-the-art technique for extracting text from photographs based on their visual features. The primary objective was to enhance content-based text extraction performance through the utilization of deep learning techniques. Rigorous testing and analysis were conducted to comprehensively compare the CRNN technique with two well-established models, namely CNN and RNN. The CRNN model exhibited distinct advantages over standalone CNN and RNN designs, as evident from the comparative research. Throughout the evaluation process using various metrics such as accuracy, recall, precision, loss score, and F1-score, the CRNN method consistently outperformed the other models. Its superior performance across these domains underscores its effectiveness in handling the complexities of text extraction tasks, particularly when faced with variations in sentence orientations, widths, and font types within images. The CRNN method demonstrated exceptional proficiency in Optical Character Recognition (OCR) by integrating sequential and spatial processing, thereby enhancing the accuracy of text identification, segmentation, and transcription.

REFERENCES

[1]. "Content-based Image Retrieval based on Convolutional Neural Networks," 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2021, pp. 149-153,doi: 10.1109/ICICIS52592.2021.9694146, M. S. Ghaleb, H. M. Ebied, H. A. Shedeed, and M. F. Tolba.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, March 2024

[2]. Devareddi R. B. and Srikrishna, A., "Review on Content-based Image Retrieval Models for Efficient Feature Extraction for Data Analysis," in 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, pp. 969-980, doi: 10.1109/ICEARS53579.2022.9752281.

[3]. "A Review on Content Based Image Retrieval Techniques," by S. K. J. and M. C. V. S., in 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, pp. 12511256;doi110.1109/ICCPCT58313.2023.10245360

[4]. Y. Zhenyu and Z. Jiao, "Research on Image Caption Method Based on Mixed Image Features," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019, pp. 1572-1576, doi: 10.1109/IAEAC47372.2019.8998010.

[5]. G. Sairam, M. Mandha, P. Prashanth, and P. Swetha, "Image Captioning using CNN and LSTM," in Bahrain, 2021, online conference, 4th Smart Cities Symposium (SCS 2021), pp. 274-277, doi: 10.1049/icp.2022.0356.

[6]. K. Wangi and A. Makandar, "Autoencoder for Image Retrieval System using Deep Learning Technique with Tensorflow and Kears," in IEEE ICICACS 2023 (Raichur, India), Proceedings, 1–5, doi: 10.1109/ICICACS57338.2023.10099675.

[7]. M Sheppard and Hinton G E 2011 ESANNEnd-to-end scene text recognition2

[8]. P. Balasundaram, S. Muralidharan and S. Bijoy, "An Improved Content Based Image Retrieval System using Unsupervised Deep Neural Network and Locality Sensitive Hashing," 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 2021, pp. 1-7, doi: 10.1109/ICCCSP52374.2021.9465496.

[9]. Theory and applications of scale invariant feature transform on the sphere, Cruz-Mota J., Bogdanova I., Paquier B., Bierlaire M., Thiran J. (2012) Int. J. Comput. Vis. 98:217–241. 10.1007/s11263-011-0505-4 is the doi.

[10]. K.N. Natei Journal of Engineering Research and Application ISSN : 2248-9622, Vol. 8, Issue5 (Part -V) May 2018, pp 27-33 USA from June 20–25, 2005,

[11]. Zhang, Xiangnan, Xinbo Gao, and Chunna Tian. "Text detection in natural scene images based on colour prior guided MSER." Neurocomputing 307 (2018):

[12]. "Feature pyramid networks for object detection," T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and H. Belongie, 2016, https://arxiv.org/abs/1612.03144.

[13]. "A faster RCNN-based pedestrian detection system," X. Zhao, W. Li, Y. Zhang, T. A. Gulliver, S. Chang, and Z. Feng, Proceedings of the IEEE 84th Vehicular Technology Conference (VTC-Fall), IEEE, Montreal, Canada, 18–September 2016.

[14]. "Arbitrary-oriented scene text detection via rotation proposals," by J. Ma, W. Shao, H. Ye, and others, IEEE Transactions on Multimedia, vol. 20, pp. 3111–3122, 2017.

[15]. "Detecting text in natural image with connectionist text proposal network," Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, Proceedings of the 14th European Conference on Computer Vision, pp. 56–72, Springer, [16]Cham, Switzerland, October 2016.

[16]. End-to-end scene text recognition by Wang, Babenko, and Belongie (2011); of 2011 International Conference on Computer Vision, Barcelona, Spain, November 6–13, 2011; pp. 1457–1464.

[17]. Theory and applications of scale invariant feature transform on the sphere, Cruz-Mota J., Bogdanova I., Paquier B., Bierlaire M., Thiran J. (2012) Int. J. Comput. Vis. 98:217–241. 10.1007/s11263-011-0505-4 is the doi.

[18]. SIn the Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), held in San Diego, California, USA from June 20–25, 2005, Dalal N