Received: 23 November 2022 | Revised: 28 March 2023 | Accepted: 4 April 2023 | Published online: 6 April 2023

RESEARCH ARTICLE

Artificial Intelligence and Applications
2023, Vol. 1(3) 129–138
DOI: 10.47852/bonviewAIA3202549

BON VIEW PUBLISHING

Adapting a Swin Transformer for License Plate Number and Text Detection in Drone Images

Srikanta Pal[1], Ayush Roy[2], Palaiahnakote Shivakumara[3,*] and Umapada Pal[2]

[1]*Maynooth International Engineering College, Maynooth University, Ireland*

[2]*Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, India*

[3]*Faculty of Computer Science and Information Technology, University of Malaya, Malaysia*

Abstract: The use of drones and unmanned aerial vehicles has significantly increased in various real-world applications such as monitoring illegal car parking, tracing vehicles, controlling traffic jams, and chasing vehicles. However, accurate detection of license plate numbers in drone images becomes complex and challenging due to variations in height distances and oblique angles during image capturing, unlike most existing methods that focus on normal images for text/license plate number detection. To address this issue, this work proposes a new model for license plate number detection in drone images using Swin transformer. The Swin transformer is chosen due to its special properties such as higher accuracy, efficiency, and fewer computations, making it suitable for license plate number/text detection in drone images. To further improve the performance of the proposed model under adverse conditions such as degradations, poor quality, and occlusion, the proposed work incorporates a maximally stable extremal region-based regional proposal network to represent text data in the images. Experimental results on both normal license plates and drone images demonstrate the superior performance of the proposed model over state-of-the-art methods.

Keywords: MSER, deep learning, Swin transformer, text detection, license plate number detection

1. Introduction

Text and license plate number detection is important for several real-world surveillance applications, where text detection facilitates text recognition to understand images and videos. Some examples of real-world cases include automatic driving without a pilot, machine translation, human–computer interactions, etc. In these applications, there are some challenges like arbitrary orientation, arbitrarily shaped text, low resolution, complex background, font variations, etc. for achieving better detection results (Mittal et al., 2022; Nadanwar et al., 2022). However, most of these challenges are addressed adequately by the existing methods using different deep learning-based approaches. But in the case of surveillance applications, drones have been used for monitoring and tracking vehicles, traffic jams, illegal parking, toll fee collection, etc. In these situations, due to variations in the heights and oblique angles of drone cameras, captured images suffer from severe degradation, poor quality, occlusion, inadequate information, etc. It is visible in Figure 1(a), where partial license plate number is visible, quality differs from one license plate number to another due to distance variations between the camera and cars and the effectiveness of perspective distortion due to oblique angle.

In contrast to drone images in Figure 1(a), normal images shown in Figure 1(b) do not suffer much from degradation. Since these challenges are different from normal scene images, the past methods may not be effective for drone images. It is evident from the results of the state-of-the-art methods (Liao et al., 2022; Zhang et al., 2020; Zhu et al., 2021) and the proposed method on drone and normal scene images shown in Figure 1(a) and (b), respectively. The methods (Liao et al., 2022; Zhang et al., 2020; Zhu et al., 2021) used a deep learning-based approach for addressing challenges of scene text detection, misses characters in the case of drone images. On the other hand, the same methods works well for normal scene text images. As a result, one can infer that the existing methods are not effective for drone images. At the same time, the results of the proposed method shown in Figure 1(a) and (b) show that the proposed method is capable of handling both drone and normal scene images. Therefore, there is a need for addressing the above challenges to achieve better results for drone images.

Previous studies have attempted to address the challenges of drone images. For instance, Kim et al. (2022) proposed a method for rescuing missing people by designing a web server that receives drone images and uses visual content to detect missing people. Mohite et al. (2022) developed hyperspectral imaging techniques to detect crop water stress from images captured by drones using visual spectral analysis. Chowdhury et al. (2022) explored gradient vector flow to detect dominant points in palm tree images captured by drones to detect crown-shaped regions and count the number of

*Corresponding author: Palaiahnakote Shivkumara, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. Email: shiva.um@edu.my

Figure 1
Challenges faced during license plate number detection in drone images

palm trees in drone images. Dwivedi et al. (2022) developed a model for estimating crop area and extraction in images captured by drones based on object-based image analysis instead of pixel-based image analysis to monitor agriculture. However, the scope of these methods is limited to general images and may not be effective for text detection in drone images, including license plate numbers. Therefore, this work aims to develop a new method for detecting license plate numbers and text in both drone and normal images.

The remarkable success of deep learning discussed in the past (Zhang et al., 2022a, 2022b) for the classification of objects and complex scene images motivated us to explore transformer architectures for license plate number detection in drone images. This is because transformer (Liu et al., 2021) has the ability to cope with the challenges posed by multiple adverse factors, and they perform better than conventional deep learning approaches; hence, we explore addressing the challenges of drone images as well as normal scene images in this work. To reduce the effect of nonuniform quality of license plate numbers due to the presence of multiple vehicles in the same image, the proposed work adapts maximally stable extremal regions (MSERs) (Gómez & Karatzas, 2014) based regional proposal network (RPN) to detect text candidates in the input images. This step helps the Swin transformer to perform better detection irrespective of drone and normal scene images.

The main contributions are as follows: (i) exploring Swin transformer for addressing challenges of both drone and normal scene images is new compared to the state-of-the-art methods and (ii) the use of the combination of MSER and RPN for reducing the effect of background complexity and the effect of nonuniform quality is new compared to the existing methods.

This paper is structured as follows: Section 2 provides a succinct overview of related works, while Section 3 delves into the proposed method. The experimental findings are presented in Section 4, and the paper concludes with Section 5.

2. Literature Review

Broadly speaking, the methods for detecting text in scene images can be classified into two categories: those designed for scene text detection and those developed specifically for detecting license plate numbers. Therefore, we review the same in this section.

Zhu et al. (2021) employed the Fourier contour embedding technique to identify text in scene images. The approach aims to create an effective text representation that can handle diverse geometric variations, which is achieved by leveraging the Fourier domain instead of spatial information. However, the method may not perform optimally in detecting text with arbitrary shapes. Similarly, Zhang et al. (2020) proposed a text detection approach for scene images using a deep relational reasoning graph (DRRG) network that employs a graph convolutional network.

Most text detection methods typically use segmentation as a strategy to improve their performance, but their effectiveness is reliant on complex postprocessing procedures. To mitigate the impact of these complicated postprocessing steps, Liao et al. (2022) developed a novel approach for text detection in scene images that leverages differential binarization and an adaptive scale fusion technique. This method aims to overcome the limitations of segmentation-based methods by integrating binarization and segmentation steps to achieve more efficient text detection. Since the scope of the method is limited to scene text images, the method may not be extended for document layout analysis. To address this challenge, Long et al. (2022) tackled this issue by creating an end-to-end unified model that addresses both scene text detection and document layout analysis challenges. Since scene text detection is a component of document image analysis, it is reasonable to expect that a text detection method that performs well in scenes would also excel in document layout analysis. The proposed approach can simultaneously detect scene text and group text into clusters.

Although various methods have been proposed for text detection in scene images, they are often sensitive to noise and low-contrast images. To address this issue, Soni et al. (2022) introduced a supervised attention network that learns multiscale edge semantics and pixel-wise spatial structure information to detect text masks in edge-fainted noisy scene images. However, many existing methods prioritize accuracy over efficiency. To achieve both accuracy and efficiency, Wang et al. (2022) developed an end-to-end approach for spotting arbitrarily shaped text in scene images using kernels that describe the text shape and distinguish it from adjacent text. While most methods require a large number of training samples, Dai et al. (2021) proposed a scale-aware data augmentation-based technique that generates synthetic samples, reducing the dependency on real samples for accurate scene text detection. Nonetheless, these methods may not perform well on images containing deformed text. To address this challenge, Ma et al. (2022) proposed a text attention network that obtains super-resolution images, significantly improving text detection performance, especially for low-contrast and spatially deformed text in scene images.

While most text detection methods use training and testing data from the same distribution to achieve optimal results, this is not always feasible for real-world applications. To address this issue, Zheng (2022) proposed a scene text detection method using cross-domain data and a domain adaptation strategy that involves both low-level and high-level alignment models for feature extraction. Additionally, transformer-based methods have been introduced to reduce computational complexity, improve text detection performance, and reduce the reliance on the number of training samples. For instance, Zeng & Song (2022) developed a Swin transformer with a feature pyramid network for scene text detection in circuit cabinet wiring images. The proposed approach leverages global self-attention context at each level of feature pyramid networks and integrates features from all levels to effectively detect text in the images.

To summarize, while the existing methods have effectively addressed many challenges of scene text detection, they have not been specifically designed to detect text in drone images. Drone-captured images present unique challenges, such as occlusion, distortion, degradations, nonuniform illumination, and multiple text instances in the same image (such as license plate numbers of multiple vehicles in the same image), which may limit the effectiveness of the discussed methods. Furthermore, these methods are primarily focused on scene text images and may not be well-suited for detecting license plate numbers.

Several methods have been developed recently for detecting license plate numbers in different situations. For instance, Bagi et al. (2021) proposed a method for multilingual-oriented scene text and traffic sign detection in adverse meteorological conditions. However, this approach does not primarily focus on license plate number detection. To improve the performance of license plate detection in adverse conditions, Lee et al. (2022) developed an information maximization-based method that uses scene text detectors for detecting license plate numbers. Srilekha et al. (2022) developed a method for license plate number detection and nonhelmet rider identification using a combination of Yolov2 and an optical character recognizer. Gizatullin et al. (2022) used an image weight model for license plate number detection that involves multiple-scale wavelet transforms and morphological gradient for improving performance. Kim et al. (2021) developed a deep learning-based model for recognizing license plate numbers in CCTV images, which includes a super-resolution technique using a generative adversarial network. However, these methods do not address the specific challenges of license plate number detection in drone images. These challenges include occlusion, distortion, degradations, nonuniform illumination, and multiple text instances in the same image.

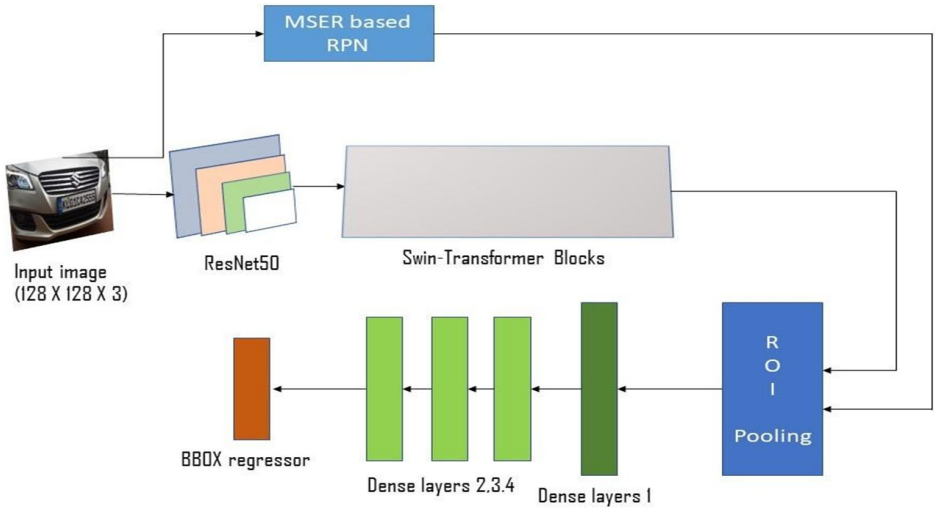
To summarize, the existing methods for license plate number detection have not addressed the challenge of detecting license plate numbers in drone images. However, Jain et al. (2022) proposed a method that adapted the Yolov5 architecture for number plate detection in drone applications, but it is not effective for images with cluttered backgrounds and multiple adverse factors. There is a need for a more robust model that can handle the challenges of drone images with high accuracy and minimal computations. Thus, the proposed work introduces the Swin transformer for detecting license plate numbers and text in scene images.

3. Proposed Method

The aim of this study is to develop a method for detecting license plate numbers and text in drone images. Unlike existing methods that only focus on text detection in normal scene images, the proposed method considers both license plate numbers and text images for detection. However, detecting text in drone images is challenging due to degradations, occlusion, and distortion caused by varying height distances and oblique angles. To address this, the proposed method uses the Swin transformer for license plate number detection, inspired by its ability to extract context and semantic features with high accuracy and fewer computations.

To deal with the complex background of drone images, the proposed method uses a combination of MSERs and RPN to extract text components. The extracted features from the Swin

Figure 2
The block diagram of the proposed model



transformer and MSER-based RPN are fused for license plate number detection. The pipeline of the proposed method can be seen in Figure 2. In the figure, the input image is fed to the backbone network (ResNet50) followed by Swin transformer layers. The feature maps of the backbone are then used in the Swin transformers. The anchor-free RPN comprises MSER-based text region detections that are projected on the feature map of the Swin transformer blocks. Region of Interest (ROI) pooling is performed to maintain a fixed feature size for input to the dense layers. The dense layers include 64 units in dense layer 1, and 32 units each in dense layers 2 and 3, while dense layer 4 has 16 units, all activated using the RELU activation function. The BBOX regressor has four units, which are the four coordinates of the bounding boxes and are activated using the linear activation function.

3.1. MSER-based RPN for text component detection

As explained in the previous section, we propose a novel approach that combines MSER-based RPN to identify text candidate components in the images. For a given input image, the proposed work employs MSER, which outputs candidate components as shown in Figure 3, whereas for the input image shown in (a), the MSER outputs candidate components by discarding nontext components as shown in Figure 3(b). Since MSER is sensitive to background components, it detects some of the nontext components as candidate components. Therefore, the proposed work obtains Canny edge image for the input image as shown in Figure 3(c), where edges are representing prominent

Figure 3
Illustrating the steps for text component detection

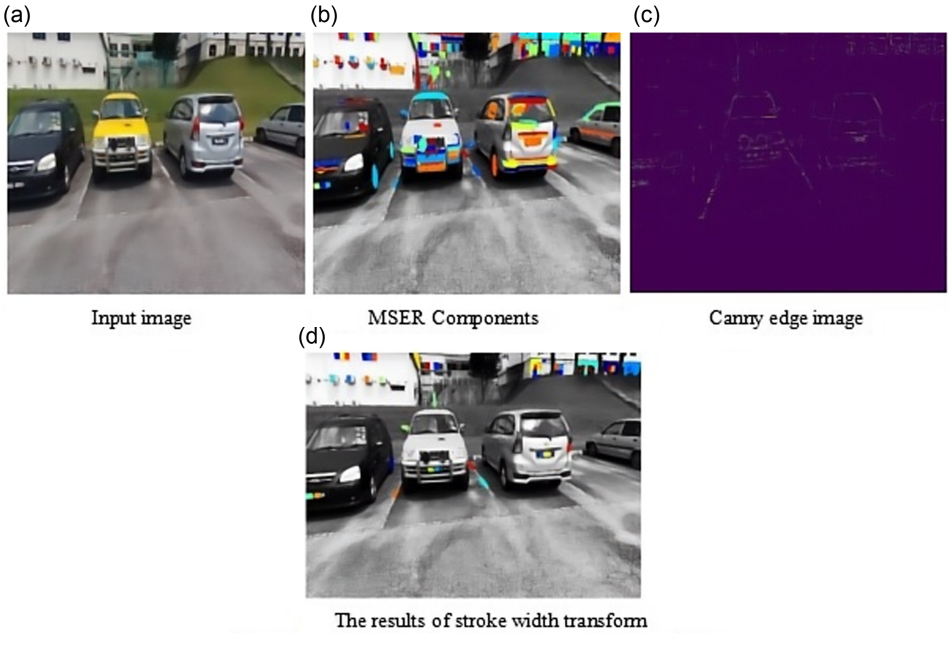
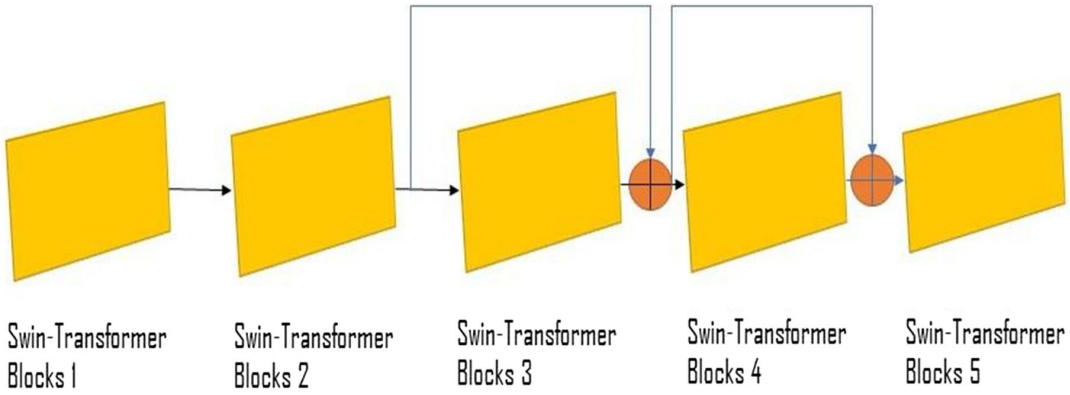


Figure 4
Swin transformer network architecture for license plate number and text detection in drone images



information (edges of text). The RPN (Chen et al., 2017) is used to fuse the output of the Canny edge image and MSER making it anchor free. It uses merged information from the Canny edge detector and the MSER region proposals (the edges obtained by the Canny edge detector are used as a boundary for the MSER regions and the background information is removed) (Tabassum & Dhondse, 2015). Furthermore, to reduce the effect of false positives, the proposed work performs stroke width transform (SWT) (Epshtein et al., 2010) over the results of the fused step. The SWT considers the boundary pixels to estimate the stroke width, and we believe that the stroke width of every character is almost similar. Based on this observation, the proposed work fixes certain thresholds to remove nontext candidates. This step helps us to eliminate most of the nontext candidates as shown in Figure 3(d), where it can be seen that most of the text candidates are retained and most of the nontext candidates are removed. The result of SWT is called text components detection. It is noted from the results (see Figure 3(d)) that the output still contains nontext components. This is because of variations in the foreground and background. Note that this step does not remove nontext components at the cost of text components. Therefore, the steps retain all the text components. Since most of the background components are removed, the complexity of the text detection is also reduced. The five reduced text candidates having the highest IoU score are considered. The proposed regions are ROI pooled with the feature map of the Swin transformer network. This leads to obtaining good results by the Swin transformer irrespective of the challenges of drone images, which is the advantage of the introduction of MSER-based RPN. Another challenge of the drone dataset is the arbitrary angles as well as the varying scales of texts presented which creates difficulties for segmentation-based methods. MSERs are immune to affine transformations and perform multiscale detection. This provides an edge over other detection models and performs superiorly.

3.2. Swin transformer

In this study, the Swin transformer is selected for license plate number and text detection in drone images, as it is well-suited for representing data and extracting high-level features. While the vision transformer (Dosovitskiy et al., 2021) and data efficient transformer (Touvron et al., 2021) are designed for specific objectives, such as visual information and data collection, respectively, the Swin transformer (Liu et al., 2021) is capable of global and local self-attention, allowing it to extract context features globally and locally. This property is particularly useful for differentiating text components from nontext components, making it an ideal choice for the proposed work's objectives. Figure 4 shows the complete architecture of the Swin transformer blocks.

Figure 5 represents two consecutive Swin transformer blocks. The input to the first block is the encoded features z (after patch partition) which are passed on to the layer normalization, followed by the weighted multihead self-attention (MSA) layer. The output from the multilayer perceptron (MLP) is fed to the next block. Instead of W-MSA, the shifted window MSA is used for computational efficiency. The SW-MSA procedure is shown where the map is shifted by two units for performing the attention mechanism. To fill up the empty space, either padding is used or a more sophisticated approach of cyclic shift is applied and indicated using the green arrows.

The images are first divided into patches by a patch partition layer (e.g., H, W, three-dimensional image is divided into H/4, W/4, 48). These partitioned patches are passed on a linear embedding layer to project it into a dimension of C (H/4, W/4, C). In between the stages, (between two subsequent blocks) patch merging is done to reduce the number of patches (H/8, W/8, 2C) resulting in lower dimensional concatenated features. Swin transformer uses a shifted window attention mechanism to effectively reduce the computational burden.

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{1}$$

Equation (1) represents MSA, where h is the height of the image, w is the width of the image, and C is the dimension of the embedding vector of the image.

$$\Omega(WMSA) = 4hwC^2 + 2M^2hwC \tag{2}$$

Equation (2) is the window attention mechanism where MSA is applied not on the entire image but rather on a local window of nonoverlapping patches (window dimension is 7×7). A cyclic shift approach is adopted which introduces connections between neighboring overlapping windows like the convolutional neural networks as shown in Figure 5. This cross-window MSA increases the accuracy while reducing computation by eliminating redundant calculations. Each transformer block consists of linear regularization, MSA layer (number of attention heads is 8), and two-layer MLP with GELU activation function. After these blocks, a patch merging layer is used. The input image of dimension (128 × 128) is passed on to the embedding layer for

Figure 5
Illustration of two consecutive Swin transformer blocks and cyclic shift mechanism

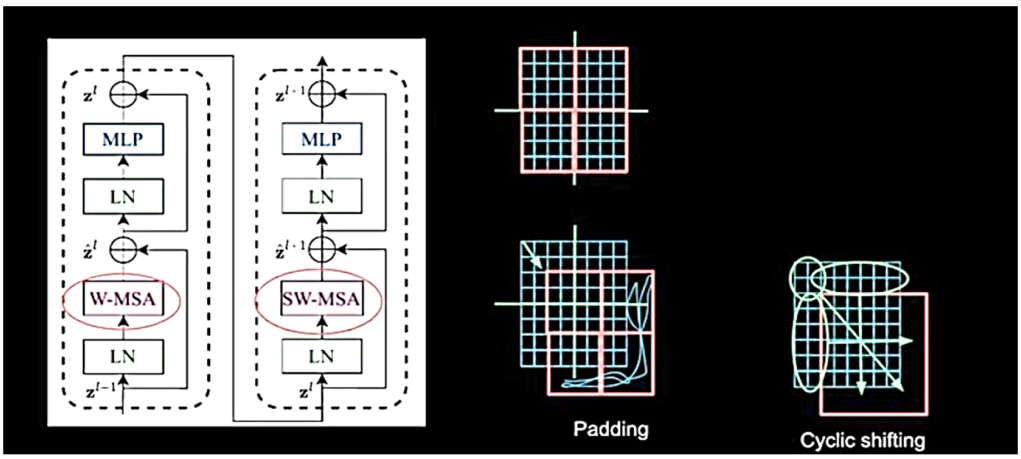


Figure 6
Text detection result of the proposed method for the images in Figure 3(d)

positional encoding of the patches (each patch is of size 4 × 4) due to the transformer's immunity to permutational changes. The addition of features of Blocks 2 and 3 before feeding them to Block 4 (similarly the features of Blocks 4 and 5 are added together) increases the accuracy of the model. The effect of the Swin transformer is illustrated in Figure 6 for the image shown in Figure 3(d), where one can see that the proposed model detects license plate numbers properly for all the vehicles in the drone image.

4. Experimental Results

To evaluate the proposed method on both drone images and normal scene images, we collected car images from Kaggle[1]. This dataset provides 432 images of cars with license plate number ground truth in the PASCAL VOC format. For the same dataset, our collected drone images are added to evaluate and validate the effectiveness of the proposed method for license plate number detection in drone and normal scene images. In total, the dataset contains 1142 images for experimentation, which include low resolution, degraded, good quality, poor quality, partially occluded license plate number images, and images with tiny text. Sample images of our dataset are shown in Figure 7(a) and (b), respectively, for normal images collected from the Kaggle dataset and drone datasets, where one can see the complexity of license plate number detection varies from one image to another. When we look at the sample images of the Kaggle and our datasets shown in Figure 7(a) and (b), the presence of multiple vehicles and background complexities is almost similar. However, the height distance varies much in the case of our dataset compared to the Kaggle dataset. To show that the proposed model works well for different situations, such as good-quality and poor-quality images, our dataset includes images of the Kaggle dataset for experimentation. Therefore, overall, we believe that the diversified images of our dataset reflect real scenarios. All the images of our dataset are resized to 256 × 256 dimensions and normalized (intensity of image normalized by dividing it by 255) before feeding the model.

To demonstrate the effectiveness of our proposed method, we conducted a comparative study with the state-of-the-art techniques that use powerful deep learning models and are robust to challenges similar to drone images. These methods include the differential binarization network (Liao et al., 2022), DRRG

[1]https://www.kaggle.com/datasets/andrewmvd/car-plate-detection

Figure 7
Examples of normal and drone license plate number images

Display images chosen from Kaggle website without detection results

Display images chosen from drone dataset without detection results

network (Zhang et al., 2020), and Fourier contour embedding network (Zhu et al., 2021). To ensure a fair comparison, we retrained these methods on our dataset and used a 70:30% split for training and testing data. We maintained a consistent experimental setup for all the experiments, which involved using an HP Laptop 15s-eq0xxx with an AMD Ryzen 5 3500U processor, 8GB RAM, and 2GB RADEON AMDA graphics card.

To evaluate and compare the performance of the proposed and existing methods, we use the commonly used metrics of precision, recall, and F1-score, which are defined in Equations (3), (4), and (5), respectively. These standard evaluation measures have also been used in the previous studies (Liao et al., 2022; Zhang et al., 2020; Zhu et al., 2021). We follow the same evaluation scheme as in these studies for calculating the metrics. By using these measures, we can assess the effectiveness of each method in terms of both accuracy and completeness of the license plate number detection. Similar to the traditional precision, recall, and F1-score, the measures are defined as follows for evaluating the performance of the methods. The pixels that are inside the ground truth bounding box are defined as true positives. The pixels that are inside the predicted box but are outside the ground truth bounding box are defined as false positives. The pixels that are outside the ground truth bounding box are defined as true negatives. The pixels that are outside the predicted bounding box but are inside the ground truth bounding box are defined as false negatives.

$$P = \frac{area(ground\ truth)\ \cap\ area(predicted\ box)}{area(predicted\ box)} \tag{3}$$

$$R = \frac{area(ground\ truth) \cap\ area(predicted\ box)}{area(ground\ truth)} \tag{4}$$

$$F1 = 2 \times \frac{(P\ \times\ R)}{(P + R)} \tag{5}$$

4.1. Ablation study

The adapted Swin transformer uses ResNet50 as backbone and the combination of MSER with RPN to improve the performance of license plate number detection in drone images. To validate the effectiveness of the above two key steps, we conducted the following experiments using our drone images dataset. The results are reported in Table 1. (i) Use the baseline RestNet101 for license plate number detection by feeding images as input. This is to test the effectiveness of the Swin transformer. (ii) In the same way, supply of input images to the ResNet50 instead of the RestNet101 for license plate number detection. (iii) Use of the ResNet50 as a backbone for Swin transformer without MSER + RPN. This is to test the effectiveness of the MSER + RPN. (iv) The proposed method that considers the ResNet50 as a backbone to Swin transformer and the steps of MSER + RPN. Table 1 shows that the baseline architecture of the ResNet50 is better than the baseline architecture of the ResNet101 in terms of F-measure. In this case, the precision increases for the ResNet50 while recall decreases for the ResNet50 compared to the ResNet101. Therefore, the ResNet50 is good for reducing the number of false positives while the ResNet101 is good for detecting text instances in the images. Since the precision of the ResNet50 gained more

Table 1
Assessing the efficacy of key steps in the proposed method for license plate detection

Experiments	Methods	Precision	Recall	F1-score
(i)	Baseline ResNet101	42.4	33.6	37.4
(ii)	Baseline ResNet50	48.6	31.9	38.5
(iii)	Proposed method without MSER-RPN	50.3	48.6	49.4
(iv)	Proposed method with MSER-RPN	79.8	77.9	78.9

Figure 8
Qualitative results of the proposed method for license plate number detection on different datasets

Detection results for the images chosen from Kaggle dataset

Detection results for the images chosen from Drone image dataset

than 6% over the ResNet101, recall of the ResNet101 gained more than 2% over the recall of the ResNet50. In addition, overall, the F1-score performance is better for the ResNet50 compared to the ResNet101. Thus, one can infer that the ResNet50 is effective for license plate number detection in drone images. However, when we compare the performance of baseline architectures and the proposed method, one can conclude that baseline architectures are not capable of achieving the best results for drone images.

The results (iii) and (iv) reveal a significant difference in the performance of the proposed method with and without MSER + RPN. Specifically, the proposed method without MSER + RPN exhibits inferior performance in comparison to the proposed method with MSER + RPN. This highlights the importance of MSER + RPN in improving the accuracy of the proposed method for detecting license plate numbers in drone images.

4.2. Experiments on license plate number detection

In order to evaluate the efficacy of the proposed method, qualitative results obtained from sample images of Kaggle and drone datasets are depicted in Figure 8(a) and (b), respectively. As observed from Figure 8, the proposed method is able to accurately detect license plate numbers in all images, even in the presence of multiple adverse factors. These results demonstrate the method's effectiveness in detecting license plate numbers in both normal and drone images. Similar conclusions can be drawn from the quantitative results presented in Table 2, which indicate that the proposed method achieves the highest recall and F1-score when compared to existing methods. The poor performance of existing methods can be attributed to their lack of suitability for drone images, as they were developed exclusively for text detection in scene images.

When we compare the results of the existing methods (Liao et al., 2022; Zhang et al., 2020; Zhu et al., 2021) reported in Table 2, the performance of the FCE (Zhang et al., 2020) is better than DRRG (Zhu et al., 2021). This is due to the model in Zhang et al. (2020) using the frequency domain to represent text instances while the model (Zhu et al., 2021) uses the spatial domain for representing text instances. It is true that the frequency domain can represent complicated shapes accurately compared to the spatial domain. However, the postprocessing steps used in Zhang et al. (2020) and Zhu et al. (2021) to improve the detection performance are not robust, and hence, the models (Zhang et al., 2020; Zhu et al., 2021) report poor precision compared to the model (Liao et al., 2022). In Liao et al. (2022), the method uses optimized postprocessing steps to overcome the limitations of the steps used in Zhang et al. (2020) and Zhu et al. (2021). However, overall, the existing models (Liao et al., 2022; Zhang et al., 2020; Zhu et al., 2021) report poor results compared to the proposed model in terms of recall and F-score.

From these results, it can be inferred that methods developed for text detection in normal scene images may not perform well when applied to drone images. Conversely, the proposed method is capable of effectively detecting license plate numbers in both types of images, thereby demonstrating its versatility and applicability across various settings. This is because of the contribution of MSER-based RPN for text component detection and the advantage of the Swin transformer. However, the method (Liao et al., 2022) reports the highest precision compared to the other existing method and the proposed method. This is because the performance of the method depends on postprocessing unlike other existing methods, and it is an end-to-end model for scene text detection. In the case of our method, the use of Canny edge

Table 2
The performance of the proposed method and existing techniques for license plate detection in both normal and drone images in (%)

Methods	Precision	Recall	F1-score
DBNet++ (Liao et al., 2022)	90.97	61.04	73.06
DRRG (Zhang et al., 2020)	64.96	54.88	59.50
FCENet (Zhu et al., 2021)	90.96	66.00	76.50
Proposed	79.86	77.99	78.91

Figure 9
Some failure cases of the proposed method

images and MSER step sometimes detects nontext components as text components. The reason is Canny and MSER are sensitive to complex backgrounds and degradations in the images, and hence, the step introduces spurious edges for the complex background. As a consequence, the proposed method may generate a higher number of false positives, which in turn can result in poor precision.

4.3. Limitations

As mentioned earlier, our proposed method may fail when the input images are hazy, have poor resolution, or are noisy in nature. This is demonstrated in sample images in Figure 9, where the method misses characters or does not correctly identify bounding boxes. This may be due to the sensitivity of the Canny edge detector used in the RPN to noise, resulting in inaccurate bounding box predictions. Therefore, there is room for improvement by replacing the step of text component detection with a new deep learning model or end-to-end transformer, which we plan to investigate in future work.

The processing time for license plate number detection using our proposed model is 7.2 FPS, which may not be optimal. This is due to the large number of parameters and computations involved in region proposal calculation. However, the processing time is affected by various factors, including system configuration, programming, and platform. Our focus in this work is to address the problem of drone images rather than achieving the lowest processing time. In future work, we aim to develop a system that can be used in real-time environments.

5. Conclusion and Future Work

The proposed method in this study combines MSER and Swin transformers to detect license plate numbers in both normal and drone images. The MSER and RPN are used for detecting text components in drone images despite the challenges that come with them. The Swin transformer is adapted to detect license plate numbers in both drone and normal images. Experimental results on our dataset, which includes license plate images from both normal and drone scenes, demonstrate that the proposed method outperforms the state-of-the-art methods in terms of recall and F1-score. However, severe degradations in images can cause the performance of the proposed method to deteriorate. Nonetheless, this issue falls outside the scope of this study. To tackle such challenges in the future, the step of text component detection could be replaced with a new transformer.

Conflicts of Interest

Palaiahnakote Shivakumara is an Editor-in-Chief and Umapada Pal is an Advisory Board Member for *Artificial Intelligence and Applications*, and were not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflicts of interest to this work.

References

Bagi, R., Dutta, T., Nigam, N., Verma, D., & Gupta, H. P. (2021). Met-MLTS: Leveraging smartphones for end-to-end spotting of multi-lingual oriented scene texts and traffic signs in adverse meteorological conditions. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 12801–12810.

Chen, X., Guo, Q., Li, S., & Zhang, J. (2017). Holistic vertical regional proposal network for scene text detection. *In 2017 2nd International Conference on Image, Vision and Computing*, 72–77.

Chowdhury, P. N., Shivakumara, P., Nandanwar, L., Samiron, F., Pal, U., & Lu, T. (2022). Oil palm tree counting in drone images. *Pattern Recognition Letters*, *153*, 1–9.

Dai, P., Li, Y., Zhang, H., Li, J., & Cao, X. (2021). Accurate text detection via scale-aware data augmentation and shape similarity constraint. *IEEE Transactions on Multimedia*, *24*, 1883–1895.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ..., & Houlsby, N. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations*, 1–21.

Dwivedi, A. K., Singh, A. K., & Singh, D. (2022). An object-based image analysis of multispectral satellite and drone images for precision agriculture monitoring. In *2022 IEEE International Geoscience and Remote Sensing Symposium*, 4899–4902.

Epshtein, B., Ofek, E., & Wexler, Y. (2010). Detecting text in natural scenes with stroke width transform. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2963–2970.

Gizatullin, Z. M., Lyasheva, M. M., Shleymovich, M. P., & Lyasheva, S. A. (2022). Automatic car license plate detection based on the image weight model. In *2022 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering*, 1346–1349.

Gómez, L., & Karatzas, D. (2014). MSER-based real-time text detection and tracking. In *2014 IEEE 22nd International Conference on Pattern Recognition*, 3110–3115.

Jain, S., Patel, S., Mehta, A., & Verma, J. P. (2022). Number plate detection using drone surveillance. In *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering*, 1–6.

Kim, S. S., Jung, H. T., Lee, S. J., Park, J. H., Yu, S. H., & Go, J. H. (2022). A study of real-time 4K drone images visualization to rescue for missing people based on web. In *2022 IEEE 13th International Conference on Information and Communication Technology Convergence*, 1594–1596.

Kim, T. G., Yun, B. Y., Kim, T. H., Lee, J. Y., Park, K. H., Jeong, Y., & Kim, H. D. (2021). Recognition of vehicle license plate based on image processing. *Applied Science*, *11*(14), 6292.

Lee, Y., Jeon, J., Ko, Y., Jeon, M., & Pedrycz, W. (2022). License plate detection via information maximization. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 14908–14921.

Liao, M., Zou, Z., Wan, Z., Yao, C., & Bai, X. (2022). Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 919–931.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., & Raptis, M. (2022). Towards end-to-end unified scene text detection and layout analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1039–1049.

Ma, J., Liang, Z., & Zhang, L. (2022). A text attention network for spatial deformation robust scene text image super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5901–5910.

Mittal, A., Shivakumara, P., Pal, U., Lu, T., & Blumenstein, M. (2022). A new method for detection and prediction of occluded text in natural scene images. *Signal Processing: Image Communication*, *100*, 116512.

Mohite, J., Sawant, S., Agrawal, R., Pandit, A., & Pappula, S. (2022). Detection of crop water stress in maize using drone based hyperspectral imaging. In *2022 IEEE International Geoscience and Remote Sensing Symposium*, 5957–5960.

Nadanwar, L., Shivakumara, P., Ramachandra, R., Lu, T., Pal, U., & Antonacopoulos, A. (2022). A new deep wavefront-based model for text localization in 3D video. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(6), 3375–3389.

Soni, A., Dutta, T., Nigam, N., Verma, D., & Gupta, H. P. (2022). Supervised attention network for arbitrarily-shaped text detection in edge-fainted noisy scene images. *IEEE Transactions on Computational Social Systems*, 1179–1188.

Srilekha, B., Kiran, K.V. D., & Pradyala, V. V. P. (2022). Detection of license plate numbers and identification of non-helmet rider using Yolov2 and OCR method. In *2022 IEEE International Conference on Electronics and Renewable Systems*, 1539–1549.

Tabassum, A., & Dhondse, S. (2015). Text Detection Using MSER and Stroke Width Transform. In *2015 IEEE Fifth International Conference on Communication Systems and Network Technologies*, 568–571. https://doi.org/10.1109/CSNT.2015.154.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357.

Wang, W., Xie, E., Li, X., Liu, X., Liang, D., Yang, Z., . . . , & Shen, C. (2022). PAN++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. *IEEE Transactions on Analysis, Pattern and Intelligence, Machine*, *44*(9), 5349–5367.

Zeng, C., & Song, C. (2022). Swin transformer with feature pyramid networks for scene text detection of secondary circuits cabinet wiring. In *2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems*, 255–258.

Zhang, R., Xu, L., Yu, A., Shi, Y., Mu, C., & Xu, M. (2022a). Deep-IRTarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation. *IEEE Transactions on Multimedia*, *24*, 1735–1749.

Zhang, R., Yang, S., Zhang, Q., Xu, L., He, Y., & Zhang, F. (2022b). Graph-based few-shot learning with transformed feature propagation and optimal class allocation. *Neurocomputing*, *470*, 247–256.

Zhang, S. X., Zhu, X., Hou, J. B., Liu, C., Yang, C., Wang, H., & Yin, X. C. (2020). Deep relational reasoning graph network for arbitrary shape text detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9699–9708.

Zheng, J. (2022). Multiple-level alignment for cross-domain scene text detection. In *2022 IEEE 2nd International Conference on Consumer Electronics and Computer Engineering*, 671–175.

Zhu, Y., Chen, J., Liang, L., Kuang, Z., Jin, L., & Zhang, W. (2021). Fourier contour embedding for arbitrary-shaped text detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3123–3131.

How to Cite: Pal, S., Roy, A., Shivakumara, P., & Pal, U. (2023). Adapting a Swin Transformer for License Plate Number and Text Detection in Drone Images. *Artificial Intelligence and Applications* 1(3), 129–138, https://doi.org/10.47852/bonviewAIA3202549