Journal of Advanced Marine Engineering and Technology, Vol. 49, No. 6, pp. 551~559, 2025
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2025.49.6.551

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

Free-form document detection and ratio-preserving rectification

Min-Jae Kim[1] · Yoon-Sang Han[2] · Dong-Hoan Seo[†]

(Received October 13, 2025 ; Revised October 29, 2025 ; Accepted December 17, 2025)

Abstract: Document detection and rectification are essential preprocessing steps for transforming mobile camera images of document into images suitable for optical character recognition (OCR). Conventional approaches rely on geometric features such as corners and edges to detect documents and assume standardized formats (e.g., A4) for rectification. Although these methods do not require a flatbed scanner, they are limited to specific paper sizes. In this paper, we propose a framework for detecting and rectifying arbitrary rectangular documents without assuming fixed aspect ratios. Our approach introduces CornerNet, which estimates the probability distributions of the four document corners, thereby enabling robust detection at arbitrary viewing angles. Based on these detected corners, we also designed a ratio-preserving rectification module that estimates the intrinsic aspect ratio of the document and normalizes it to a reference size. Finally, a sub-pixel refinement step corrects the rectified corners to increase precision. The experimental results demonstrate that the proposed method achieves accurate corner localization and rectification that preserves the aspect ratio, enabling reliable document scanning from unconstrained mobile phone images.

Keywords: Document detection, Document rectification, Free-form document

1. Introduction

Optical character recognition (OCR) is a widely used technology that digitizes documents, which are the primary medium of information in human society **[1][2]**. Traditionally, OCR has relied on flatbed scanners to acquire document images. However, with the recent advances in smartphone camera capabilities, research has increasingly focused on digitizing documents by mobile devices. Early approaches assumed that documents were photographed from a frontal view under controlled conditions, yielding images similar to those from scanners, albeit at a reduced resolution **[3]**. More recently, the research has expanded to address documents captured from a distance or at arbitrary angles, aiming to detect documents, reduce noise, and enhance resolution to produce scan-quality results. Such technological progress has enabled OCR-based content extraction to be broadly applied in various fields, such as on-site operations, education, and research, and has thus become an active area of investigation worldwide.

On mobile devices, it is inherently difficult for users to capture documents in a perfectly vertical orientation, and the workload increases significantly when many documents are processed. As a result, document detection must assume a freeform environment, in which documents are captured at arbitrary angles. A document is essentially a rectangle with four right-angled corners and edges, each corresponding to a specific aspect ratio that depends on the format. Early studies therefore adopted template-based approaches to identify rectangular shapes **[4][5]**. However, in freeform environments, perspective distortions caused by camera pose introduce geometric deformations. Consequently, the corners are no longer right-angled and the edge lengths differ, rendering the template-based approaches ineffective.

To address these challenges, traditional studies have employed edge-based approaches to detect document corners. Even under perspective distortion, the edges remain linear; thus, bottom-up methods analyze edge intersections as corners to infer the shape of a

† Corresponding Author (ORCID: http://orcid.org/0000-0003-3610-0356): Professor, Division of Electronics and Electrical Information Engineering & Interdisciplinary Major of Maritime AI Convergence, National Korea Maritime and Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: dhseo@kmou.ac.kr, Tel: 051-410-4412

1 M. S. Candidate, Department of Electrical and Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, National Korea Maritime and Ocean University, E-mail: kminjae2926@gmail.com, Tel: +82-51-410-4822

2 M. S., Department of Electrical and Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, National Korea Maritime and Ocean University, E-mail: hanysang@gmail.com, Tel: +82-51-410-4822

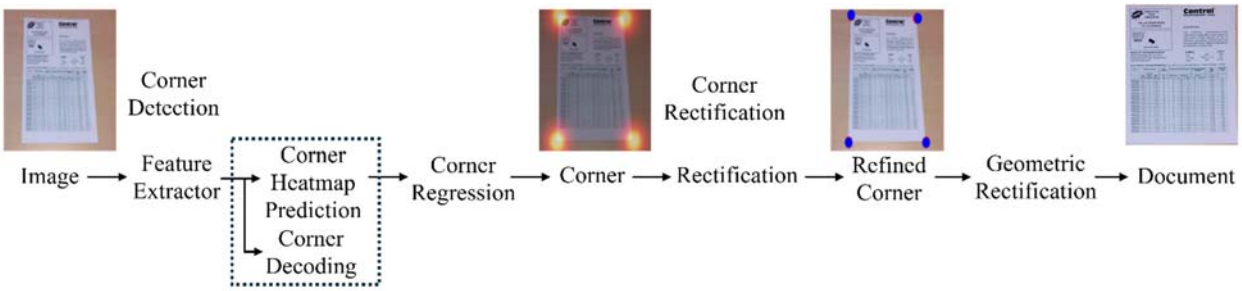

Figure 1: Overall pipeline of the proposed document detection and rectification framework.

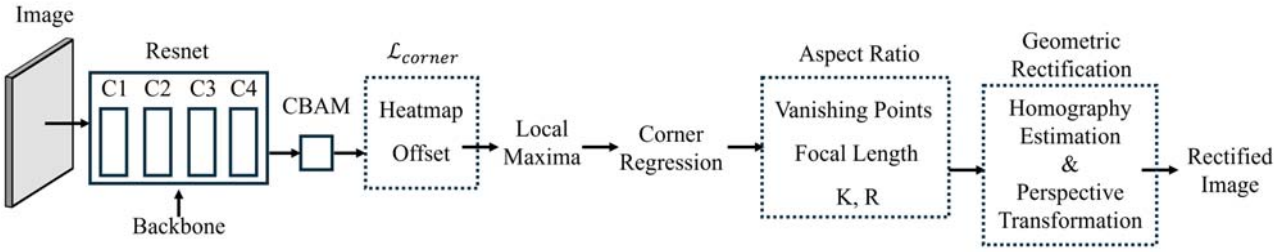

Figure 2: Detailed architecture of the document detection module in the proposed framework

document **[6]-[8]**. Although these methods are robust against perspective distortion, they are vulnerable to background noise, which limits their accuracy.

With the advent of deep learning, bottom-up approaches have been further advanced through keypoint detection, which leverages feature representations to localize corners more accurately **[9]-[11]**. Feature extraction using powerful backbone models such as ResNet is highly effective because corners are relatively simple features to detect, and more complex representations can also be exploited to determine the presence of documents. However, most of these approaches regress corner coordinates directly, which can be less accurate than more probabilistic formulations.

In contrast, top-down approaches such as template matching directly detect entire documents. Recent advances in object detection have been extended to instance segmentation, which localizes object regions at the pixel level, and several studies have attempted to apply this paradigm to document detection **[12]-[15]**. In this approach, the document region is segmented and the detected edges are refined to obtain more accurate document boundaries. These methods exhibit strong robustness against background noise and camera pose variations. However, aliasing frequently appears along the edges, thereby reducing the accuracy of document detection. Because document detection is inevitably followed by planar rectification, the precise acquisition of boundary information is critical.

Document rectification generally refers to the process of restoring a document to a predefined format based on its detected corners. Unlike flatbed scanning, freeform capture inevitably requires perspective correction, making rectification an essential step. In most cases, rectification assumes a standardized format, such as A4. However, in practice, documents may be folded or nonstandard in size, resulting in inaccurate aspect ratios when such assumptions are applied.

To this end, we propose a novel framework that enables the accurate detection and rectification of documents in images captured at arbitrary angles and aspect ratios. In contrast to conventional methods that assume fixed aspect ratios, the proposed framework can handle documents of any rectangular shape. First, CornerNet, which is proposed in this study, estimates the probability distribution of the four document corners, thereby enabling robust detection under free-form capture conditions. Using the detected corners, the ratio-preserving rectification module estimates the intrinsic aspect ratio of the document and normalizes it to a reference size. Finally, a subpixel refinement step further adjusts the rectified corner positions to increase precision.

2. Related Work

Recent studies have proposed various methods that jointly perform document detection and rectification, with representative examples including FDRNet **[16]**, RDLNet **[17]**, DocReal **[18]**, and DocScanner **[19]**. FDRNet **[16]** integrates control point prediction, Thin-Plate Spline(TPS)-based rectification, Fourier-based restoration, and recognition to enhance OCR performance. RDLNet **[17]** employs a Light-SAM ViT-based encoder and a masked-attention decoder to extract

document corner coordinates, instance-level segmentation masks, and document categories. DocReal **[18]** adopts a segmentation-based detection step followed by an attention-enhanced control point module and linear interpolation for rectification, whereas DocScanner **[19]** predicts a warping flow using a progressive correction module and applies bilinear interpolation to rectify distorted documents.

Document detection has traditionally relied on segmentation-based approaches that approximate document boundaries as quadrilaterals. However, under perspective distortions, the boundaries often appear as irregular polygons, making it difficult to consistently extract four stable corner points. Furthermore, noise introduced during contour approximation can result in mismatches between the detected corners and actual document corners, thereby degrading detection accuracy.

Document rectification aims to correct perspective distortions and curvatures that occur in captured documents, serving as a critical step for improving OCR performance. Existing approaches are broadly categorized into geometry-based **[20][21]** and learning-based **[22][23]** methods. Geometry-based methods typically estimate a perspective transformation from four detected corners, which is simple and efficient but limited in handling curved documents. By contrast, learning-based methods leverage predicted warping maps or control points to perform TPS-based rectification, thereby enabling recovery from both perspective and nonlinear distortions. These methods, however, are computationally expensive and highly dependent on training data. More recently, end-to-end architectures that integrate detection and rectification have been introduced, enabling simultaneous document localization and rectification **[22][23]**.

Accordingly, this study adopts a learning-based keypoint detection strategy that directly predicts the four document corners. This choice ensures both computational efficiency and robustness during the rectification stage, thereby providing a stable foundation for subsequent perspective corrections.

3. Proposed Method

In this paper, we proposed a novel framework for accurate document detection and rectification from images captured at arbitrary angles and aspect ratios. Unlike conventional methods, the proposed framework does not assume a fixed document aspect ratio and can handle any rectangular document. It employs CornerNet to estimate the probability distribution of four document corners, thereby enabling robust detection under freeform capture conditions. Using the detected corners, the ratio-preserving rectification module estimates the intrinsic aspect ratio and normalizes it to the reference size. Finally, a subpixel refinement step further adjusts the corner positions to increase precision. **Figure 1** illustrates the overall workflow of the proposed framework, which consists of corner detection, corner rectification, and geometric rectification, while **Figure 2** details the internal network architecture and processing steps corresponding to these stages.

3.1 Overview

As **Figure 1** shows, the input is an RGB image of size W×H, and a document is defined by its four corner points in the image coordinate system (x, y), ordered clockwise as top left (TL), top right (TR), bottom-right (BR), and bottom-left (BL).

To extract document features, a ResNet-based backbone is employed in combination with an attention mechanism. From the extracted features, document corners are detected using two parallel heads that estimate the heat maps and offsets. The heat maps select the top four candidate corners in the image, whereas the offsets refine the detected integer corner positions to subpixel accuracy based on the ground-truth values.

The four detected corners are then used to compute the homography matrix with respect to the reference coordinate system. Using this homography, a warp transformation aligns the document corners across multiple images, correcting perspective distortions. Finally, a subpixel refinement step adjusts the corner positions with high precision, resulting in accurately localized document regions.

3.2 Document Corner Detection and Rectification

Figure 2 illustrates the architecture of the document detection component in the proposed framework. For clarity, the framework is divided into a corner detection and rectification stage. In the detection stage, we adopt a CornerNet-based architecture, in which a ResNet-34 backbone is employed for feature extraction. The stride of the first block in layer 4 is adjusted from 2 to 1, preserving the spatial resolution of the output feature map at 1/8th of the input resolution. All other layer configurations remain identical to the standard ResNet-34, and no additional architectural modifications are introduced because of the relatively simple nature of the single-document detection task. The convolutional block attention module (CBAM) is applied to effectively capture fine-grained features, even in the presence of complex backgrounds or ambiguous document

boundaries. The feature maps from the C4 stage of ResNet are passed through CBAM to obtain the final C5 feature maps. In this work, the standard CBAM configuration is adopted in which channel attention precedes spatial attention. For the channel attention module, the bottleneck multilayer perceptron uses a reduction ratio of 16, whereas for the spatial attention module, the convolution employs a kernel size of seven. CBAM is applied as a lightweight auxiliary module without additional structural modifications because the standard configuration is sufficient to enhance feature saliency with minimal computational overhead for the single-document detection task. The overall document shape is represented as a pixel-level segmentation mask that allows the module to learn spatial attention and accurately distinguish the document region. This process enhances document-relevant features while suppressing background noise.

The framework employs a heatmap-based representation to precisely locate corners. Each corner candidate is mapped to a two-dimensional Gaussian distribution to generate a heat map, and the four peaks with the highest values in the image are defined as the document corners. Because heatmap-based predictions are limited to integer pixel coordinates, offset regression is also performed to refine the corner positions to subpixel accuracy.

The detected corners then undergo a ratio-preserving rectification process to ensure geometric consistency. This step corrects document tilt by estimating two vanishing points from pairs of parallel edges, where each vanishing point is computed as

$$v_x = (p_1 \times p_2) \times (p_3 \times p_4),\ v_y = (p_2 \times p_3) \times (p_4 \times p_1) \tag{1}$$

Using the orthogonality constraint

$$(\mathrm{K}^{-1}v_{x)}{}^{T}, \tag{2}$$

the focal length f is estimated, and the intrinsic matrix

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

is constructed using the principal point (c_x, c_y). The camera rotation matrix R that aligns the document plane with a fronto-parallel view is then derived, and the rectification transform

$$H_c = KR \tag{4}$$

maps the detected corners into a metric-rectified coordinate space. In this space, the true aspect ratio is obtained from the Euclidean distances of adjacent rectified corners as

$$\text{aspect} = \frac{\|x_2' - x_1'\|}{\|x_3' - x_2'\|}, \quad x_i' = H_c^{-1} p_i, \tag{5}$$

Using this physically consistent aspect ratio, a target rectangle is defined, and a homography is estimated between the original corner coordinates and this rectangle to perform the final perspective warp. This produces a geometrically consistent rectification and makes the corner configuration closer to an ideal rectangle, improving the stability of the subsequent corner refinement.

In the refinement stage illustrated in **Figure 2**, the proposed model applies a sub-pixel based correction to estimate document corners with higher precision. The sub-pixel correction first transforms the document into an approximately rectified rectangular form through a primary perspective warp using the initially predicted four corner coordinates. In this rectified space, ideal rectangular corner positions are initialized, and the corner locations are re-estimated in continuous coordinate space by exploiting local image gradient information at sub-pixel resolution. This refinement step is not a simple integer-grid matching process, but rather an optimization procedure that adjusts the corner positions by searching for extrema guided by local brightness gradients. The sub-pixel refined corner coordinates are subsequently mapped back to the original image coordinate space, thereby reducing residual geometric distortions that cannot be eliminated by the initial rectification alone and ultimately increasing the precision of document boundary estimation.

3.3 Training Loss

To train the document detection network, we designed a loss function that combines two outputs. The overall loss $\mathcal{L}_{Corner}$ is defined as

$$\mathcal{L}_{corner} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}^{(i)}{}_{heatmap} + \mathcal{L}^{(i)}{}_{offset}), \tag{6}$$

where $\mathcal{L}_{heatmap}$ and $\mathcal{L}_{offset}$ correspond to the losses of the segmentation map and the offset map, respectively. Since the outputs are normalized, no additional weighting coefficients are required.

The segmentation map loss $\mathcal{L}_{heatmap}$, which learns the center locations, is based on the focal loss **[21]** and is formulated as

$$\mathcal{L}_{heatmap} = -\frac{1}{N}\sum_{x,y}\begin{cases}(1-\hat{H}_{xy})^{\alpha}\log(\hat{H}_{xy}) & \text{if } H_{xy}=1\\ (1-H_{xy})^{\beta}(\hat{H}_{xy})^{\alpha}\log(1-\hat{H}_{xy}) & \text{otherwise}\end{cases}, \quad (7)$$

where $\hat{H}_{xy}$ and H_{xy} denote the predicted and ground-truth segmentation maps, respectively, and α, β are hyperparameters. N represents the number of positive corner points.

Finally, the offset loss $\mathcal{L}_{offset}$ is formulated as an L1 loss as follows:

$$\mathcal{L}_{offset} = \frac{1}{N}\sum_{x,y}\left[T_{xy}\cdot\left\|\hat{O}_{xy}-O_{xy}\right\|_1\right], \quad (8)$$

where only the offsets corresponding to the true centers are considered via the mask T_{xy}. Here, $\hat{O}_{xy}$ and O_{xy} represent the predicted and ground-truth offsets, respectively.

4. Experiment

4.1 Dataset and Implementation

Recent research has mainly focused on document detection for OCR; however, datasets of single-document images captured in mobile environments remain limited. To address this limitation, we conducted single-document detection experiments using the publicly available SmartDoc dataset. SmartDoc was originally used in the ICDAR 2015 SmartDoc Challenge 1 and consists of videos recorded with five different simple backgrounds and six types of document forms for each background. The videos include variations in document orientation, position, and perspective. The entire dataset is composed of 150 videos with an average length of 10 s.

To ensure experimental reliability, the dataset was split into training/validation/test sets using an 8:1:1 ratio, ensuring that the same file belonged to only one split. All videos were sampled at 10-frame intervals.

Table 1: Comparison of performance on the SmartDoc dataset

Method	Recall	Precision	F1
Zhu *et al.* **[5]**	0.89	0.89	0.89
EAST **[4]**	0.80	0.78	0.79
MS RCNN **[13]**	0.92	0.89	0.90
SOLO **[15]**	0.93	0.92	0.92
Pan *et al.* **[6]**	0.95	0.94	0.94
Ours	0.99	0.99	0.99

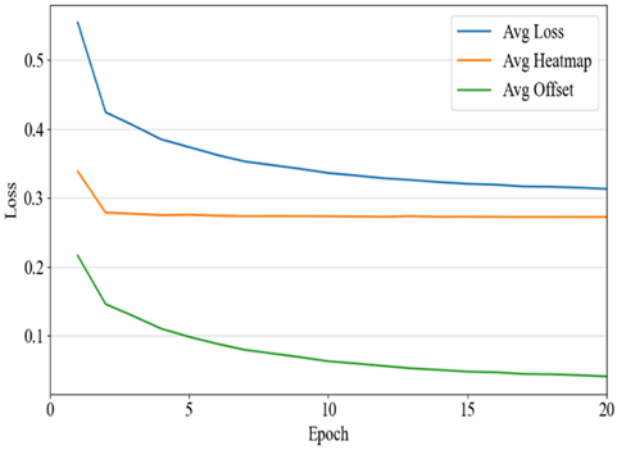

Figure 3: Training losses for corner detection

In addition, we independently collected approximately 2,000 single-document images using mobile devices under more challenging capture conditions, including larger tilt angles, to better simulate realistic scenarios. These images were annotated following the same protocol as SmartDoc and were used as an additional evaluation set.

The annotation for these images followed the same protocol as that of the original SmartDoc dataset, labeling the four corners in clockwise order (TL, TR, BR, and BL) and adding coordinate information. All final images had a resolution of 1920×1080, and for memory efficiency, model inputs were normalized to 960×540. The model was trained using the Adam optimizer with a learning rate of 1×10^{-4}. Experiments were conducted on an NVIDIA GeForce GTX TITAN X GPU.

Although, document detection methods can be evaluated using the SmartDoc dataset, as document scanning using mobile devices has become more common, it has become necessary to consider a wider range of capture angles. Therefore, in this paper, we additionally images captured under larger tilt angles in the experiments, and evaluated the performance of the proposed model using the mean absolute error (MAE), root mean squared error (RMSE), and mean IoU.

4.2 Performance Evaluation of the Proposed Framework

Figure 3 presents the variation in corner prediction losses during the training process. In the graph, the x-axis represents the number of training epochs, whereas the y-axis denotes the prediction loss values for the document corners. The average loss is computed as a weighted sum of the heatmap loss and the offset loss, serving as an indicator to evaluate whether both losses contribute in a balanced manner to updating the model parameters. The heatmap loss decreases sharply within the first

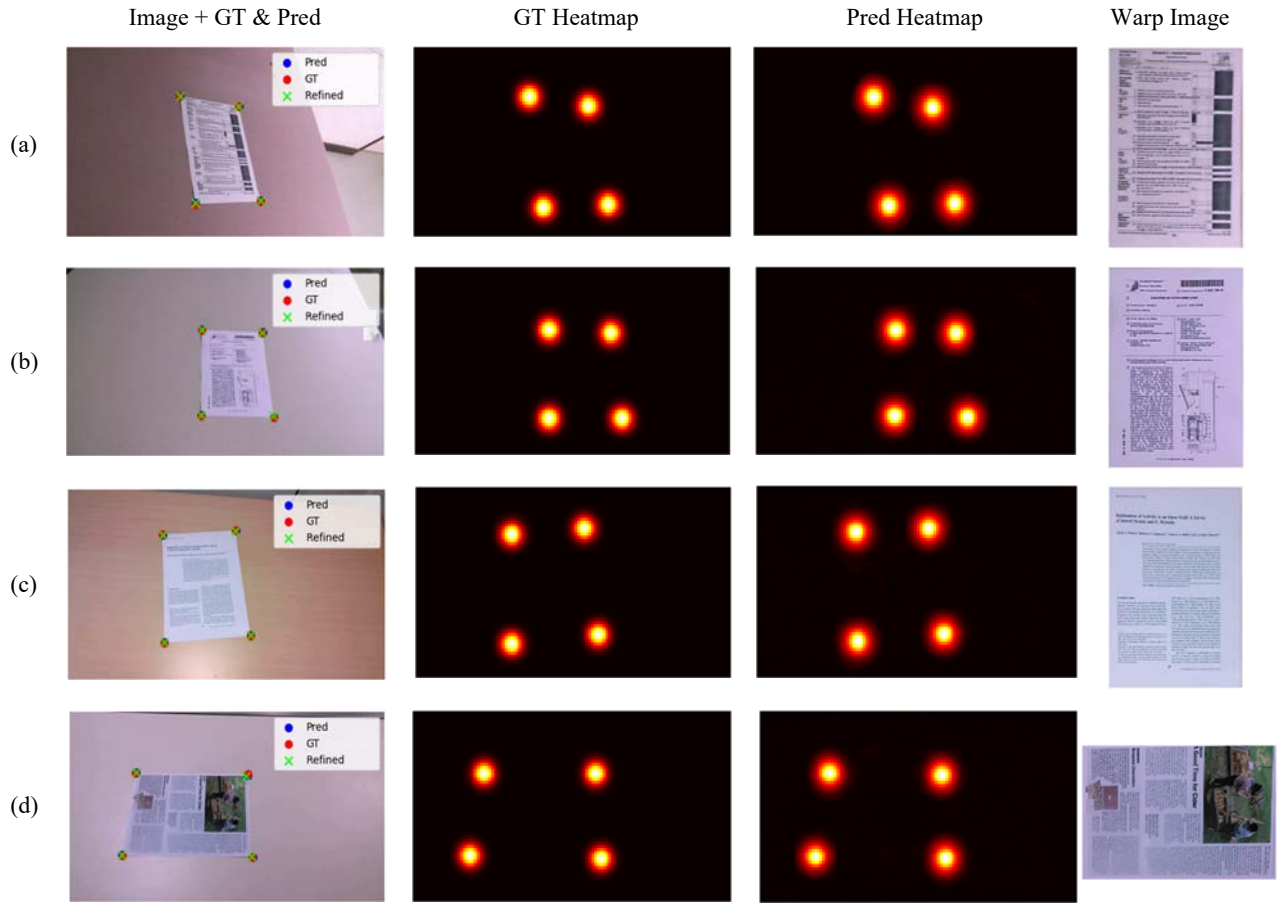

Figure 4: Qualitative results of the proposed framework

five epochs and tends to converge rapidly, whereas the offset loss converges to a stable value within approximately ten epochs. Therefore, training the model for around ten epochs is sufficient; however, performing a few additional epochs may further enhance stability.

In this study, the performance of single-document recognition was evaluated using the SmartDoc dataset, and the results for the proposed method and five comparison methods (Zhu *et al.* **[5]**, EAST **[4]**, MS RCNN **[13]**, SOLO **[5]**, and Pan *et al.* **[6]**) are presented in **Table 1**. Detections with an IoU of 0.9 or higher were considered correct, and the precision, recall, and F1-score metrics were calculated.

The method of Zhu *et al.* **[5]**, which directly regresses corner coordinates, performed poorly due to its sensitivity to geometric variations. The EAST method adopts a rectangle-based detection approach; however, due to the characteristics of regional semantic segmentation, the outer boundaries of documents were blurred, resulting in low overall accuracy. MS RCNN is based on instance segmentation and achieves relatively good results but suffers from a significant drop in precision because document shapes were not fully preserved under background interference conditions. SOLO, an instance segmentation-based method, performed well; however, its detection was limited when the object's estimated position was uncertain. Pan *et al.*, which attempts to improve detection by combining corners and edges, remained vulnerable to incorrect matches.

In contrast to the comparison approaches, the proposed model demonstrated stable and consistent performance in noisy environments by leveraging corner heatmaps using an approach grounded in the structural properties of documents. Moreover, because it detects document regions by extracting boundaries based on corner points, it achieved stable performance that is superior to those of existing approaches.

Table 2: Performance analysis on an additional evaluation dataset collected by the authors

Method	Ours
MAE	5.9012
RMSE	7.8026
mIoU	0.9538

Table 2 presents the quantitative results of the error between the predicted corner coordinates and the ground truth. The extended SmartDoc dataset, which includes more variations in angle, was used for this experiment. In conventional studies, document detection typically evaluates accuracy based on whether the predicted region and the ground truth overlap sufficiently to classify them as the same object. However, in such cases, the precision for downstream tasks such as document rectification cannot be guaranteed. Hence, we directly verify the accuracy of the predicted corner coordinates instead using MAE and RMSE. In addition, by reporting the mIoU, we report more quantitative values that are more precise than those of the conventional approach, which considers an IoU above 0.9 as the same object. The proposed method achieved an MAE of 5.9012 pixels, RMSE of 7.8026 pixels, and mIoU of 0.9538. The pixel-level MAE and RMSE correspond to an accuracy within 1% on a full high definition screen. Furthermore, the results of the proposed method exceed the conventional IoU threshold of 0.9, indicating that the proposed model accurately captured the detected document region. These results show that the corner prediction performance and the transformed results obtained by rectifying the corner placement meet the purpose of the design.

Figure 4 shows the qualitative detection results for a single example document. The left column presents the predicted coordinates, ground truth coordinates, and refined coordinates in the input images, while the middle two columns visualize the probability maps of the four document corners. The right column shows examples of the final results. As can be seen in the figure, the document is accurately detected, and the corner positions are predicted reliably. In particular, even when the document's size and position vary due to perspective distortion, the corner refinement allows for precise estimation of the entire document area. These results demonstrate that the proposed model provides high-precision detection and refinement performance in single-document scenarios.

4.3 Limitation and Discussion

Accurate detection alone is not sufficient for the OCR processing of document images; precise normalization is also essential. In the method proposed in this paper, documents are analyzed based on the horizontal and vertical axes, which do not necessarily align with the actual text orientation. Therefore, for future applications, an additional text orientation determination step will be required.

4. Conclusion

In this study, we proposed a novel framework for accurate document detection and rectification from images captured at arbitrary angles and with various aspect ratios. In the proposed framework, the novel method CornerNet predicts the probability distribution of the four document corners, while the ratio-preserving rectification module and sub-pixel refinement enable precise estimation and correction of document regions. Experiments on the SmartDoc dataset demonstrate that the proposed model reliably detects corners and achieves stable rectification performance in single-document scenarios, with an MAE of 5.9012 pixels, RMSE of 7.8026 pixels, and mIoU of 0.9538, indicating highly accurate boundary estimation.

Future work will extend the framework to account for document folds and physical damage. It will also consider robust detection and rectification in multi-document scenarios, thereby further enhancing its applicability in practical OCR pipelines.

Acknowledgement

This research was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (RS-2021-KA162922602820634902O1).

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (RS-2024-00352187).

This work was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE). (RS-2024-00424595, Regional Residency Program for Cultivating Advanced Research Talent in Next-Generation Marine Mobility Industry Innovation).

Author Contributions

Conceptualization, M. J. Kim and Y. S. Han; Methodology, M. J. Kim; Software, Y. S. Han; Formal Analysis, M. J. Kim; Investigation, M. J. Kim; Resources, Y. S. Han; Data Curation Y. S. Han; Writing-Original Draft Preparation, M. J. Kim; Writing-Review & Editing, Y. S. Han; Visualization, Y. S. Han; Supervision, D. H. Seo; Project Administration, D. H. Seo; Funding Acquisition, D. H. Seo.

References

[1] A. Poznanski and L. Wolf, "CNN-n-gram for handwritingword recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[2] X. Zhang, Y. Bengio, and C. Liu, “Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark,” Pattern Recognition, vol. 61, pp. 348-360, 2017.

[3] J. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. Luqman, and M. Mehri, “ICDAR2015 competition on smartphone document capture and OCR,” in International Conference on Document Analysis and Recognition, pp. 1161-1165, 2015.

[4] K. Javed and F. Shafait, “Real-time document localization in natural images by recursive application of a CNN,” in IEEE Conference on Document Analysis and Recognition, 2017.

[5] A. Zhu, C. Zhang, Z. Li, and S. Xiong, “Coarse-to-fine document localization in natural scene image with regional attention and recursive corner refinement,” International Journal on Document Analysis and Recognition, vol. 22, no. 3, pp. 351-360, 2019.

[6] R. Pan and A. Zhu, “Multi-document detection via corner localization and association,” Neurocomputing, vol. 466, pp. 37-48, 2021.

[7] G. Binmakhashen and S. Mahmoud, “Document layout analysis: A comprehensive survey,” ACM Computing Surveys, vol. 52, no. 6, pp. 1-36, 2019.

[8] Y. Qiao, Q. Hu, G. Qian, S. Luo and W. Nowinski, “Thresholding based on variance and intensity contrast,” Pattern Recognition, vol. 40, no. 2, pp. 596-608, 2007.

[9] R. Gioi, J. Jakubowicz, J. Morel, and G. Randall, “LSD: A fast line segment detector with a false detection control,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 4, pp. 722-732, 2010.

[10] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476-3483, 2013.

[11] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “Centernet: Keypoint triplets for object detection,” in IEEE/CVF International Conference on Computer Vision, pp. 6568-6577, 2019.

[12] X. Li, C. Lv, W. Wang, G. Li, L. Yang, and J. Yang, “Generalized focal loss: Towards efficient representation learning for dense object detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 3, pp. 3139-3153, 2023.

[13] K. He, G. Gkiozari, P. Dollar, and R. Girshick, “Mask r-cnn,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386-397, 2020.

[14] A. Kirillov, Y. Wu, K. He, and R. Girshick, “Pointrend: image segmentation as rendering,” in IEEE Conference on Computer Vision and Pattern Recognition, pp. 9796-9805, 2020.

[15] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, “Solo: Segmenting objects by locations,” in European Conference on Computer Vision, 2020.

[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W. Lo, P. Dollar, and R. Girshick, “Segment anything,” in IEEE/CVF International Conference on Computer Vision, 2023.

[17] H. Bandyopadhyay, T. Dasqupta, N. Das, and M. Nasipuri, “RectiNet-v2: A stacked network architecture for document image dewarping,” Pattern Recognition Letters, vol. 155, pp. 41-47, 2022.

[18] F. Hertlein, A. Naumann, and Y. Sure-Vetter, “DocMatcher: Document image dewarping via structural and textual line matching,” in IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5771-5780, 2025.

[19] K. Ma, Z. Shu, X. Bai, J. Wang, and D. Samaras, “Docunet: Document image unwarping via a stacked u-net,” in IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700-4709, 2018.

[20] S. Das, K. Ma, Z. Shu, D. Samaras, and R. Shilkrot, “Dewarpnet: Single-image document unwarping with stacked 3D and 2D regression networks,” in IEEE/CVF International Conference on Computer Vision, pp.131-140, 2019.

[21] P. Kumari and S. Das, “Am I readable? Transfer learning based document image rectification,” International Journal on Document Analysis and Recognition, vol. 27, no. 3, pp. 433-446, 2024.

[22] C. Xue, Z. Tian, F. Zhan, S. Lu, and S. Bai, “Fourier document restoration for robust document dewarping and recognition,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4563-4572, 2022.

[23] Y. Wu, Z. Xu, Y. Duan, Y. Wu, Q. Zheng, H. Li, *et al.*, “RDLNet: a novel and accurate real-world document localization method,” In Proceedings of the 32nd ACM

International Conference on Multimedia, pp. 9847-9855, 2024.

[24] F. Yu, Y. Xie, L. Wu, Y. Wen, G. Wang, S. Ren, and W. Li, “Docreal: Robust document dewarping of real-life images via attention-enhanced control point prediction,” in IEEE/CVF Winter Conference on Applications of Computer Vision, 2024.

[25] H. Feng, W. Zhou, J. Deng, Q. Tian, and H. Li, “DocScanner: Robust document image rectification with progressive learning,” International Journal of Computer Vision, vol. 133, pp. 5343-5362, 2025.