Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 5 (No. 2), pp. 81-86, 2025
doi: 10.54569/aair.1845016
Published online: Dec 25, 2025

An End-to-End Deep Learning Architecture for Information Extraction from Turkish Identity Cards

Ömer Can Eskicioğlu [1*], Ali Hakan Işık [2]

[1] Burdur Mehmet Akif Ersoy University, Faculty of Bucak Computer and Informatics, Department of Software Engineering, Burdur, Türkiye
[2] Burdur Mehmet Akif Ersoy University, Faculty of Engineering, Department of Computer Engineering, Burdur, Türkiye

Abstract

With the acceleration of digital transformation in the service sector, remote customer acquisition and identity verification processes have become the cornerstone of secure ecosystems. Particularly in internet-based services, image distortions, perspective errors, and variable lighting conditions encountered during the transfer of physical documents to the digital environment are the most significant factors complicating data extraction. In this study, a deep learning-based end-to-end architecture is proposed that enables fast, secure, and high-accuracy information extraction from Turkish Republic Identity Cards. In the proposed system, while CURL is used to enhance image quality, a YOLOv8m-based instance segmentation model is preferred for detecting the boundaries of the card. For the determination of card orientation and perspective correction, a novel hybrid approach has been developed that analyzes the cosine distance between face biometrics obtained via RetinaFace and the segmentation mask. This structure, in which PAN for text detection and Transformer-based TrOCR models for character recognition are integrated, was tested on a unique dataset augmented with the CLoDSA library. Experimental results indicated that the YOLOv8m model exhibited success in card detection with a 99.5% mAP score. Our proposed model demonstrates that it offers an efficient solution for digital identity verification processes with an overall accuracy rate of 92.6%.

Keywords: *Digital Identity Verification; YOLOv8; RetinaFace; Perspective Correction; TrOCR; Deep Learning.*

1. Introduction

The digitalization process of societies has become an important research topic in recent years in both academia and the service sector. The digitalization of service infrastructures has brought about the conversion of physical documents into digital formats. In this newly created ecosystem, the use of digital identity is a necessity for users to integrate into the system and establish a presence in the digital world [1]. Therefore, the reliable verification of these identities has become an inevitable imperative for service providers.

Traditional identity verification methods fall short against today's sophisticated fraud techniques and cannot keep up with the pace of digital transformation. In the present day, where digital literacy is accepted as a core competency, users demand that services be much more efficient, secure, and accessible compared to traditional methods.

The secure and autonomous transfer of physical identity cards to the digital environment is a necessity today. This process is a significant part of the transformation from physical assets to digital assets. This study covers the stages of high-accuracy detection of identity cards and the secure extraction of the data contained therein. Upon examining current research on the subject, new approaches and development processes in the literature are discussed below.

The MIDV-2020 dataset was proposed by Bulatov et al. [2] with the aim of addressing data diversity and the lack of privacy, which are among the biggest constraints in extracting and processing features in identity documents. The proposed work contains 72,409 annotated images in scan, photo, and video formats belonging to 1000 different mock identity documents. This study constitutes one of the most comprehensive open datasets in the literature. The authors analyzed document and text field detection under different lighting, blur, and angle conditions. Similarly, in another study focusing on data scarcity and privacy [3], 90,000 blurred and noisy identity card images belonging to identity cards such as UIDAI and PAN were generated. In the study, the authors used the Differentiable Binarization algorithm for the detection of texts on the card. In the recognition stage, they proposed an OpenNMT architecture with an Attention mechanism. It was observed that the proposed approach yielded more successful results than traditional systems.

Although attempts are made in the literature to overcome the data scarcity problem through synthetic methods, the complex layout features of documents are considered a significant challenge in the information extraction process. In this context, Yu et al. [4] proposed the PICK framework by focusing on visual and

*Corresponding author
E-mail address: eskiciogluomer@gmail.com

Received: 19/Dec/2025; Accepted: 23/Dec/2025.

structural features in addition to textual features. In the study, the relationships between text segments were modeled using Graph Learning and Convolution techniques. In this way, higher performance was achieved compared to baseline methods. However, although graph-based approaches like PICK are successful in structural relationships, they execute information extraction and text reading processes sequentially. Evaluating these processes independently of each other presents disadvantages, such as the propagation of errors to the subsequent stage. In a study by Zhang et al. [5], a unified structure was proposed in which the aforementioned processes mutually reinforce each other. The study combined the textual and visual features from the text reading stage with the information extraction stage. It was observed that this method yielded better results than existing methods in terms of efficiency and accuracy. Although Transformer architecture-based networks possess advanced capabilities, they rely on OCR outputs for operations on text sequences. This dependency significantly hinders the model's optimal performance in cases of noisy or incorrect ordering. Addressing this ordering problem, Gu et al. [6] proposed the XYLayoutLM model. In the study, accurate and ordered layout information extraction was performed using the Augmented XY Cut algorithm. Furthermore, a flexible solution capable of processing variable-length inputs was presented using the Dilated Conditional Position Encoding module.

In the training of multimodal models, the use of different methods for visual and textual expressions can lead to misalignment issues. To address this problem, the LayoutLMv3 model was proposed by Huang et al. [7]. In their study, the authors did not evaluate textual and visual expressions separately but trained them with a common unified masking. Additionally, a specialized alignment method was used to find the equivalent of the word on the image. The model achieved high performance in document understanding and classification tasks. One of the biggest obstacles for such models is their dependency on the language used. To overcome training difficulties in documents of different languages or texts with different layouts, the LiLT (Language-independent Layout Transformer) architecture was proposed by Wang et al. [8]. In this study, the learning of the document layout was addressed independently of the text content. Thus, layout information learned in a single language can be applied and used in different languages. In the study tested in eight different languages, it was observed that high performance was obtained. Apart from language features, the interactions between text and images in visual document analysis directly affect model success. In this context, the DocFormer architecture was proposed by Appalaraju et al. [9]. In their study, text and visual features found in the document were combined using a multi-modal self-attention layer. Furthermore, a shared spatial embeddings method was applied to strengthen the relationship between text and visual expressions. With this architecture, more successful results were obtained compared to larger models.

While document understanding models are developing, traditional approaches based on architectures like CNN and RNN continue to be used in text recognition processes. Efficiency problems may arise due to the need for extra language models and the model complexity of these architectures. To solve this problem, the TrOCR model was proposed by Li et al. [10]. In this model, an end-to-end structure based on the Transformer architecture was used. In the proposed architecture, the Image Transformer and Text Transformer operate in conjunction. It was observed that the model performed better than existing models in the literature in operations such as printed text and handwriting recognition.

Although basic text recognition models have an end-to-end structure, a trend towards integrated models capable of working simultaneously has begun in the literature. In a study by Tang et al. [11], the UDOP (Universal Document Processing) model, which combines text, layout, and image models under a single framework, was proposed. The study was presented with a prompt-based approach using a Vision-Text-Layout Transformer. The proposed work achieved good results in different tasks and ranked first in the Document Understanding Benchmark. Despite the successes of these integrated universal models, LLM (Large Language Model) integration becomes critical due to challenges such as image distortions and data imbalance in real-world scenarios. In this context, in the thesis study conducted by Freitter [12], a three-stage structure was proposed. The study primarily consists of segmentation, classification, and text extraction modules. A YOLOv8-based approach was presented for card detection, and a ResNet-based approach for classification. For text recognition and extraction, it was proposed to support the PaddleOCR model with the Llama 3.2 Vision model. In the study, an accuracy rate of 96.72% was obtained in identity card information extraction.

When this development process in the literature is examined, it is seen that there is an evolution from approaches focusing on singular tasks towards integrated systems. However, factors such as language support constraints and layout differences limit the direct applicability of universal models in local and real-world problems. In this study, an architecture specific to Turkish Republic Identity Cards is proposed by integrating object segmentation, image enhancement, and optical character recognition (OCR) techniques. With this proposed approach, it is aimed to develop a high-accuracy, reliable, and practical method in the process of digitizing physical identities.

2. Materials and Methods

This study presents a multi-stage architecture aimed at extracting text data by resolving structural

distortions on Turkish Republic Identity Cards. The method begins with instance segmentation for the precise determination of the card's boundaries. Subsequently, perspective distortions were corrected using the spatial relationships between the face biometrics and corner coordinates on the card. PAN (Pixel Aggregation Network) was integrated for the detection of text fields in the obtained rectified images, while TrOCR models were integrated for character recognition and extraction. Figure 1 illustrates the general pipeline of the proposed model.

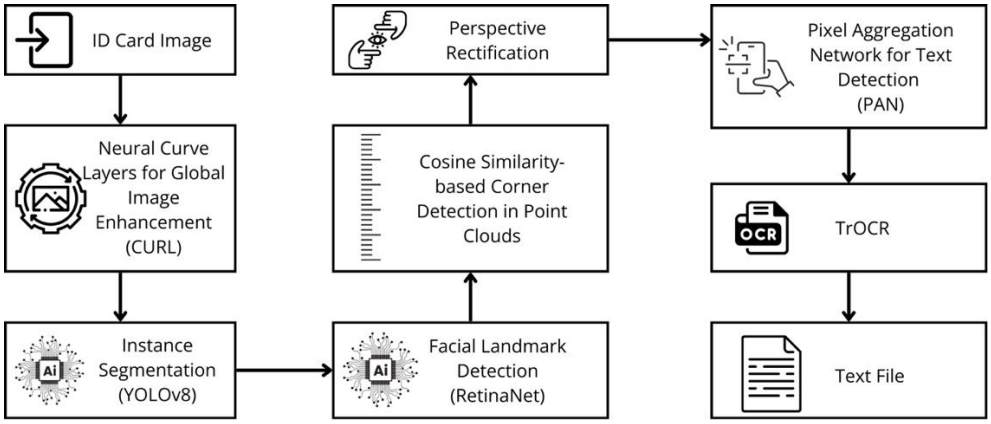

Figure 1. *General Pipeline of the Proposed Model*

2.1. Dataset and Preprocessing

During the dataset creation phase, three videos, each two minutes in length and recorded against varying backgrounds, were utilized as source material. The videos were sampled in accordance with the frames per second (FPS) rate, and raw images were obtained. The obtained images were annotated for the instance segmentation task using CVAT (Computer Vision Annotation Tool) [13], an open-source annotation tool developed for image and video analysis. In this process, 700 images were processed, with each containing an average of 50 coordinate points.

To enhance image quality and minimize noise resulting from environmental factors, the CURL (Neural Curve Layers for Image Enhancement) [14] model, which utilizes neural curves in illumination and color space adjustments, was applied to the raw images; thereby, clarity and contrast optimization were achieved.

To increase data diversity on the annotated and enhanced images, CLoDSA [15], a data augmentation library specialized for object detection tasks, was employed. According to the configuration detailed in Table 1, 10 new derived images were generated from each original image. As a result of the augmentation process, the total number of data points reached approximately 7000. In the final stage, the images and label files were converted into the YOLO format, which enables single-stage and real-time object detection, and were prepared for model training.

Table 1. *CLoDSA Data Augmentation Configuration*

	Technical Name	Code Function	Parameters	Description
1	Horizontal Flip	flip	{"flip": 1}	Flips the image along the horizontal axis.
2	Vertical Flip	flip	{"flip": 0}	Flips the image along the vertical axis.
3	Rotation (+15°)	rotate	{"angle": 15}	Rotates 15 degrees clockwise.
4	Rotation (-15°)	rotate	{"angle": -15}	Rotates 15 degrees counter-clockwise.
5	Rotation (90°)	rotate	{"angle": 90}	Rotates 90 degrees.
6	Shearing	shearing	{"a": 0.1}	Slightly shifts the perspective of the image.
7	Darkening	gamma	{"gamma": 1.5}	Darkens the image/increases contrast via gamma correction.
8	Brightening	gamma	{"gamma": 0.6}	Makes the image brighter via gamma correction.
9	Noise	salt_and_pepper	{"low": 0, "high": 255}	Adds random black and white dots to the image.
10	Horizontal Flip	flip	{"flip": 1}	Flips the image along the horizontal axis.

The dataset was partitioned into 75% training, 10% validation, and 15% test sets. Accordingly, model training was conducted with 5250 images, while 700 images were reserved for validation and 1050 images for the testing phase. A 99.5% mAP score was achieved as a result of training lasting 50 epochs with the YOLOv8m architecture.

Following the instance segmentation process performed on the identity card, the deep learning-based

RetinaFace architecture [16], capable of detecting faces at various scales and under challenging conditions with high success, was utilized for the precise detection of the face region on the card. In this process, the WIDER FACE [17] dataset was utilized to evaluate the performance of the model on noisy and low-resolution images. The said dataset consists of 393,703 labeled face images within 32,203 images possessing extreme pose, illumination, and scale variations.

2.2. Detecting ID Card Process

The identity card detection and alignment process consists of sequential steps. First, the boundaries of the card were determined via Instance Segmentation, and subsequently, the location of the face photograph on the card was detected using RetinaFace. A hybrid approach was adopted for the verification of the card's corner points. The cosine distance between the characteristic features found on the face (eyes, nose, etc.) and the boundary points obtained from the segmentation output was analyzed across four separate regions with vertical and horizontal axes. As a result of this analysis, the four corner points and the orientation of the card were determined with high accuracy. An exemplary illustration of the process is provided in Figure 2.

Figure 2. *Face and Corner Detection on the Identity Card*

Following the determination of corner coordinates and card orientation, perspective transformation was applied to the image. Through this procedure, angled or warped card images were flattened, making them suitable for OCR processing. The rectified card image is shown in Figure 3.

Figure 3. *Application of the Perspective Correction Process on the Identity Card*

For the detection of text fields on the aligned and flattened identity card image, the segmentation-based

PAN [18] architecture, which is capable of detecting texts by clustering pixel density, was utilized. The recognition and interpretation of characters within the detected text boxes were carried out using the TrOCR [10] model, which combines image and text processing pipeline with an end-to-end Transformer structure. The text detection and recognition stages are presented in Figure 4.

Figure 4. *Text Detection on the Identity Card*

3. Results and Discussion

The end-to-end identity card analysis architecture developed in this study was trained and tested on a unique dataset containing varying lighting conditions, background clutter, and perspective distortions.

When the training process of the model is examined, it is observed that the loss values (box_loss, seg_loss, cls_loss) on the training and validation sets show a steady downward trend. The validation box loss (val/box_loss), which was at the level of 0.34 at the beginning of training, decreased to 0.119 at the end of the 50th epoch; while the segmentation loss (val/seg_loss) declined from 0.32 to 0.148. This decrease indicates that the model's capacity to learn card boundaries and masks increased as training progressed and that it extracted generalizable features without experiencing the overfitting problem.

Detection and Segmentation Performance When the performance metrics of the YOLOv8m instance segmentation model are evaluated, it is seen that the model's detection precision reached 0.999 and its recall reached 0.995. The balanced and high values of precision and recall indicate that the model minimizes false positives and possesses high coverage regarding card detection.

The mAP@50 value, accepted as the general performance metric, was measured as 99.5% at the end of training. Achieving a high score like 0.995 in the mAP@50-95 metric reveals that the model not only locates the card but also overlaps the card boundaries (IoU - Intersection over Union) with very high precision.

End-to-End System Performance Alongside segmentation success, card orientation and perspective errors were successfully resolved with the RetinaFace-supported biometric alignment module. This alignment success had a direct positive impact on the OCR stage. It was determined that the system, integrated with the Transformer-based TrOCR model, achieved an information extraction accuracy of 92.6% even in noisy and low-resolution images. These obtained results demonstrate that the proposed architecture offers a more stable and high-performance solution compared to traditional methods in the literature.

4. Conclusions

In this study, a deep learning-based architecture is proposed for the rapid and automated transfer of Turkish Republic Identity Cards to the digital environment. Within the scope of the study, an end-to-end structure encompassing image enhancement, object detection, biometric-referenced perspective correction, and OCR processes has been developed.

The results obtained in the study demonstrated that the YOLOv8m-based segmentation model exhibited high success in card detection with a 99.5% mAP score and operated robustly despite complex backgrounds. The hybrid alignment approach, which is one of the most original contributions of the study and utilizes face biometrics obtained via RetinaFace, minimized layout deviations and perspective distortions. Thanks to the input image provided by this rectification process, the overall information extraction accuracy of the system integrated with the Transformer-based TrOCR model resulted in 92.6%. The obtained findings demonstrate that the proposed architecture offers a fast and high-accuracy solution. The developed system has the potential

to increase operational efficiency in finance, banking, telecommunications, e-commerce, public services, insurance, and tourism processes. In future studies, it is aimed to apply optimization techniques for the model to operate in real-time on mobile devices, and to integrate modules such as blockchain technology, liveness analysis, and fraud detection into the system.

Declaration of interest

No conflict of interest was declared by the authors.

Acknowledgements

This study was supported by Burdur Mehmet Akif Ersoy University Scientific Research Projects Commission. Project Number: 0832-YL-22. This study was produced from the first author's thesis titled 'Image Processing Based Identity Recognition and Liveness Analysis System with Microservice'.

References

[1] U. Mir, A. K. Kar, and M. P. Gupta, "AI-enabled digital identity–inputs for stakeholders and policymakers," Journal of Science and Technology Policy Management, vol. 13, no. 3, pp. 514-541, 2022.

[2] B. K. Bulatovich et al., "MIDV-2020: A comprehensive benchmark dataset for identity document analysis," Компьютерная оптика, vol. 46, no. 2, pp. 252-270, 2022.

[3] M. K. Gupta, R. Shah, J. Rathod, and A. Kumar, "Smartidocr: Automatic detection and recognition of identity card number using deep networks," in 2021 Sixth International Conference on Image Information Processing (ICIIP), 2021, vol. 6: IEEE, pp. 267-272.

[4] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao, "Pick: processing key information extraction from documents using improved graph learning-convolutional networks," in 2020 25th International conference on pattern recognition (ICPR), 2021: IEEE, pp. 4363-4370.

[5] P. Zhang et al., "Trie: end-to-end text reading and information extraction for document understanding," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1413-1422.

[6] Z. Gu et al., "Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4583-4592.

[7] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," in Proceedings of the 30th ACM international conference on multimedia, 2022, pp. 4083-4091.

[8] J. Wang, L. Jin, and K. Ding, "Lilt: A simple yet effective language-independent layout transformer for structured document understanding," arXiv preprint arXiv:2202.13669, 2022.

[9] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "Docformer: End-to-end transformer for document understanding," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 993-1003.

[10] M. Li et al., "Trocr: Transformer-based optical character recognition with pre-trained models," in Proceedings of the AAAI conference on artificial intelligence, 2023, vol. 37, no. 11, pp. 13094-13102.

[11] Z. Tang et al., "Unifying vision, text, and layout for universal document processing," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 19254-19264.

[12] S. Freitter, "Automating ID Card Verification leveraging Deep Learning," Technische Universität Wien, 2025.

[13] B. Sekachev et al., "opencv/cvat: v1. 1.0," Zenodo, 2020.

[14] S. Moran, S. McDonagh, and G. Slabaugh, "Curl: Neural curve layers for global image enhancement," in 2020 25th International Conference on Pattern Recognition (ICPR), 2021: IEEE, pp. 9796-9803.

[15] Á. Casado-García et al., "CLoDSA: a tool for augmentation in classification, localization, detection, semantic segmentation and instance segmentation tasks," BMC bioinformatics, vol. 20, no. 1, p. 323, 2019.

[16] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5203-5212.

[17] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5525-5533.

[18] W. Wang et al., "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8440-8449.