INSTITUTE OF INFORMATION
AND COMMUNICATION TECHNOLOGIES
BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **26**, No 2
Sofia • 2026 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2026-0015

A Modular Privacy-Preserving Framework for Travel Document Segmentation and Information Extraction Using Synthetic Data

Plamen Nakov[1], *Petar Petrov*[2], *Georgi Kotov*[1], *Milena Lazarova*[1], *Ognyan Nakov*[1]

[1]*Technical University of Sofia, 8 Kl. Ohridski Blvd, 1000 Sofia, Bulgaria*
[2]*DSS, 5A Baku str, 1700 Sofia, Bulgaria*
E-mails: p.nakov@tu-sofia.bg petar.petrov@dss.bg gkotov@tu-sofia.bg milaz@tu-sofia.bg nakov@tu-sofia.bg

***Abstract*:** *Automated travel document recognition is a key technology for digital identity verification. However, robust extraction of structured information from images captured in unconstrained conditions remains challenging due to perspective distortion, background clutter, motion blur, and heterogeneous lighting that often degrade the performance of the systems. The paper proposes a modular pipeline for automated travel document segmentation and data extraction that integrates instance segmentation, perspective rectification, optical character recognition, and rule-based field parsing. In order to avoid the use of sensitive personal data, the segmentation model is trained exclusively on a synthetic dataset generated in Blender that comprises 2500 annotated images with diverse variations in lighting, viewpoint, blur, and background. The experimental results demonstrate strong generalization from synthetic to real data with 99.50% mAP50, 99.22% mAP50-95, 90% character-level Optical Character Recognition (OCR), and 90% MRZ field extraction accuracy on synthetic data, and 88% MRZ extraction accuracy on a dataset with real documents.*

***Keywords*:** *YOLOv11, Document segmentation, Perspective rectification, PaddleOCR, Travel document OCR, Structured data extraction, Computer vision.*

1. Introduction

Automated recognition of travel documents such as passports, identity cards, and visas has become a critical element of modern digital identity systems. Services like electronic onboarding, border management, access control, and remote account verification depend on high-accuracy document reading. However, in practical deployments and real-world scenarios, travel documents are often captured with mobile devices or built-in cameras by non-expert users under unconstrained conditions rather than scanned in controlled environments. As a result, the acquired images frequently exhibit perspective distortion, non-uniform illumination, motion blur, glare, background clutter, partial occlusion, and other artifacts, which can substantially reduce the reliability of the recognition systems even for machine-readable travel documents with standardized formats [1-4].

Document image analysis has tackled tasks such as localization, rectification, enhancement, Optical Character Recognition (OCR), and structured information extraction. Classic OCR approaches are usually applied directly to raw images and

often produce unstable or incomplete outputs, especially when the document region is poorly localized or geometrically misaligned. Travel documents such as passports and identity cards contain both free-form visually printed text and highly structured Machine-Readable Zones (MRZs) that have different requirements on extraction accuracy. Minor recognition errors may invalidate extracted fields.

Moreover, travel documents contain sensitive personal information that limits the assembly and annotation of large-scale real-world datasets for model training and evaluation. Recent studies suggest the utilization of synthetic datasets in order to avoid sensitive information exposure and lower the data annotation cost [5-7].

Deployment-ready systems for travel document recognition require accurate localization and text recognition, reliable geometric normalization, structured field extraction, as well as high accuracy and fast inference, interpretable outputs, and reliable integration in a privacy-sensitive environment. End-to-end models that are accurate in laboratory settings are difficult to deploy in production contexts as they usually require computationally expensive training on a vast amount of real-world training data.

The paper proposes a modular pipeline for automated travel document segmentation and data extraction that integrates four sequential stages: (1) YOLO-based model for document localization by instance segmentation; (2) perspective rectification based on quadrilateral approximation of the segmentation mask; (3) PaddleOCR-based text recognition; (4) rule-based structured field parsing based on regular expressions and validation logic. The proposed framework balances the robustness, transparency, and deployment efficiency in a privacy-sensitive setting by combining data-driven visual localization with deterministic geometric correction and interpretable field extraction.

A central aspect of the proposed approach is the utilization of synthetic training data at the segmentation stage that comprises rendered images with automatically generated ground-truth masks and domain randomization to simulate realistic acquisition conditions, including changes in viewpoint, lighting, motion blur, and background complexity. The used approach for synthetic travel data generation creates high-quality pixel-level annotations without exposing sensitive real personal data during model development.

The main contributions of the paper are as follows:

- A modular and deployment-oriented framework is proposed that integrates segmentation, rectification, OCR, and structured parsing into a coherent travel-document recognition pipeline;
- A privacy-preserving strategy for training a segmentation model on a synthetic dataset is utilized that allows effective model generalization to real travel-document imagery.
- A focused, interpretable, and computationally efficient framework architecture is proposed that combines instance segmentation with geometric normalization to improve the reliability of text extraction under unconstrained imaging conditions.

The rest of the paper is organized as follows. Section 2 reviews related work on travel document image analysis. Section 3 describes the proposed framework.

Section 4 presents the experimental evaluation. Section 5 discusses the results, and Section 6 concludes the paper.

2. Related work

Automated travel document recognition requires identity-document image processing and analysis as well as text recognition under unconstrained acquisition conditions. In the digital identity verification systems, documents such as passports, identity cards, and visas are frequently captured with mobile devices rather than flatbed scanners that might introduce blur, glare, skew, background clutter, and illumination changes. These factors make travel document recognition a more specialized and demanding problem than generic OCR.

Document recognition systems evolved from template matching and strict layout rules that expect scanned images under ideal conditions [8] towards more robust and accurate deep learning-based systems. Classical OCR engines assume minimal distortion and uniform backgrounds and are applied directly to cropped document regions. These systems are effective under constrained imaging conditions such as flatbed scanning; however, they show significant performance degradation in real-world scenarios. Furthermore, template-based approaches lack robustness to layout variations across document types and issuing authorities.

An important subtask in travel document analysis is the recognition of the MRZ. ICAO Doc 9303 [1] defines the structure, encoding rules, and field constraints for machine-readable travel documents. MRZ lines are central to the automated processing of passports, visas, and other travel documents because they encode highly structured identity information in a standardized format and allow rule-based parsing. At the same time, accurate MRZ extraction from mobile images remains difficult due to geometric distortion, blur, reflections, and low image quality that can corrupt character sequences and degrade recognition accuracy. Robust localization and rectification are critical for reliable recognition, especially in unconstrained imaging scenarios [9, 10]. The structure of MRZ content makes post-processing particularly important, as the strict formatting rules and positional regularities enable rule-based validation and correction after the OCR stage.

Document image analysis systems are usually multi-stage and comprise subtasks such as document detection, skew correction, binarization, layout analysis, OCR, and structured information extraction. Moreover, failures in early stages, such as localization or rectification, can directly degrade text recognition and field extraction, especially for identity document recognition, where text is dense, layouts might be semi-structured, and both visual and machine-readable content must be processed reliably.

Reliable OCR of travel documents depends heavily on accurate localization and geometric normalization of the captured images. Text recognition accuracy drops if the document is not isolated from the scene or if perspective distortion is not corrected. Perspective correction and document rectification are crucial for robust recognition outside scanner-based settings. More recent deep learning approaches for object detection, instance segmentation, and lightweight OCR architectures

demonstrate significant improvement in the robustness of document recognition systems and the localization of documents in cluttered scenes. At the same time, travel document recognition remains a challenging field as it combines strict structural requirements, visually sensitive content, and privacy constraints that limit the availability of large annotated datasets. In [11], identity-document recognition is described as a distinct research problem with its own constraints, including sensitive content, diverse sources, and the need to generalize across different document types and image acquisition scenarios.

With the advancement of Convolutional Neural Networks (CNNs), document analysis moved toward data-driven visual localization and text recognition approaches. Region-based detectors and segmentation models such as Faster R-CNN and Mask R-CNN enable improved localization of document regions and text blocks and provide high detection accuracy, but often demand significant computational overhead that limits real-time deployment in mobile and embedded systems [12, 13]. Modern OCR frameworks utilizing various CNN and transformer architectures [14], such as Tesseract OCR [15], PP-OCR [16], PaddleOCR [17], EasyOCR [18], AWS Textract [19], have significantly lowered the technical barriers for end-to-end document understanding systems that combine detection, recognition, and post-processing in practical deployment environments. Models like YOLO have gained popularity for fast and accurate real-time object detection due to the single-stage architecture and favorable speed-accuracy trade-off [20]. In document analysis, YOLO-based models are applied for document boundary detection and text region localization, which enables pixel-level object masks in addition to bounding box prediction [21, 22].

The document recognition pipeline is composed of multiple interdependent stages; thus, performance degradation in one stage, such as localization or rectification, propagates directly to text recognition and field extraction [3, 23-25]. However, many previous works focus only on limited subtasks or narrow capture conditions without integrating subsequent geometric rectification and structured data extraction into a complete processing pipeline [25-27]. The existing approaches address isolated components of the pipeline, such as text detection, recognition, or document classification, rather than the full problem of robust end-to-end document analysis [5, 28]. However, deployable systems for travel and identity documents must function across diverse document types, image sources, and uncontrolled acquisition environments.

End-to-end monolithic document recognition models demonstrate strong performance on structured document benchmarks, but they typically require large annotated datasets, significant training capacity, and design choices that may reduce interpretability and modular flexibility in deployment-oriented applications and incremental deployment [24, 29]. Practical identity-document systems have to support reliable integration into operational workflows, interpretable intermediate outputs, predictable latency, and adaptability to multiple document types and acquisition settings. Modular architecture for production-oriented travel document analysis offers a practical path for real-world systems as it allows localization, geometric correction, OCR, and field parsing to be optimized and validated

independently [10, 21, 26, 27]. The modularity improves maintainability and debugging while preserving the option to replace components without redesigning the full system.

Furthermore, generic OCR datasets do not reflect the challenges of mobile identity document capture [27, 28]. Dedicated datasets relevant for mobile document analysis are utilized, such as MIDV-500, which contains video-based identity-document imagery with multiple document types and acquisition conditions [27]. A major bottleneck in travel document recognition systems is the scarcity of public training data. Passports, IDs, and visas contain personal data, thus making large-scale real travel document datasets difficult to collect, annotate, and release. This limitation has driven substantial interest in synthetic data generation for document analysis that reduces annotation cost by delivering exact pixel-level or field-level annotations, provides strong control over scene variability, including camera viewpoint, illumination, blur, and background composition, and mitigates privacy risks when real data are unavailable or sensitive [29, 30].

3. Proposed framework

The proposed modular privacy-preserving framework for travel document segmentation and information extraction from images captured under unconstrained conditions is designed as a pipeline that combines deep learning-based document localization with deterministic geometric correction, optical character recognition, and rule-based semantic parsing. In contrast to end-to-end architectures that directly predict structured outputs from raw images, the proposed framework decomposes the problem into a sequence of specialized stages, thus improving the maintainability, allowing component-level optimization, and reducing error propagation by ensuring that each subsequent stage operates on progressively cleaner and more structured inputs.

3.1. System overview

The framework consists of four sequential stages as illustrated in Fig. 1: (1) document segmentation; (2) perspective rectification; (3) text recognition; (4) structured field extraction. Given an input image acquired in a real-world setting, the first stage localizes the document and predicts its segmentation mask. The second stage uses the mask geometry to estimate document boundaries and rectify the document into a canonical frontal view. The third stage applies OCR to the rectified image in order to obtain textual content and associated detection confidences. Finally, the recognized text is processed by a rule-based parser that extracts structured fields such as document number, holder name, date of birth, expiration date, and nationality. Raw OCR applied directly to unconstrained camera images is highly sensitive to perspective distortion, cluttered backgrounds, and non-uniform lighting. The pipeline reduces irrelevant visual variation before text recognition by prior document region isolation and rectification. The final parsing layer then exploits the semi-structured nature of travel documents, especially the regular formatting of identity fields and MRZ content, and converts OCR outputs into reliable machine-readable fields. In this

way, the proposed framework combines the strengths of learning-based visual perception with the precision and interpretability of deterministic post-processing and leverages the strengths of each component, thus improving robustness and interpretability.

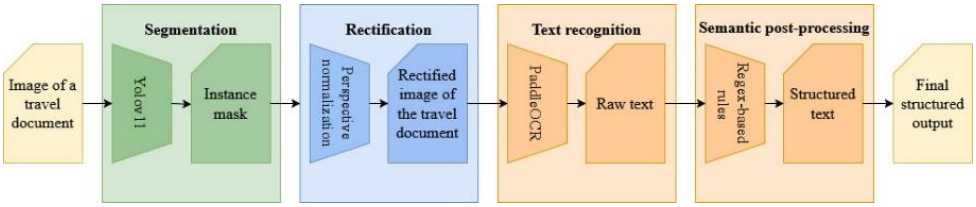

Fig. 1. Processing stages of the proposed modular privacy-preserving framework for travel document recognition

3.2. Document segmentation

The first stage of the pipeline performs document localization using the YOLOv11-seg architecture that combines hierarchical feature extraction, multi-scale feature aggregation, and separate prediction heads for detection and segmentation:

- CSP-Darknet backbone extracts high-level hierarchical visual features from the input image [31];
- PAN + FPN neck aggregates the features across scales to improve robustness to size and viewpoint variation, thus enhancing detection and mask quality;
- detection head predicts bounding boxes and class confidences;
- segmentation head generates prototype masks and per-instance mask coefficients to reconstruct the final document mask.

This segmentation-driven approach is superior to simple object detection for the travel document recognition problem as it provides a more accurate representation of the document outline in the presence of rotations and perspective skew. Precise boundary estimation is especially important in travel document analysis as the quality of the document's outline directly affects the success of the later rectification step. The model predicts a bounding box and class confidence as well as an instance mask corresponding to the visible document region. Thus, the output of the segmentation stage serves both as a document localization result as well as a geometric basis for the next rectification stage.

The presented framework uses the medium variant of YOLOv11-seg, selected based on its trade-off between inference speed and segmentation quality for real-time deployment. Larger model variants are also considered, but the additional computational cost did not provide a sufficiently strong practical benefit for the intended deployment scenario.

3.3. Document rectification

OCR performance is strongly affected by image geometric inconsistency, and even high-quality OCR models may degrade when text lines are slanted, warped, or unevenly scaled due to perspective distortion. The rectification stage of the pipeline standardizes the document layout before the text recognition stage in order to ensure that textual regions occupy more predictable positions and orientations.

The predicted document mask at the output of the segmentation stage is transformed into a geometrically normalized representation. The goal of the document rectification stage is to remove perspective distortion and to produce a front-facing rectangular image of the travel document that is better suited for the text recognition stage. The mask is first converted into a polygonal contour, a convex hull is computed, and the resulting shape is simplified to a quadrilateral approximation that provides the coordinates required for the geometric rectification. The corner coordinates define a homography to map the document from its observed projective view in the input image to a fixed rectangular target plane. A perspective warp standardizes the document layout and eliminates perspective distortion. The rectified image provides a consistent frontal view that is an optimal input for the text recognition stage.

The perspective rectification acts as a normalization layer between visual localization and text recognition. It reduces intra-class variation caused by capture conditions and enables the OCR system to focus on character recognition rather than geometric compensation. This stage is therefore a central part of the proposed framework rather than a minor preprocessing step.

3.4. Text recognition

Once the document is rectified, the text recognition is performed using PaddleOCR, specifically a PP-OCRv3-based configuration that integrates text detection and recognition in a single framework. The OCR stage takes the normalized document image as input and produces detected text lines together with bounding boxes and confidence scores. The model is configured to operate on Latin-script travel documents and is tuned for high recall in unconstrained conditions, including variations in font style, illumination, and image quality. Travel documents contain multiple structured fields that may be short, densely printed, or affected by local degradations in image quality. A recall-oriented configuration allows candidate text to be preserved for later validation and parsing rather than potentially useful information to be discarded too early in the pipeline.

The choice of PaddleOCR is consistent with the design of the framework targeted to real-time or near-real-time utilization in mobile and embedded identity verification settings. PP-OCR is developed as a lightweight OCR system intended for practical deployment with a focus on balancing the recognition performance and the computational efficiency. The text recognition is not considered as an isolated end solution within the proposed pipeline. Its performance depends on the recognizer itself as well as on the quality of the preceding segmentation and rectification stages. Therefore, the OCR subsystem benefits directly from the geometric and visual cleanup performed earlier in the pipeline.

3.5. Semantic post-processing

At the semantic post-processing stage, the raw OCR output is converted into structured travel document data through a regular expression-based parsing module. This stage is responsible for mapping the recognized text fragments to semantically meaningful fields such as document number, holder name, date of birth, expiration

date, and nationality. The extraction procedure is based on regular expressions, positional patterns, and format validation rules derived from the expected structure of travel documents, thus allowing the system to identify valid field candidates, reject implausible text fragments, and standardize recognized values into consistent machine-readable representations. Examples include normalization of dates into a common YYYY-MM-DD format and cleanup of personal names by removing extraneous symbols or OCR artifacts.

The semantic post-processing stage is especially important for MRZ and identity-related fields, where the document format imposes strong syntactic regularities. Rather than relying entirely on the raw text sequence, the parser uses domain knowledge to improve the reliability of the final outputs. It also mitigates common OCR failure modes such as character confusions, fragmented detections, or partial field matches. As a result, the final system output is a structured representation of the document's key identity attributes, thus improving both usability and interpretability in the document verification workflows.

3.6. Design rationale and framework properties

The proposed framework is intentionally modular and separates segmentation, rectification, OCR, and field extraction into distinct stages, thus leveraging the strengths of each component. Each stage performs a distinct subtask and passes a more refined representation to the next stage. YOLOv11-seg model isolates the document and focuses on the difficult visual localization task. The perspective warp rectification standardizes image geometry and ensures geometric consistency. PaddleOCR handles text recognition under variable imaging conditions. Regex-based parsing provides precise and interpretable structured outputs. The modular design has several practical advantages:

- improves interpretability because intermediate outputs such as masks, rectified documents, OCR detections, and parsed fields can be inspected independently;
- improves maintainability because individual components can be updated or replaced without redesigning the entire system;
- supports deployment efficiency and enables reliable real-time extraction of travel document data under unconstrained conditions because the pipeline combines a fast segmentation model with lightweight OCR and low-cost rule-based post-processing.

The framework is also aligned with the constraints of privacy-sensitive document processing. Since real travel documents contain personally identifiable information, the system is designed to minimize dependence on large volumes of real annotated data. The segmentation component is trained on synthetic data, while the next stages exploit geometric normalization and document structure to generalize across real capture conditions. This makes the framework especially suitable for environments where data access is restricted but robust field extraction is still required, providing a deployment-oriented strategy for privacy-preserving travel document analysis.

4. Experimental evaluation

The evaluation of the proposed travel document analysis pipeline assesses both the performance of the document segmentation component and the overall effectiveness of the end-to-end system, including perspective rectification, OCR, and structured field extraction. As the target application involves privacy-sensitive travel documents captured under unconstrained conditions, the experimental design combines controlled evaluation on synthetic data with practical validation on a smaller set of real document images. This two-level evaluation protocol measures segmentation quality under controlled conditions and assesses the proposed framework's generalization to real-world inputs.

4.1. Training datasets

Due to the sensitive nature of the travel documents, no real personal data is used for training the segmentation model. A synthetic dataset generated through a Blender-based rendering pipeline is used in order to enable privacy-preserving model development and to provide accurate pixel-level annotations for the document region. The synthetic images are created by programmatically rendering multiple travel document templates under varied acquisition conditions that include changes in illumination, camera viewpoint, motion blur, and background composition. Polygon masks corresponding to the document instance are generated automatically during the image rendering in order to derive precise ground-truth annotations for instance segmentation.

The final synthetic dataset consists of 2500 images split into training and validation subsets using a 90/10 ratio that is intentionally selected to maximize the amount of training data, given the relatively limited dataset size. Since the model is data-driven and benefits significantly from additional training samples, prioritizing training data is considered beneficial for the overall model performance. To ensure that the smaller validation set does not bias the evaluation, the validation subset is balanced and representative of the dataset distribution, and the reported metrics are stable across multiple training runs.

The synthetic dataset is used to train and validate the YOLOv11-seg model responsible for document localization. The utilization of synthetic data is particularly appropriate for the travel document analysis problem as it enables systematic control over visual variability and avoids the collection, storage, and annotation of sensitive personal identity data. A sample synthetic image and the corresponding instance mask are shown in Fig. 2.

The segmentation model relies exclusively on synthetic data to learn robust features required for document segmentation and text recognition. It ensures full compliance with privacy regulations and allows the system to generalize effectively across variations in perspective, illumination, and background clutter that are challenging to capture with real data.

In order to assess the practical generalization beyond the synthetic domain, the complete pipeline is additionally evaluated on a small set of 100 real travel document images captured under unconstrained conditions. These image samples are not used during the training of the segmentation model and serve only for framework

evaluation. The real images include imperfections typically encountered in deployment scenarios, such as reflections, shadows, mild occlusions, background clutter, and variable capture angles.

(a) (b)

Fig. 2. Sample images from the dataset: synthetic image (a); instance mask (b)

4.2. Training strategy

The document segmentation model is initialized from weights pretrained on the COCO segmentation dataset in order to leverage transferable low-level and mid-level visual features and to improve convergence during training. The model is trained using the AdamW optimizer with an initial learning rate of 0.01 and a cosine decay learning-rate schedule. Input images are resized to 640×640 pixels, the batch size is set to 16, and training is performed for 250 epochs [32]. These settings are selected to provide sufficient optimization time for the model to adapt from general segmentation features to the specific visual characteristics of the synthetic travel document images.

The loss function is a composite YOLOv11 objective that combines bounding box regression, classification, distribution focal loss, and segmentation mask losses. Extensive augmentation utilizing mosaic and mixup augmentation, random HSV color jitter, horizontal and vertical flipping, scaling, and perspective transformations is also applied during the segmentation model training in order to improve its generalization. These augmentations are intended to simulate the visual variability of the real camera-captured document images, such as rotation, lighting changes, and partial occlusion, and to reduce overfitting to the synthetic rendering distribution.

4.3. Experimental setup and evaluation protocol

The evaluation protocol is designed to examine the pipeline at two complementary levels: document segmentation performance is measured on the synthetic validation set using standard detection and segmentation metrics, and the complete end-to-end pipeline is assessed by measuring text and field extraction accuracy after the rectification stage. For the real document images, the evaluation focuses primarily on MRZ recognition since the MRZ provides a well-defined and standardized textual target against which OCR outputs can be compared reliably. The evaluation protocol reflects the framework architecture itself: the segmentation quality determines the quality of the rectification, the rectification influences OCR reliability, and the text recognition performance affects the final field extraction stage.

The performance of the proposed pipeline is evaluated on the two datasets: the synthetic test set generated using the Blender-based rendering pipeline and a small set of 100 real travel document images. The synthetic test set allows controlled evaluation of segmentation and OCR performance under systematically varied conditions, including perspective distortions, illumination changes, and background complexity. The real documents captured in unconstrained conditions are used to assess the practical applicability of the pipeline and to benchmark field extraction accuracy against ground-truth MRZ data.

All experiments are conducted on a high-performance system equipped with NVIDIA L4 GPUs with 24 GB of video memory that is used for model training and evaluation of the suggested pipeline and does not represent the target deployment platform. The YOLOv11-seg model is implemented using the Ultralytics framework with a PyTorch backend. The OCR inference is performed using the official Paddle library implementation.

The segmentation performance on the synthetic validation dataset is quantified using standard metrics, precision and recall, as well as mean Average Precision (mAP). These metrics are computed over the predicted document masks and provide a standard assessment of both detection quality and mask overlap quality. Strong segmentation performance is essential as a prerequisite for successful processing at the next stage, as the segmented document contour is used directly at the perspective rectification stage.

For end-to-end evaluation, the rectified images produced by the segmentation and the perspective correction stages are processed by PaddleOCR, and the extracted fields are compared against ground-truth annotations. The evaluation metrics for the synthetic dataset are character-level OCR accuracy and MRZ field extraction accuracy. For the real image dataset, the evaluation focuses on MRZ extraction accuracy by comparing the recognized MRZ lines against the printed ground-truth MRZ. Inference speed is also measured in order to evaluate the suitability of the pipeline for real-time deployment on mobile and embedded platforms.

The evaluation setup provides a comprehensive assessment of the pipeline's robustness, accuracy, and practical utilization in real-world travel document recognition tasks by combining evaluation on both synthetic and real documents.

5. Results and discussion

5.1. Segmentation performance on synthetic data

The document segmentation results indicate that the proposed YOLOv11-seg model performs extremely well on previously unseen synthetic validation images. Table 1 summarizes the evaluation results, and Fig. 3 illustrates sample segmentation masks obtained from the trained YOLOv11-seg model. The model achieves training and validation precision and recall values close to 100% with validation mAP@50 of 99.50% and mAP@50–95 of 99.22%. The results show that the model learns highly accurate document masks from the synthetic training distribution and can generalize effectively within the synthetic domain. As can be seen in Fig. 3, the qualitative

examples shown further suggest that the predicted masks are sufficiently precise for subsequent corner estimation and perspective correction.

Table 1. Training and validation metrics for YOLO segmentation

Split	Precision	Recall	mAP50	mAP50-95
Training	99.972%	99.671%	99.67%	99.50%
Validation	99.972%	99.671%	99.50%	99.22%

Fig. 3. Document detection and mask segmentation results for a sample synthetic image from the validation dataset

The high values of the evaluation metrics for the document segmentation on the synthetic data are particularly important because even small segmentation errors near document boundaries can lead to inaccurate homography estimation at the rectification stage. The very high mask accuracy, therefore, supports the use of deep learning-based instance segmentation as the front-end stage of the proposed framework rather than coarse bounding-box localization. The segmentation model detects the localization of the document in the image and provides the geometric fidelity required for reliable normalization of the document plane before the text recognition stage is applied.

5.2. End-to-end OCR and field extraction on synthetic data

The evaluation of the end-to-end OCR stage demonstrates that the complete pipeline can successfully provide robust text recognition using PaddleOCR on the rectified images produced from the segmentation masks and the structured extraction performed using regular expression parsing. The experimental results on the synthetic dataset achieve character-level OCR accuracy above 90% and MRZ field extraction accuracy of 90%. These results indicate that the combination of accurate segmentation, perspective normalization, and text recognition is sufficient to recover structured travel document information from the rendered images that include diverse imaging variations introduced through domain randomization. Fig. 4 provides qualitative examples of the document segmentation, the rectification, and the extracted text lines under various synthetic conditions.

(a)

(b)

REPUBLICADEGUINEA ECUATORIAL
PISAPORIE-PASPOR!
P
GNQ
P99990093
ISABEL
ONDO MANGUE
ECUATOGUINEANA
11ABR/AVR84
ANALISTAINFORMATICO
F
ASONGA-BATA
000119983
24 JUN/JUIN 18
MALABO
bd
N
23JUN/JUIN23

(c)

P<GNQISABEL<<ONDO<MANGUE<<<<<<<<<
<<<<<<<<
P999900931GNQ8404110F2306234000119983<<<
90

(d)

(a)

(b)

SPECIMEN
15.072021Luxembourg
3Hu
000100034
Taina Bofferding
Mnistre de finteneur

(c)

IDLUXSPEC073476210715<<<<<<<<<
8308193F3107151LUX<<<<<<<<<<6
SPECIMEN<<JEANNE<<<<<<<<<<<<<

(d)

(a)

(b)

SPECIMEN
16.09.2013
MPB/MUP/MIA
0123456789
N
Olal
ID0000000
PRISTINA

(c)

IDRKSID00000005<<<<<<<<<<<
8906120F2309154RKS01234567891
HOXHA<<MERITA<<<<<<<<<<<<<<<<<

(d)

Fig. 4. Results on samples from the synthetic dataset for the full processing pipeline: source synthetic image (a); rectified travel document (b); recognized text (c); MRZ lines (d)

5.3. Generalization to real travel document images

The most important practical test of the proposed framework is its behavior on real images since the system is intended for deployment on real travel documents rather than synthetic renderings. Evaluation focuses on MRZ extraction accuracy by comparing recognized MRZ lines to the ground-truth printed MRZ on a set of 100 real travel document images. The recognition accuracy is 88% and demonstrates that the trained segmentation model generalizes effectively from synthetic training data to real-world inputs. The result is notable given that the segmentation model is trained without any real data, and the real evaluation images are captured under unconstrained conditions.

The observed performance also provides evidence that the synthetic rendering pipeline captures relevant geometric and visual variation to support transfer to real images. At the same time, the remaining error rate highlights the persistent domain gap between synthetic and real imagery. Most errors on real images are caused by reflections, shadows, and slight occlusions. These failure modes are consistent with the known vulnerabilities of the OCR-based systems and suggest that the main limitations of the current framework arise from the adverse image quality conditions that degrade text readability even after correct document segmentation and rectification.

5.4. Runtime performance and deployment considerations

The experimental results for the evaluation of the suggested modular privacy-preserving framework for travel document segmentation and information extraction from images captured under unconstrained conditions support three main conclusions:

- YOLOv11-seg model achieves very high-quality segmentation on synthetic data and provides instance masks that are accurate enough to reliably support the perspective rectification.
- the complete pipeline delivers robust OCR and MRZ extraction results on synthetic images that reveal an internally coherent and effective modular sequence of the segmentation, rectification, OCR, and structured parsing stages;
- the system generalizes well to real travel document images, achieving 88% MRZ extraction accuracy despite the training only using synthetic data at the segmentation stage.

The proposed framework is intended for inference-time deployment in production environments on mobile and embedded devices where the trained models are executed using optimized runtime configurations. That's why, in addition to the recognition accuracy, the framework is deployed in a production environment for automated travel document verification on various mobile devices and is evaluated in terms of inference speed to determine its suitability for utilization in real-time or near-real-time environments. The measured inference speed indicates that the framework is compatible with deployment-oriented workflows, including mobile-assisted and embedded identity-verification scenarios: the document segmentation and perspective rectification together require ~40 ms, while PaddleOCR and regex-based extraction run for an additional ~140 ms, corresponding to a total processing

time of 180 ms per 1 image. The experimental results indicate that the framework is compatible with the requirements for interactive identity-verification workflows in which a user expects prompt feedback after capturing a document image. Table 2 summarizes end-to-end recognition performance on synthetic and real data and the inference runtime of the pipeline stages. The measured runtime demonstrates that the modular architecture is suitable for near-real-time document processing workflows under deployment-oriented conditions.

Table 2. Quantitative summary of the proposed pipeline on synthetic and real travel document images

Evaluation setting	Metric	Reported result
Synthetic test set	Character-level OCR accuracy	> 90%
Synthetic test set	MRZ field extraction accuracy	90%
Real travel document set	MRZ extraction accuracy	~88%
Inference runtime	Segmentation + perspective rectification	~40 ms
Inference runtime	OCR + regex extraction	~140 ms
Inference runtime	Total pipeline time per image	~180 ms

The reported runtime also supports one of the main design motivations of the framework: a modular pipeline can remain computationally efficient and ensure consistent performance under variations in document orientation, perspective distortion, lighting, and background clutter while still providing interpretable intermediate outputs. Rather than using a heavier monolithic architecture, the system distributes the task across several efficient, specialized components. Thus, the framework is capable of meeting the requirements for operational deployment, especially in mobile-assisted or embedded identity verification scenarios where latency and maintainability are both important.

The modularity of the framework also provides an important practical advantage since for the document recognition systems deployed in operational settings is usually appropriate to inspect intermediate outputs, replace components, or adapt the stages without redesigning the entire architecture. The experimental results show that the modular design that combines learned document localization with deterministic geometric and semantic post-processing is a productive compromise between robustness, interpretability, and computational efficiency. The pipeline achieves practical privacy-preserving automated travel document recognition suitable for production systems.

The remaining limitations of the proposed framework concern failure cases under difficult imaging conditions. Extreme lighting artifacts, strong reflections, and partial document occlusions continue to affect the text recognition confidence and, therefore, structured extraction accuracy. Future improvements may be obtained through additional OCR refinement as well as through better illumination handling, reflection suppression, confidence-aware post-processing, or targeted adaptation using limited real-world samples.

6. Conclusion

The paper presents a modular framework for automated travel document analysis that integrates high-precision document segmentation, geometric rectification, optical

character recognition, and rule-based structured field extraction to achieve robust and interpretable results. The framework is designed to address the practical challenges of travel document recognition under unconstrained acquisition conditions where mobile captured images frequently exhibit perspective distortion, blur, illumination variation, and background clutter. The framework separates the task into interpretable stages and thus enables robust document localization, normalized visual representation, and reliable extraction of structured identity information.

The document segmentation model is trained exclusively on a Blender-generated synthetic dataset, thus avoiding direct reliance on sensitive real personal data during the model development. The framework evaluation achieves very high segmentation performance on synthetic test data and generalizes successfully to real travel document images with MRZ extraction accuracy of 88%. The experimental results highlight the effectiveness of the staged design of the framework. The framework provides high recognition quality and maintains efficient runtime characteristics, supports real-time deployment scenarios, and offers a favorable balance between accuracy, interpretability, privacy preservation, and computational efficiency.

At the same time, the observed failure cases, mainly associated with reflections, shadows, and slight occlusions, further indicate that the OCR accuracy under different imaging conditions remains an important challenge. Future work will therefore focus on the improvement of the text recognition robustness through stronger image enhancement, confidence-aware post-processing, and targeted handling of difficult illumination conditions. The framework could also be extended to a wider variety of document classes, including travel document types with different layouts and scripts. Limited-domain adaptation with carefully controlled real-world samples could be explored to reduce the residual gap between synthetic and real imagery while maintaining privacy constraints. Future research will also include stronger baseline comparisons and larger-scale real-world validation to further establish the generality and the operational value of the privacy-preserving synthetic-data-driven document recognition pipeline.

***Acknowledgments*:** The research work presented in the paper is funded by the European Union-NextGenerationEU via the National Recovery and Resilience Plan of the Republic of Bulgaria under project BG-RRP-2.004-0005 "Improving the research capacity and quality to achieve international recognition and reSilience of TU-Sofia (IDEAS)". The research that led to these results was carried out using the infrastructure purchased under the National Roadmap for RI, financially coordinated by the MES of the Republic of Bulgaria (Grant No DO1-98/26.06.2025).

References

1. Machine Readable Travel Documents. Doc Series. Doc 930. International Civil Aviation Organization.
https://www.icao.int/publications/doc-series/doc-9303
2. Liu, Y., H. Joren, O. Gupta, D. Raviv. MRZ Code Extraction from Visa and Passport Documents Using Convolutional Neural Networks. – International Journal on Document Analysis and Recognition, Vol. **25**, 2022, No 1, pp. 29-39.

3. A b d a l l a h, A., D. E b e r h a r t e r, Z. P f i s t e r, A. J a t o w t. A Survey of Recent Approaches to Form Understanding in Scanned Documents. – Artificial Intelligence Review, Vol. **57**, 2024, No 12.
4. G h a i, D., S. S a x e n a, G. D h i n g r a, S. L. T r i p a t h i. A Comprehensive Review on Performance-Based Comparative Analysis, Categorization, Classification, and Mapping of Text Extraction System Techniques for Images. – Multimedia Tools and Applications, Vol. **84**, 2025, No 5, pp. 2327-2484.
5. B o n e d, C., et al. Synthetic Dataset of ID and Travel Documents. – Scientific Data, Vol. **11**, 2024, No 1.
6. M a n, K., J. C h a h l. A Review of Synthetic Image Data and Its Use in Computer Vision. – Journal of Imaging, Vol. **8**, 2022, No 11.
7. M u m u n i, A., F. M u m u n i, N. K. G e r r a r. A Survey of Synthetic Data Augmentation Methods in Machine Vision. – Machine Intelligence Research, Vol. **21**, 2024, No 5, 831-869.
8. K l i n k, S., T. K i e n i n g e r. Rule-Based Document Structure Understanding with a Fuzzy Combination of Layout and Textual Features. – International Journal on Document Analysis and Recognition Article, Vol. **4**, 2001, pp. 18-26.
9. V e e r a s e k a r, P. A., M. M. R. S i n d h a, U. M. P a n d i y a n, V. V i j a y a r a j a. A Transfer Learning Approach for College ID Card Detection. – In: Proc. of AIP Conf., Vol. **3258**, 2025, No 1.
10. G a y e r, A. V., Y. S. C h e r n y s h o v a, V. V. A r l a z a r o v. Recognition of Machine-Readable Zone in Identity Documents: A Review. – IEEE Access, 2025.
11. B u l a t o v, K. B., P. V. B e z m a t e r n y k h, D. P. N i k o l a e v, V. V. A r l a z a r o v. Towards a Unified Framework for Identity Documents Analysis and Recognition. – Computer Optics, Vol. **46**, 2022, No 3, pp. 436-454.
12. X u, X., et. al. Crack Detection and Comparison Study Based on Faster R-CNN and Mask R-CNN. – Sensors, Vol. **22**, 2022.
13. B h a r a t i, P., A. P r a m a n i k. Deep Learning Techniques – R-CNN to Mask R-CNN: A Survey. – In: A. Das, J. Nayak, B. Naik, S. Pati, D. Pelusi, Eds. Computational Intelligence in Pattern Recognition. Advances in Intelligent Systems and Computing. Vol. **999**. 2020.
14. K i m, G., et al. OCR-Free Document Understanding Transformer. – In: S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner, Eds. Computer Vision – ECCV, Lecture Notes in Computer Science. 2022. Cham, Springer, pp. 498-517, 13688.
15. P a d m a j a, D. L., S. N. P r a s a d, G. S. H a r s h a, G. S. D e e p a k. A Comparative Analysis of Tesseract OCR. – In: Proc. of International Conference on Technology Advances for Green Solutions and Sustainable Development: ICT4GS-2024, Springer, Nature, 2025.
16. D u, Y., et. al. PP-OCR: A Practical Ultra-Lightweight OCR System. – arXiv Preprint, 2020, arXiv:2009.09941.
17. C u i, C., et. al. PaddleOCR 3.0 Technical Report. – arXiv Preprint, 2025, arXiv:2507.05595.
18. C h a m p a, H. N. Scene Text Detection and Recognition Using EasyOCR: Performance Analysis and Evaluation. – In: Proc. of 2nd IEEE Asian Conference On Intelligent Technologies (ACOIT'25), 2025.
19. R a j a S h r e e, S., C. N u t h a n a k a l u v, L. S h a r m i l a, J. S. D u e l a, N. M u n e e r a. Extraction of Data from Documents Using AWS Textract. – In: Proc. of AIP Conference, Vol. **3257**, 2025, No 1.
20. R a m a c h a n d r u l a, P., J. K u n i s e t t y, S. S u b r a m a n i a n, S. P a l a n i s w a m y, P. B. P a t i. A Novel System for Authentication of Identity Cards Using YOLOv5 and YOLOv8. – In: Advances in Data Science and Artificial Intelligence, Lecture Notes in Electrical Engineering. Vol. **1399**. 2025, Singapore, Springer.
21. R a n j a n, S., K. M a n o j. Comparing YOLOv11 and YOLOv8 for Instance Segmentation of Occluded and Non-Occluded Immature Green Fruits in a Complex Orchard Environment. – arXiv Preprint, 2025, arXiv:2410.19869.
22. K h a n a m, R., M. H u s s a i n. YOLOv11: An Overview of the Key Architectural Enhancements. – arXiv Preprint, 2024, arXiv:2410.17725.
23. M a t a l o v, D., V. A r l a z a r o v. Model-Driven Approach to Creating ID Document Templates for Localization and Classification Based on a Single Image – Computer, Vol. **49**, 2026, No 6.

24. A r l a z a r o v, V., et. al. Mobile ID Recognition: Coarse-to-Fine Approach. – In: Image Analysis and Pattern Recognition: State of the Art in the Russian Federation, 2025, pp. 827-914.
25. S k o r y u k i n a, N., D. T r o p i n, J. S h e m i a k i n a, V. A r l a z a r o v. Document Localization and Classification as Stages of a Document Recognition System. – Pattern Recognition and Image Analysis, Vol. **33**, 2023, No 4, pp. 699-716.
26. F a n, G. DocPINN: A Neural PDE-Based Framework for Document Image Dewarping. – In: Proc. of International Conference on Document Analysis and Recognition, Cham, Springer Nature Switzerland, 2025, pp. 382-397.
27. A r l a z a r o v, V. V., K. B u l a t o v, T. C h e r n o v, V. L. A r l a z a r o v. MIDV-500: A Dataset for Identity Document Analysis and Recognition on Mobile Devices in a Video Stream. – Computer Optics, Vol. **43**, 2019, No 5, pp. 818-824.
28. B u l a t o v, K., et. al. MIDV-2020: A Comprehensive Benchmark Dataset for Identity Document Analysis. – Computer Optics, Vol. **46**, 2022, No 2, pp. 252-270. DOI: 10.18287/2412-6179-CO-1006.
29. F e r n a n d e s, L., et. al. BRIDP: Dataset and Validation Method for Brazilian Identity Document Parsing. – In: Proc. of International Conference on Intelligent Systems Design and Applications, Cham, Switzerland, Springer Nature, 2023, pp. 445-454.
30. Z e n g, Q., J. E.T a p i a, I.G a r c í a, J. M. E s p í n, C. B u s c h. ID-Card Synthetic Generation: Toward a Simulated Bona Fide Dataset. – arXiv Preprint, 2025, arXiv:2508.13078.
31. Ultralytics YOLO.
https://www.ultralytics.com
32. M e n g, L. K., H. H. Y i, N. B. W e i, L. J. X i n, Z. A. A b d u l S a l a m. A Machine Learning Approach for a Face Mask Detection System with AdamW Optimizer. – Journal of Advanced Trends in Information Technology, Vol. **7**, 2024, No 1.

Fast-track. Received: 01.04.2026, Revised version: 03.05.2026, Accepted: 10.05.2026