

# TransDocAnalyser: A Framework for Offline Semi-structured Handwritten Document Analysis in the Legal Domain

Sagar Chakraborty<sup>1,2</sup>, Gaurav Harit<sup>2</sup>, and Saptarshi Ghosh<sup>3</sup>

<sup>1</sup> Wipro Limited, Salt Lake, Kolkata, India

<sup>2</sup> Department of Computer Science and Engineering, Indian Institute of Technology, Jodhpur, Rajasthan, India

chakraborty.4@iitj.ac.in, gharit@iitj.ac.in

<sup>3</sup> Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

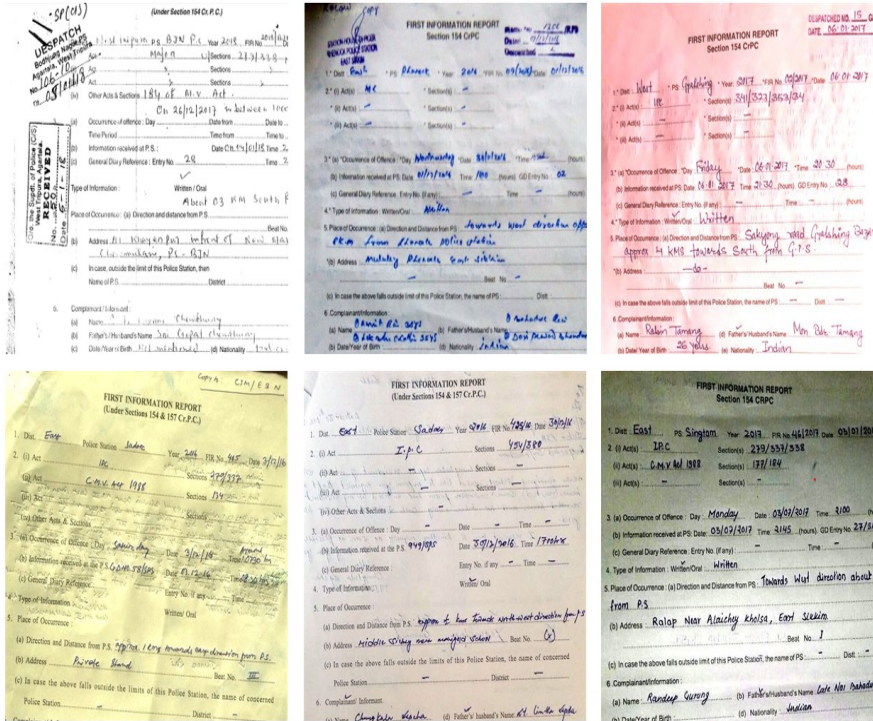
saptarshi@cse.iitkgp.ac.in

**Abstract.** State-of-the-art offline Optical Character Recognition (OCR) frameworks perform poorly on semi-structured handwritten domain-specific documents due to their inability to localize and label form fields with domain-specific semantics. Existing techniques for semi-structured document analysis have primarily used datasets comprising invoices, purchase orders, receipts, and identity-card documents for benchmarking. In this work, we build the first semi-structured document analysis dataset in the legal domain by collecting a large number of First Information Report (FIR) documents from several police stations in India. This dataset, which we call the FIR dataset, is more challenging than most existing document analysis datasets, since it combines a wide variety of handwritten text with printed text. We also propose an end-to-end framework for offline processing of handwritten semi-structured documents, and benchmark it on our novel FIR dataset. Our framework used Encoder-Decoder architecture for localizing and labelling the form fields and for recognizing the handwritten content. The encoder consists of Faster-RCNN and Vision Transformers. Further the Transformer-based decoder architecture is trained with a domain-specific tokenizer. We also propose a post-correction method to handle recognition errors pertaining to the domain-specific terms. Our proposed framework achieves state-of-the-art results on the FIR dataset outperforming several existing models.

**Keywords:** Semi-structured document · Offline handwriting recognition · Legal document analysis · Vision Transformer · FIR dataset

## 1 Introduction

Semi-Structured documents are widely used in many different industries. Recent advancement in digitization has increased the demand for analysis of scanned or mobile-captured semi-structured documents. Many recent works have used different deep learning techniques to solve some of the critical problems in processing and layout analysis of semi-structured documents [34,16,23]. Semi-structured



**Fig. 1.** Examples of First Information Report (FIR) documents from different police stations in India. The FIR dataset developed in this paper consists of a wide variety of such semi-structured FIR documents containing both printed and handwritten text.

documents consist of printed, handwritten, or hybrid (both printed and handwritten) text forms. In particular, hybrid documents (see Figure 1) are more complex to analyze since they require segregation of printed and handwritten text and subsequent recognition. With recent advancements, the OCR accuracy has improved for printed text; however, recognition of handwritten characters is still a challenge due to variations in writing style and layout.

Earlier works have focused on techniques for layout analysis, named-entity recognition, offline handwriting recognition, etc., but sufficient work has *not* been done on developing an end-to-end framework for processing semi-structured documents. A general end-to-end framework can be easily fine-tuned for domain-specific requirements. In this paper we present the first framework for semi-structured document analysis applied to legal documents.

There have been many works on legal documents, such as on case document summarization [6], relevant statute identification from legal facts [31], pretraining language models on legal text [32] and so on. But almost all prior research in the legal domain has focused on textual data, and *not* on document images. In

particular, the challenges involved in document processing and layout analysis of legal documents is unattended, even though these tasks have become important due to the increasing availability of scanned/photographed legal documents.

In this work, we build the first dataset for semi-structured document analysis in the legal domain. To this end, we focus on **First Information Report (FIR)** documents from India. An FIR is usually prepared by police stations in some South Asian countries when they first get a complaint by the victim of a crime (or someone on behalf of the victim).<sup>4</sup> An FIR usually contains a lot of details such as the date, time, place, and details of the incident, the names of the person(s) involved, a list of the statutes (written laws, e.g., those set by the Constitution of a country) that might have been violated by the incident, and so on. The FIRs are usually written on a printed form, where the fields are filled in by hand by police officials (see examples in Figure 1). It is estimated that more than 6 million FIRs are filed every year across thousands of police stations in various states in India. Such high volumes lead to inconsistent practices in-terms of handwriting, layout structure, scanning procedure, scan quality, etc., and introduce huge noise in the digital copies of these documents.

Our target fields of interest while processing FIR documents are the handwritten entries (e.g., name of the complainant, the statutes violated) which are challenging to identify due to the wide variation in handwriting. To form the dataset, which we call the **FIR dataset**, we created the meta-data for the target fields by collecting the actual text values from the police databases, and also annotated the documents with layout positions of the target fields. The FIR dataset is made publicly available at [https://github.com/LegalDocumentProcessing/FIR\\_Dataset\\_ICDAR2023](https://github.com/LegalDocumentProcessing/FIR_Dataset_ICDAR2023).

The FIR dataset is particularly challenging since its documents are of mixed type, with both printed and handwritten text. Traditional OCR identifies blocks of text strings in documents and recognizes the text from images by parsing from left to right [19]. NLP techniques like named-entity recognition (NER), which uses raw text to find the target fields, cannot be applied easily, since traditional OCRs do not work well in recognition of mixed documents with handwritten and printed characters occurring together. Another drawback of traditional OCRs in this context is their inability to recognise domain-specific words due to their general language-based vocabulary. In this work, we propose a novel framework for analysing such domain-specific semi-structured documents. The contributions of the proposed framework as follows:

1. We use a FastRCNN + Vision Transformer-based encoder trained for target field localization and classification. We also deploy a BERT-based text decoder that is fine-tuned to incorporate legal domain-specific vocabulary.
2. We use a domain-specific pretrained language model [32] to improve the recognition of domain-specific text (legal statutes, Indian names, etc.). This idea of using a domain-specific language model along with OCR is novel, and has a wider applicability over other domains (e.g., finance, healthcare,

---

<sup>4</sup> [https://en.wikipedia.org/wiki/First\\_information\\_report](https://en.wikipedia.org/wiki/First_information_report)

etc) where this technique can be used to achieve improved recognition from domain-specific documents.

3. We improve the character error rate (CER) by reducing the ambiguities in OCR through a novel domain-specific post-correction step. Using domain knowledge, we created a database for each target field (such as Indian names, Indian statutes, etc.) to replace the ambiguous words from OCR having low confidence using a combination of TF-IDF vectorizer and K-Nearest Neighbour classifier. This novel post-correction method to handle recognition errors pertaining to proper nouns, enables our proposed framework to outperform state-of-the-art OCR models by large margins.

To summarize, in this work we build the first legal domain-specific dataset for semi-structured document analysis. We also develop a framework to localise the handwritten target fields, and fine-tune a transformer-based OCR (TrOCR) to extract handwritten text. We further develop post-correction techniques to improve the character error rate. To our knowledge, the combination of Faster-RCNN and TrOCR with other components, such as Vision Transformer and legal domain-specific tokenizers, to create an end-to-end framework for processing offline handwritten semi-structured documents is novel, and can be useful for analysis of similar documents in other domains as well.

## 2 Related Work

We briefly survey four types of prior works related to our work – (i) related datasets, (ii) works addressing target field localization and classification, (iii) handwritten character recognition, and (iv) works on post-OCR correction methods

**Related Datasets:** There exist several popular datasets for semi-structured document analysis. FUNSD [22] is a very popular dataset for information extraction and layout analysis. FUNSD dataset is a subset of RVL-CDIP dataset [17], and contains 199 annotated financial forms. The SROIE dataset [21] contains 1,000 annotated receipts having 4 different entities, and is used for receipt recognition and information extraction tasks. The CloudSCan Invoice dataset [29] is a custom dataset for invoice information extraction. The dataset contained 8 entities in printed text.

Note that no such dataset exists in the legal domain, and our FIR dataset is the first of its kind. Also, the existing datasets contain only printed text, while the dataset we build contains a mixture of printed and hand-written text (see Table 2 for a detailed comparison of the various datasets).

**Localization and Labelling of field components:** Rule-based information extraction methods (such as the method developed by Kempf et al. [10] and many other methods) could be useful when documents are of high quality and do not contain handwritten characters. But when document layouts involve huge variations, noise and handwritten characters, keyword-based approaches fail to provide good results. Template-based approaches also fail due to scanning errors and layout variability [36,1,2].

Srivastava et al. [12] developed a graph-based deep network for predicting the associations between field labels and field values in handwritten form images. They considered forms in which the field label comprises printed text and field value can be handwritten text; this is similar to what we have in the FIR dataset developed in this work. To perform association between the target field labels and values, they formed a graphical representation of the textual scripts using their associated layout position.

In this work, we tried to remove the dependency on OCR of previous works [12] by using layout information of images to learn the positions of target fields and extract the image patches using state-of-the-art object detection models such as [37,35,33].

Zhu et. al. [37] proposed attention modules that only attend to a small set of key sampling points around a reference, which can achieve better performance than baseline model [8] with  $10\times$  less training epochs. Tan et. al. [35] used weighted bi-directional feature pyramid network (BiFPN), which allows easy and fast multi-scale feature fusion. Ren et al [33] proposed an improved version of their earlier work [14] provides comparative performances with [37,35] with lower latency and computational resources on FIR dataset. Hence, we use Faster RCNN model in this framework for localization and classification of the field component.

**Handwritten Character Recognition:** Offline handwriting recognition has been a long standing research interest. The works [3,4,5] presented novel features based on structural features of the strokes and their spatial relations with a character, as visible from different viewing directions on a 2D plane. Diesendruck et al. [11] used Word Spotting to directly recognise handwritten text from images. The conventional text recognition task is usually framed as an encoder-decoder problem where the traditional methods[19] leveraged CNN-based [24] encoder for image understanding and LSTM-based [20] decoder for text recognition.

Chowdhury et al. [9] combined a deep convolutional network with a recurrent Encoder-Decoder network to map an image to a sequence of characters corresponding to the text present in the image. Michael, Johannes et al. [28] proposed a sequence-to-sequence model combining a convolutional neural network (as a generic feature extractor) with a recurrent neural network to encode both the visual information, as well as the temporal context between characters in the input image. Further, Li et al. [25] used for the first time an end-to-end Transformer-based encoder-decoder OCR model for handwritten text recognition and achieved SOTA results. The model [25] is convolution-free unlike previous methods, and does not rely on any complex pre/post-processing steps. The present work leverages this work and extends its application in legal domain.

**Post-OCR correction:** Rectification of errors in the recognised text from the OCR would require extensive training which is computation heavy. Further, post-OCR error correction requires a large amount of annotated data which may not always be available. After the introduction of the Attention mechanism and BERT model, many works have been done to improve the results of the

OCR using language model based post-correction techniques. However, Neural Machine Translation based approaches as used by Duong et al. [13] are not useful in the case of form text due to the lack of adequate context and neighbouring words. We extend the idea used in the work of Trstenjak et al. [7] where they used edit distance and cosine similarity to find the matching words. In this paper we used K-nearest neighbour with edit distance to find best matches for the words predicted with low confidence score by the OCR.

### 3 The FIR Dataset

First Information Report (FIR) documents contain details about incidents of cognisable offence, that are written at police stations based on a complaint. FIRs are usually filed by a police official filling up a printed form; hence the documents contain both printed and handwritten text. In this work, we focus on FIR documents written at police stations in India. Though the FIR forms used across different Indian states mostly have a common set of fields, there are some differences in their layout (see examples in Fig. 1). To diversify the dataset, we included FIR documents from the databases of various police stations across several Indian states – West Bengal<sup>5</sup>, Rajasthan<sup>6</sup>, Sikkim<sup>7</sup>, Tripura<sup>8</sup> and Nagaland<sup>9</sup>.

As stated earlier, an FIR contains many fields including the name of the complainant, names of suspected/alleged persons, statutes that may have been violated, date and location of the incident, and so on. In this work, we selected *four target fields* from FIR documents for the data annotation and recognition task – (1) *Year* (the year in which the complaint is being recorded), (2) *Complainant’s name* (name of the person who lodged the complaint), (3) *Police Station* (name of the police station that is responsible for investigating the particular incident), and (4) *Statutes* (Indian laws that have potentially been violated in the reported incident; these laws give a good indication of the type of the crime). We selected these four target fields because we were able to collect the gold standard for these four fields from some of the police databases. Also, digitizing these four fields would enable various societal analysis, such as analysis of the nature of crimes in different police stations, temporal variations in crimes, and so on.

**Annotations:** We manually analysed more than 1,300 FIR documents belonging to different states, regions, police stations, etc. We found that FIR documents from the same region / police station tend to have the similar layout and form structure. Hence we selected a subset of 375 FIR documents with reasonably varying layouts / form structure, so that this subset covers most of the different variations. These 375 documents were manually annotated. Annotations were

<sup>5</sup> [http://bidhannagarcitypolice.gov.in/fir\\_record.php](http://bidhannagarcitypolice.gov.in/fir_record.php)

<sup>6</sup> <https://home.rajasthan.gov.in/content/homeportal/en.html>

<sup>7</sup> <https://police.sikkim.gov.in/visitor/fir>

<sup>8</sup> <https://tripurapolice.gov.in/west/fir-copies>

<sup>9</sup> <https://police.nagaland.gov.in/fir-2/>



**Fig. 2.** Sample of various entities present in First Information Reports with different writing styles, distortions and scales.

**Table 1.** FIR Dataset statistics

Split	Images	Layout	Words	Labels
Training	300	61	1,830	1,230
Testing	75	18	457	307

done on these documents using LabelMe annotation tool<sup>10</sup> to mark the bounding boxes of the target fields.

Figure 2 shows some samples of various entities present in our dataset, and Figure 3 shows examples of ground truth annotations for two of the entities in Figure 2. In the ground truth, each bounding box has four co-ordinates (X\_left, X\_width, Y\_right, Y\_height) which describe the position of the rectangle containing the field value for each target field.

**Train-test split:** During the annotation of our dataset, we identified 79 different types of large scale variations, layout distortions/deformations, which we split into training and testing sets. We divided our dataset (of 375 document images) such that 300 images are included in the training set and the other 75 images are used as the test set. During training, we used 30% of training dataset as a validation set. Table 1 shows the bifurcation statistics for training and test sets.

**Preprocessing the images:** For Faster-RCNN we resized the document images to a size of  $1180 \times 740$ , and used the bounding boxes and label names to train the model to predict and classify the bounding boxes. We convert the dataset into IAM Dataset format [27] to fine-tune the transformer OCR.

<sup>10</sup> <https://github.com/wkentaro/labelme>

```

{
  "First Information Report": [
    {
      "id": 0,
      "image_id": 127,
      "text": "Randeep Gurung",
      "box": [
        [
          293.56, 84.76, 357.43, 108.84
        ]
      ],
      "label": "complaint_name"
    },
    {
      "id": 1,
      "image_id": 127,
      "text": "122/269/270",
      "box": [
        [
          469.47, 104.65, 628.63, 130.83
        ]
      ],
      "label": "section"
    }
  ]
}

```

**Fig. 3.** Examples of ground truth annotations for two of the entities shown in Figure 2

**Table 2.** Comparison of the FIR dataset with other similar datasets

Dataset	Category	#Images	Text Type		#Entites
			Printed	Handwritten	
FUNSD [22]	Form	199	✓	x	4
SROIE [21]	Receipt	1000	✓	x	4
Cloud Invoice [29]	Invoice	326571	✓	x	8
FIR (Ours)	Form	375	✓	✓	4

**Novelty of the FIR dataset:** We compare our FIR dataset<sup>11</sup> with other datasets for semi-structure document analysis in Table 2. The FIR dataset contains both printed and handwritten information which makes it unique and complex compared to several other datasets. Additionally, the FIR dataset is the first dataset for semi-structured document analysis in the legal domain.

## 4 The TransDocAnalyser Framework

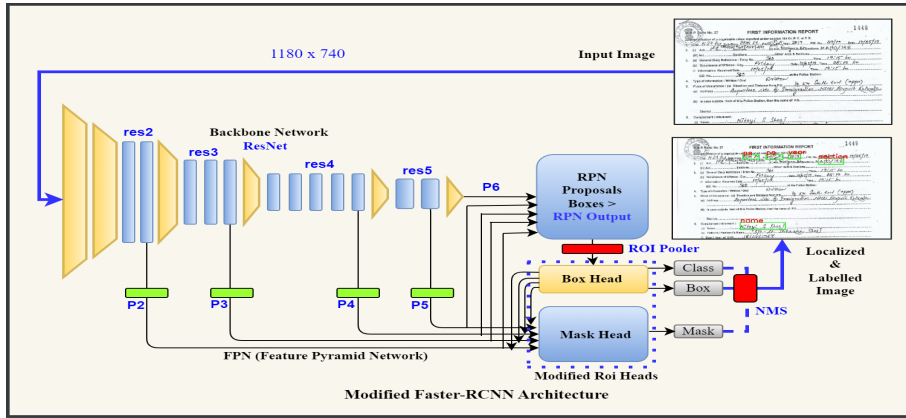
We now present TransDocAnalyser, a framework for offline processing of handwritten semi-structured documents, by adopting Faster-RCNN and Transformer-based encoder-decoder architecture, with post-correction to improve performance.

### 4.1 The Faster-RCNN architecture

Faster-RCNN [33] is a popular object detection algorithm that has been adopted in many real-world applications. It builds upon the earlier R-CNN [15] and Fast R-CNN [33] architectures. We pass the input images through the Faster-RCNN network to get the domain-specific field associations and extract the image patches from the documents.

<sup>11</sup> [https://github.com/LegalDocumentProcessing/FIR\\_Dataset\\_ICDAR2023](https://github.com/LegalDocumentProcessing/FIR_Dataset_ICDAR2023)





**Fig. 4.** Modified Faster-RCNN based architecture for target field localization and labelling

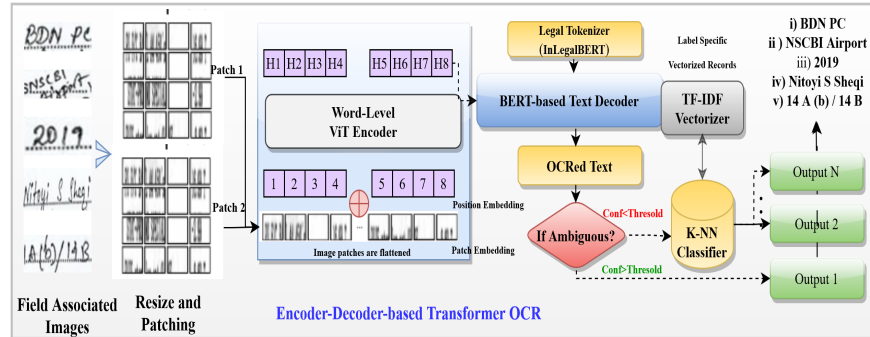
Our modified Faster-RCNN architecture consists of three main components (as schematically shown in Figure 4)– (1) Backbone Network , (2) Region Proposal Network (RPN), and (3) ROI Heads as detailed below.

**(1) Backbone Network:** ResNet-based backbone network is used to extract multi-scaled feature maps from the input – that are named as P2, P3, P4 , P8 and so on – which are scaled as 1/4th, 1/8th, 1/16th and so on. This backbone network is FPN-based (Feature Pyramid network) [26] which is multi-scale object detector invariant to the object size.

**(2) Region Proposal Network (RPN):** Detects ROI (regions of interest) along with a confidence score, from the multi-scale feature maps generated by the backbone network. A fixed-size kernel is used for region pooling. The regions detected by the RPN are called *proposal boxes*.

**(3) ROI Heads:** The input to the box head comprises (i) the feature maps generated by a Fully Connected Network (FCN), (ii) the *proposed boxes* which come from the RPN. These are 1,000 boxes with their predicted labels. Box head uses the bounding boxes proposed by the RPN to crop and prepare the feature maps. (iii) ground truth bounding boxes from the annotated training datasets. The ROI pooling uses the proposed boxes detected by RPN, crops the rectangular areas of the feature maps, and feeds them into the head networks. Using Box head and mask head together in Faster-RCNN network, inspired by He et al. [18] improves the overall performance.

During training, the box head makes use of the ground truth boxes to accelerate the training. The mask head provides the final predicted bounding boxes and confidence scores during the training. At the time of inference the head network uses non-maximum suppression (NMS) algorithm to remove the overlapping boxes and selects the top-k results as the predicted output based on thresholds on their confidence score and intersection over union (IOU).



**Fig. 5.** TrOCR architecture with custom enhancements. The Text decoder uses a domain-specific InLegalBert [32] based tokenizer. OCR predictions go for post-correction if the confidence score is less than the threshold. We convert the OCR prediction into a TF-IDF vector and search in the domain-specific field database to find the Nearest Match.

## 4.2 The TrOCR architecture

Once the localized images are generated for a target field (e.g., complainant name) by Faster-RCNN, the image patches are then flattened and sent to the Vision Transformer (ViT) based encoder model. We use TrOCR [25] as the backbone model for our finetuning (see Figure 5). TrOCR [25] is a Transformer-based OCR model which consists of a pretrained vision Transformer encoder and a pretrained text decoder. The ViT encoder is trained on the IAM handwritten dataset, which we fine-tune on our FIR dataset. We use the output patches from the Faster-RCNN network as input to the ViT encoder, and fine-tune it to generate features. As we are providing the raw image patches received from Faster-RCNN into the ViT encoder, we did not apply any pre-processing or layout enhancement technique to improve the quality of the localised images. On the contrary, we put the noisy localised images cropped from the *form fields* directly, which learns to suppress noise features by training.

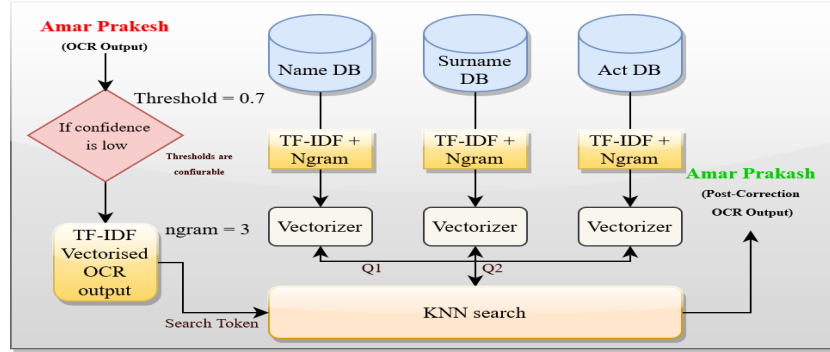
We also replace the default text decoder (RoBERTa) with the Indian legal-domain specific BERT based text decoder InLegalBERT [32] as shown in Fig. 5. InLegalBert [32] is pre-trained with a huge corpus of about 5.4 million Indian Legal documents, including court judgements of the Indian Supreme Court and other higher courts of India, and various Central Government Acts.

To recognize characters in the cropped image patches, the images are first resized into square boxes of size  $384 \times 384$  pixels and then flattened into a sequence of patches, which are then encoded by ViT into high-level representations and decoded by InLegalBERT into corresponding characters step-by-step.

We evaluate and penalise the model based on the Character Error Rate (CER). CER calculation is based on the concept of Levenshtein distance, where we count the minimum number of character-level operations required to transform the ground truth text into the predicted OCR output. CER is computed as  $CER = (S + D + I)/N$  where  $S$  is the number of substitutions,  $D$  is the number

**Table 3.** Excerpts from field-specific databases used to prepare TF-IDF vectorized records for KNN search. All databases contain India-specific entries.

Names	Surnames	Police Stations	Statutes / Acts
Anamul	Haque	Baguiati	IPC (Indian Penal Code)
Shyam	Das	Airport	D.M. Act (Disaster Management Act)
Barnali	Pramanik	Newtown	D.C. Act (Drug and Cosmetics Act)
Rasida	Begam	Saltlake	NDPS Act

**Fig. 6.** Term Frequency and Inverse Document frequency (TF-IDF) Vectorizer based K-Nearest Neighbour model for post-correction on OCR output

of deletions,  $I$  is the number of Insertions, and  $N$  is the number of characters in the reference text.

### 4.3 KNN-based OCR Correction

For each predicted word from OCR, if the confidence score is less than a threshold 0.7, we consider the OCR output to be ambiguous for that particular word. In such cases, the predicted word goes through a post-correction step which we describe now (see Figure 6).

For each target field, we create a database of relevant values and terms (which could be written in the field) from various sources available on the Web. Table 3 shows a very small subset of some of the field-specific databases such as Indian names, Indian surnames, Indian statutes (Acts and Sections), etc. We converted each database into a set of TF-IDF vectors (see Figure 6). Here TF-IDF stands for Term Frequency times Inverse Document Frequency. The TF-IDF scores are computed using n-grams of groups of letters. In our work we used  $n = 3$  (trigrams) for generating the TF-IDF vectors for OCR predicted words as well as for the entities in the databases.

For a given OCR output, based on the associated field name which is already available from the field classification by Faster-RCNN, we used the K-Nearest Neighbour (KNN) classifier to select the appropriate vectorized database. KNN

**Table 4.** Faster-RCNN model training parameters

Base Model	Base Weights	Learning Rate	Epoch #	# of Class	IMS/batch	Image Size
ResNet 50	Mask RCNN	0.00025	2500	4	4	1180 × 740

**Table 5.** Transformer OCR (TrOCR) parameters used for model fine-tuning

Feature Extractor	Tokenizer	Max Len	N-gram	Penalty	# of Beam	Optimizer
google-vit-patch16-384	InLegalBERT	32	3	2.0	4	AdamW

returns best matches with a confidence score based on the distance between the search vector (OCR output) and the vectors in the chosen database. If the confidence score returned by KNN is greater than 0.9, then the OCR predicted word gets replaced with the word predicted by the K-Nearest Neighbour search.

## 5 Experimental settings

We ran all experiments on a Tesla T4 GPU with CUDA version 11.2. We used CUDA enabled Torch framework 1.8.0.

In the first stage of the TransDocAnalyser framework, we trained the Faster RCNN from scratch using the annotated dataset (the training set). Table 4 shows the settings used for training the Faster-RCNN model. Prior to the training, input images are resized in 1180 × 740. For memory optimization, we run the model in two steps, first for 1500 iteration and then for 1000 iteration on the stored model. We tried batch sizes (BS) of 16, 32 and 64, and finalized BS as 64 because of the improvement in performance and training time. We used the trained model Faster-RCNN model to detect and crop out the bounding boxes of each label from the original document (as shown in Fig. 2) and created our dataset to fine-tune the ViT encoder.

We also created a metadata file mapping each cropped image (as shown in Fig. 2) with its corresponding text as described in [27] to fine-tune the decoder.

Table 5 shows the parameter settings used for fine-tuning the TrOCR model. Image patches are resized to 384 × 384 dimension to fine-tune ViT encoder. In the TrOCR model configuration, we replaced the tokenizer and decoder settings based on InLegalBert. We tried with batch size (BS) of 2, 4, 8, 16, 32, 64, and BS = 8 provided the best result on the validation set. We fine-tuned the Encoder and Decoder of the OCR for 40 epochs and obtained the final results.

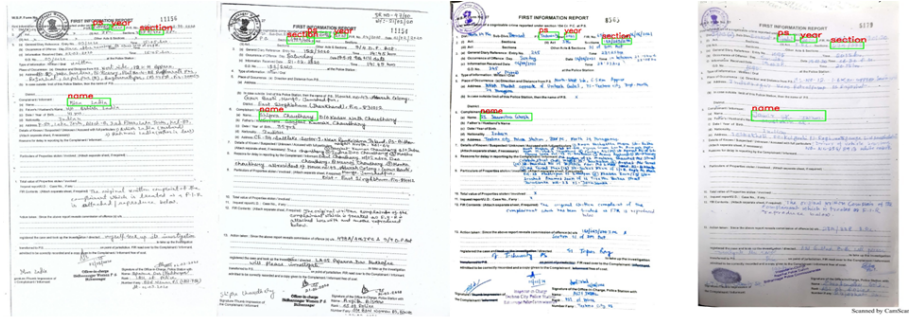
The KNN-based OCR correction module used n-grams with  $n = 1, 2, 3, 4$  to generate the TF-IDF vectors of the field-specific databases. Using  $n = 3$  (trigrams) and KNN with  $K = 1$  provided the best results.

## 6 Results

In this section, we present the results of the proposed framework TransDocAnalyser in three stages – (i) The performance of Faster-RCNN on localization and

**Table 6.** Performance of field labelling on the FIR dataset (validation set and test set). Re: Recall, Pr: Precision, F1: F1-score, mAP: mean average precision.

Results on dataset	Target field	Faster R-CNN			
		Re ↑	Pr ↑	F1 ↑	mAP ↑
Validation	Year	0.98	0.96	0.97	0.97
	Statute	0.85	0.82	0.83	0.84
	Police Station	0.96	0.90	0.93	0.93
	Complainant Name	0.84	0.76	0.80	0.77
Test	Year	0.97	0.96	0.97	0.96
	Statute	0.84	0.87	0.86	0.80
	Police Station	0.93	0.88	0.91	0.91
	Complainant Name	0.80	0.81	0.81	0.74




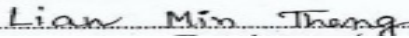



**Fig. 7.** Examples of localization and labelling of target fields by Faster-RCNN. The predicted bounding boxes are highlighted in green on the images. The associated class labels are highlighted in red.

labelling of the target fields (Table 6); (ii) Sample of OCR results with Confidence Scores (Table 7); and (iii) Comparison of the performance of the proposed framework with existing OCR methods (Table 8).

Table 6 shows the results of field label detection using Faster-RCNN on both test and validation sets of the FIR dataset. The performance is reported in terms of Recall (Re), Precision (Pr), F1 (harmonic mean of Recall and Precision) and mean Average Precision (mAP). For the localization and labelling, a prediction is considered correct if both the IOU (with the ground truth) and the confidence threshold are higher than 0.5. The results show that our model is performing well, with the best and worst results for the fields ‘Year’ (F1 = 0.97) and ‘Name’ (F1 = 0.8) respectively. This variation in the results is intuitive, since names have a lot more variation than the year.

Figure 7 shows examples of outputs of Faster-RCNN on some documents from the test set of the FIR dataset. The predicted bounding boxes are highlighted in green rectangles, and the predicted class names are marked in red on top of each bounding box.

**Table 7.** Finetuned (TrOCR) predictions on the generated image patches shown below

Image Patches	OCR Results	Confidence Score
	2019	0.89
	Lian Min Thang	0.77
	Nscbi Airport	0.79
	Amar <b>Prakesh</b>	<b>0.63</b>
	379	0.96

The output of Faster-RCNN provides bounding boxes and field names for each image, using which image patches are generated and sent to the Encoder-Decoder architecture. Table 7 shows some examples of image patches and the finetuned TrOCR predictions for those image patches. It is seen that the name ‘‘Amar Prakash’’ is predicted as ‘Amar Prakesh’’ with confidence score below a threshold of 0.7 (which was decided empirically). As the prediction confidence is below the threshold, this output goes to the post-correction method proposed in this work.

Table 8 compares the final performance of our proposed framework TransDocAnalyser, and compares our model with Google-Tesseract and Microsoft-TrOCR for handwritten recognition on proposed FIR dataset.<sup>12</sup> The performances are reported in terms of Character Error Rate (CER), Word Error Rate (WER), and BLEU scores [30]. Lower values of CER and WER indicate better performance, while higher BLEU scores are better.

We achieve state-of-the-art results using the proposed TransDocAnalyser framework which outperforms the other models with quite a good margin (see Table 8). While the TrOCR + InLegalBert model also performed well, our proposed framework TransDocAnalyser (consisting of vision transformer-based encoder, InLegalBert tokenizer and KNN-based post-correction) achieved the best results across all the four target fields of the FIR dataset.

## 7 Conclusion

In this work, we (i) developed the first dataset for semi-structured handwritten document analysis in the legal domain, and (ii) proposed a novel framework for offline analysis of semi-structured handwritten documents in a particular domain. Our proposed TransDocAnalyser framework including Faster-RCNN, TrOCR, a domain-specific language model/tokenizer, and KNN-based post-correction outperformed existing OCRs.

We hope that the FIR dataset developed in this work will enable further research on legal document analysis which is gaining importance world-wide and specially in developing countries. We also believe that the TransDocAnalyser

<sup>12</sup> We initially compared Tesseract with TrOCR-Base, and found TrOCR to perform much better. Hence subsequent experiments were done with TrOCR only.

**Table 8.** Benchmarking state-of-the-art TrOCR and our proposed framework TransDocAnalyser on the FIR dataset (best values in boldface)

OCR models	Target Field	Evaluation Metrics		
		CER ↓	WER ↓	BLEU ↑
Tesseract-OCR	Year	0.78	0.75	0.14
	Statute	0.89	0.83	0.12
	Police Station	0.91	0.89	0.10
	Complainant Name	0.96	0.87	0.9
TrOCR-Base	Year	0.38	0.32	0.72
	Statute	0.42	0.38	0.68
	Police Station	0.50	0.44	0.62
	Complainant Name	0.62	0.56	0.56
TrOCR-Large	Year	0.33	0.32	0.75
	Statute	0.34	0.33	0.73
	Police Station	0.36	0.38	0.65
	Complainant Name	0.51	0.50	0.57
TrOCR-InLegalBert	Year	0.17	0.17	0.84
	Statute	0.19	0.21	0.92
	Police Station	0.31	0.26	0.78
	Complainant Name	0.45	0.39	0.72
<b>TransDocAnalyser (proposed)</b>	Year	<b>0.09</b>	<b>0.02</b>	<b>0.96</b>
	Statute	<b>0.11</b>	<b>0.10</b>	<b>0.93</b>
	Police Station	<b>0.18</b>	<b>0.20</b>	<b>0.83</b>
	Complainant Name	<b>0.24</b>	<b>0.21</b>	<b>0.78</b>

framework can be easily extended to semi-structured handwritten document analysis in other domains as well, with a little fine-tuning.

**Acknowledgement:** This work is partially supported by research grants from Wipro Limited ([www.wipro.com](http://www.wipro.com)) and IIT Jodhpur ([www.iitj.ac.in](http://www.iitj.ac.in)).

## References

1. Amano, A., Asada, N.: Complex table form analysis using graph grammar. In: Proceedings of the 5th International Workshop on Document Analysis Systems V. p. 283–286. Springer-Verlag, Berlin, Heidelberg (2002)
2. Amano, A., Asada, N., Mukunoki, M., Aoyama, M.: Table form document analysis based on the document structure grammar. International Journal on Document Analysis And Recognition (IJ DAR) (2006)
3. Bag, S., Harit, G.: A medial axis based thinning strategy and structural feature extraction of character images. In: Proceedings of IEEE International Conference on Image Processing. pp. 2173–2176 (2010)
4. Bag, S., Harit, G.: An improved contour-based thinning method for character images. Pattern Recogn. Lett. **32**(14), 1836–1842 (oct 2011)
5. Bag, S., Harit, G.: Topographic feature extraction for bengali and hindi character images. Signal & Image Processing : An International Journal **2** (07 2011)
6. Bhattacharya, P., Hiwara, K., Rajgaria, S., Pochhi, N., Ghosh, K., Ghosh, S.: A comparative study of summarization algorithms applied to legal case judgments.

- In: Proceedings of European Conference on Information Retrieval (ECIR). pp. 413–428 (2019)
7. Bruno, T., Sasa, M., Donko, D.: KNN with TF-IDF based framework for text categorization. *Procedia Engineering* **69**, 1356–1364 (2014)
  8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *Computer Vision – ECCV 2020*. pp. 213–229. Springer International Publishing (2020)
  9. Chowdhury, A., Vig, L.: An efficient end-to-end neural model for handwritten text recognition. In: *British Machine Vision Conference* (2018)
  10. Constum, T., Kempf, N., Paquet, T., Tranouez, P., Chatelain, C., Brée, S., Merveille, F.: Recognition and information extraction in historical handwritten tables: Toward understanding early 20th century paris census. In: *Proceedings of IAPR Workshop on Document Analysis Systems (DAS)*. p. 143–157 (2022)
  11. Diesendruck, L., Marini, L., Kooper, R., Kejriwal, M., McHenry, K.: A framework to access handwritten information within large digitized paper collections. In: *Proceedings of IEEE International Conference on E-Science*. pp. 1–10 (10 2012)
  12. Divya, S., Gaurav, H.: Associating field components in heterogeneous handwritten form images using graph autoencoder. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. vol. 5, pp. 41–46 (2019)
  13. Duong, Q., Hämmäläinen, M., Hengchen, S.: An unsupervised method for OCR post-correction and spelling normalisation for Finnish. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. pp. 240–248 (2021)
  14. Girshick, R.: Fast r-cnn. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. pp. 1440–1448 (2015)
  15. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. pp. 580–587 (2014)
  16. Ha, H.T., Medved’, M., Nevěřilová, Z., Horák, A.: Recognition of ocr invoice metadata block types. In: *Proceedings of Text, Speech, and Dialogue*. pp. 304–312 (2018)
  17. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. pp. 991–995 (2015)
  18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. pp. 2980–2988 (2017)
  19. Hegghammer, T.: OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *Journal of Computational Social Science* **5**(1), 861–882 (May 2022)
  20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (nov 1997)
  21. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.V.: Competition on scanned receipt ocr and information extraction. In: *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*. pp. 1516–1520 (2019)
  22. Jaume, G., Kemal Ekenel, H., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. vol. 2, pp. 1–6 (2019)
  23. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) *Computer Vision – ECCV 2022*. pp. 498–517. Springer Nature Switzerland (2022)



24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*. vol. 25. Curran Associates, Inc. (2012)
25. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. In: *Proceedings of AAAI* (2023)
26. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 936–944 (2017)
27. Marti, U.V., Bunke, H.: The iam-database: An english sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition* **5**, 39–46 (11 2002)
28. Michael, J., Labahn, R., Grüning, T., Zöllner, J.: Evaluating sequence-to-sequence models for handwritten text recognition. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. pp. 1286–1293 (2019)
29. Palm, R.B., Winther, O., Laws, F.: Cloudscan - a configuration-free invoice analysis system using recurrent neural networks. In: *Proceedings of IAPR International Conference on Document Analysis and Recognition (ICDAR)*. pp. 406–413 (2017)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. p. 311–318 (2002)
31. Paul, S., Goyal, P., Ghosh, S.: LeSICiN: A Heterogeneous Graph-Based Approach for Automatic Legal Statute Identification from Indian Legal Documents. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 36 (2022)
32. Paul, S., Mandal, A., Goyal, P., Ghosh, S.: Pre-trained Language Models for the Legal Domain: A Case Study on Indian Law. In: *Proceedings of the International Conference on Artificial Intelligence and Law (ICAAIL)* (2023)
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Proceedings of the International Conference on Neural Information Processing Systems - Volume 1*. p. 91–99. MIT Press (2015)
34. Subramani, N., Matton, A., Greaves, M., Lam, A.: A survey of deep learning approaches for OCR and document understanding. *CoRR* **abs/2011.13534** (2020)
35. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10778–10787 (2020)
36. Watanabe, T., Luo, Q., Sugie, N.: Layout recognition of multi-kinds of table-form documents. *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(4), 432–445 (apr 1995)
37. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. In: *Proceedings of International Conference on Learning Representations (ICLR)* (2021)