

SmartDoc: An AI-Assisted Document Expiry Tracking and Reminder System Using OCR-Based Classification and Multi-Date Detection

Bhoomi Deore¹, Preet Kaur², Nikita Bankar³

¹ Student, ² Student, ³ Lecturer

Department of Computer Engineering

Guru Gobind Singh Polytechnic, Nashik, Maharashtra, India

Abstract: Passports, driving licences, insurance policies and certificates all expire, and a missed renewal can mean fines, refused services or legal trouble. Most reminder tools still ask the user to type in every expiry date by hand. Automating that job touches several research areas: image pre-processing and optical character recognition (OCR), document classification, date extraction, and reminder scheduling with some form of human check. In this paper we survey representative work in each area, group it into five themes and compare the studies in a literature matrix. OCR engines and document-understanding models have improved a great deal, but the studies we reviewed deal mostly with scanned pages, generic document classes and business documents such as invoices. Little attention goes to identity-type documents, which show several dates whose roles (issue, expiry, date of birth) have to be told apart. Temporal taggers normalise dates but do not assign these roles, reminder research comes mostly from health behaviour, and user correction is rarely part of the pipeline. These gaps motivate SmartDoc, a proposed web application that combines OCR, keyword-based classification, multi-date detection, priority-based reminders and a confirmation step. We close with directions for future work.

Keywords: document expiry tracking, optical character recognition, document classification, date extraction, reminder scheduling, literature survey

I. INTRODUCTION

Most of us hold a handful of documents that expire: a passport, a driving licence, an insurance policy, perhaps a few certificates. Organisations hold hundreds. Keeping track of the dates by hand is tedious and easy to get wrong, and a missed renewal can bring fines, refused services or legal complications. Reminder apps help only a little, because the user still has to type each expiry date in.

Digitisation makes it realistic to automate this. OCR turns a document image into text, and open-source engines such as Tesseract [1] and EasyOCR [2] are freely available. Reading the text is, however, only the first step. A tracking system also has to recognise what kind of document it is looking at, work out which of several dates is the expiry date, and remind the user at a sensible time through a suitable channel.

This paper is a literature survey, not an experimental study. It asks what is already known instead of reporting new results. We look at work on image pre-processing, OCR, document classification, date and information extraction, and reminder scheduling with human verification, always with expiry tracking in mind. Our contributions are a thematic organisation of the literature, a comparison matrix of representative studies, a set of research gaps, and a proposed direction, SmartDoc, that follows from those gaps.

Section II explains how we selected and analysed the literature. Section III reviews and compares it. Section IV states the research gaps and the proposed direction, and Section V concludes with future scope.



II. SURVEY METHODOLOGY

A. Search Strategy

We searched by pairing the topic with a method or application term, for example “document image OCR”, “mobile document capture”, “document classification”, “temporal expression extraction”, “key information extraction invoices receipts”, “reminder effectiveness” and “human-in-the-loop”. The sources were Google Scholar, IEEE Xplore, the ACM Digital Library, SpringerLink, ScienceDirect and the proceedings of the International Conference on Document Analysis and Recognition (ICDAR). For each result we read the title, abstract, year and venue before deciding whether to shortlist it for detailed reading.

B. Selection Criteria

We included a study if it was peer-reviewed or an established open-source technical source, addressed at least one of the five themes, was written in English and described its method clearly. We left out marketing material, non-technical product pages, duplicates and papers with no described method. From the shortlist we kept the ten sources that best represent the themes.

C. Analysis Procedure

Each shortlisted paper was read in a fixed order: title, abstract, introduction, method, results and limitations. We recorded the essentials in a literature matrix (Table I) and then followed five steps: collect the papers, classify them into themes, compare similarities and differences, critique strengths and limitations, and conclude with research gaps and future scope.

III. LITERATURE REVIEW

A. Image Acquisition and Pre-processing

Documents photographed with a phone or camera pose problems that flatbed scans do not: perspective distortion, shadows, blur and uneven lighting all hurt recognition. The SmartDoc competition [3] grew out of this. Its organisers pointed out that existing work lacked databases and baseline benchmarks for digitising mobile-captured documents, so they released datasets and evaluation protocols for both smartphone document capture and smartphone OCR. Thresholding, filtering and geometric correction are the usual first defence against poor captures, but the studies we reviewed say little about how much pre-processing contributes to the final extraction accuracy.

B. OCR Engines and Text Recognition

Tesseract is a widely used open-source engine, originally developed at HP; Smith [1] describes its line finding, feature classification and adaptive classifier. EasyOCR [2] is a ready-to-use OCR library that advertises support for more than 80 languages, which is attractive for multilingual use.

Comparisons of whole engines tell us more for practical design. Hegghammer [4] benchmarked Tesseract against Amazon Textract and Google Document AI on English book scans and Arabic article scans, each replicated with artificial noise. Both cloud services clearly beat Tesseract, particularly on noisy images, and accuracy was considerably higher for English than for Arabic. This points to a trade-off that matters for document tracking. A local open-source engine is free and keeps personal documents away from an external provider, but Tesseract lost accuracy on noisy images. The cloud services in that study were more accurate, but they are paid per page and run on remote servers.

C. Document Type Classification

Classification can work on the OCR text or on the page image. Harley et al. [5] showed that deep convolutional networks trained on document images beat hand-crafted features on RVL-CDIP, a collection of sixteen generic classes such as letters, forms, questionnaires and resumes. Multimodal models go further: LayoutLM [6] pre-trains on text together with its two-dimensional position on the page and improved results on form understanding, receipt understanding and document classification. Accuracy across these studies is high, but the models depend on large labelled or pre-training collections, and their classes are generic. Passports, driving licences and insurance policies, the very categories that matter for expiry tracking, are not among the standard benchmark classes, and labelled collections



of such sensitive documents are probably hard to assemble. In that situation, transparent keyword-based rules are a reasonable starting point. They need no training data and are easy to inspect and correct, though they break when the wording varies.

D. Date and Key Information Extraction

Rule-based temporal taggers are the most established tools for finding dates in text. HeidelTime [7] combines regular expressions with knowledge resources to extract and normalise temporal expressions, and was developed for the SemEval-2010 TempEval-2 task. It labels expressions by type (date, time, duration or set) and converts them to a standard form. It does not decide whether a date is an issue date, an expiry date or a date of birth, and its evaluation uses text corpora rather than OCR output with recognition errors.

A second group of studies pulls labelled fields out of business documents. CloudScan [8] uses a recurrent neural network to extract invoice fields without per-template configuration, trained on 326,471 invoices with data derived automatically from end-user feedback. Identity and licence documents are different: a single card may show a date of birth, an issue date and an expiry date side by side, usually next to a printed label such as “valid till”. Keyword-proximity rules look natural for this, but none of the surveyed studies evaluates them.

E. Reminder Scheduling and Human-in-the-Loop Verification

Most of the evidence on reminders comes from health research. Fry and Neff [9] reviewed 19 studies of periodic prompts used to promote diet, physical activity and weight loss. Eleven reported positive findings; the review also tried to identify which prompt characteristics went with better results, and it could not assess long-term effects. So reminders can help, but they are not automatically effective, and timing, frequency and channel deserve deliberate design. In the studies we surveyed, none adapts the schedule to how important the tracked item is, and none links reminders to automatic extraction from document images.

OCR is never error-free, so a misread expiry date would quietly produce wrong reminders. The guidelines for human–AI interaction by Amershi et al. [10] include making clear what a system can do and how well, and supporting efficient correction when it is wrong. CloudScan [8] goes a step further by using end-user feedback as training data, which shows that corrections can act both as a safeguard and as a learning signal. For our problem, a confirm-or-correct step before a record is saved is a design requirement, not an optional extra.

F. Literature Comparison Matrix

Table I records the studies discussed above under the headings recommended for a literature review matrix: problem, method, dataset or setting, key finding and limitation.

TABLE I
LITERATURE COMPARISON MATRIX OF THE SURVEYED STUDIES

Author & Year	Problem	Method	Dataset / Setting	Key Finding	Limitation
Smith, 2007 [1]	Open-source OCR engine	Line finding, feature classification, adaptive classifier	UNLV Annual Test of OCR Accuracy	Documents the engine’s design	Not evaluated on identity documents
Burie <i>et al.</i> , 2015 [3]	Capture and OCR on smartphones	Competition with shared datasets and protocols	Smartphone-captured document images	Supplied the missing benchmark for mobile capture	No classification or date extraction
Hegghammer, 2022 [4]	Comparing OCR engines	Tesseract vs. two cloud services under synthetic noise	English book scans; Arabic article scans	Cloud services more accurate, especially on noise	Not identity documents; cloud services paid and server-based



Harley <i>et al.</i> , 2015 [5]	Document image classification	Deep CNN features	RVL-CDIP (400,000 images, 16 classes)	Beat hand-crafted features	Generic classes, not expiry-bearing documents
Xu <i>et al.</i> , 2020 [6]	Document understanding	Joint text and layout pre-training	FUNSD, SROIE, RVL-CDIP	New best results on form, receipt and classification tasks	Needs OCR text and boxes as input; large pre-training
Strötgen & Gertz, 2010 [7]	Temporal expression extraction	Rule-based tagger (regular expressions + resources)	SemEval-2010 TempEval-2 data	Two rule sets: high precision, high recall	Labels expression types, not date roles; clean text
Palm <i>et al.</i> , 2017 [8]	Invoice field extraction	Recurrent neural network	326,471 invoices; 8 fields	Works on unseen layouts without templates	Invoices only; needs a very large collection
Fry & Neff, 2009 [9]	Effect of periodic reminders	Systematic review of 19 studies	Diet, activity and weight-loss interventions	11 of 19 studies reported positive findings	Health domain; long-term effects not examined
Amershi <i>et al.</i> , 2019 [10]	Human-AI interaction design	Eighteen design guidelines	Tested against 20 AI-infused products	Includes clear capability and easy correction	General; not specific to document tracking

G. Comparative Analysis

Three things stand out in Table I. First, accuracy and convenience pull in opposite directions: Tesseract is free and runs locally, whereas the cloud services in the cited benchmark were more accurate but are paid and server-based. Second, the surveyed studies were evaluated on book pages, generic document classes, invoices or health interventions. None uses identity documents, insurance policies or certificates, which are the documents that matter most here. Third, each stage is well studied on its own, but none of the surveyed studies connects capture, recognition, classification, date interpretation and reminder scheduling in one pipeline.

Table II summarises the OCR options most relevant to a low-budget web application.

TABLE II
COMPARISON OF OCR APPROACHES

Approach	Strengths	Limitations
Tesseract [1], [4]	Free; runs locally; 116 languages	Less accurate than cloud services, especially on noisy images
EasyOCR [2]	Free; runs locally; 80+ languages	Not included in the cited benchmark [4]
Cloud OCR services [4]	Most accurate in the benchmark; robust to Noise	Paid per page (US\$1.50 per 1,000 pages in 2021); server-based

H. Challenges

Four challenges keep coming up. (1) Image quality: shadows, skew, blur and low resolution degrade recognition [3], [4]. (2) Language and script: accuracy varies by language, as the English–Arabic gap shows [4], and regional-language identity documents are not covered by the surveyed benchmarks. (3) Privacy and cost: the most accurate services are external and paid, while personal documents are sensitive. (4) Data and evaluation: deep models need labelled data, and we did not find public datasets of identity documents, probably because of privacy, which makes results hard to compare across studies.



IV. RESEARCH GAP

A. Identified Gaps

Existing studies have addressed OCR accuracy, document classification and field extraction for scanned pages, generic document classes and business documents; however, limited attention has been given to turning photographs of expiry-bearing personal documents into reliable, timely reminders. We see five specific gaps.

G1. Role assignment among several dates. Temporal taggers [7] and invoice field extractors [8] find or label dates, but the surveyed studies do not address separating issue date, expiry date and date of birth on one document.

G2. Expiry-relevant document types. Classification work uses generic classes [5]. None of the surveyed studies targets passports, licences, policies or certificates, or looks at lightweight rule-based classifiers for them.

G3. Robustness on these documents. The effects of noise and script are documented [3], [4], but they have not been measured on the document types and regional languages that matter here.

G4. Importance-aware reminders. Reminder research [9] comes from health promotion. None of the surveyed studies ties reminder frequency and channel to how critical a document is.

G5. An integrated, verifiable pipeline. The stages are studied separately, and user confirmation, although recommended [10], has not been evaluated in an expiry-tracking pipeline.

B. Proposed Direction: SmartDoc

To address these gaps we propose SmartDoc, a Flask-based web application with a MySQL database. The user uploads a photo or scan of a document (Fig. 1). The image is pre-processed with OpenCV and Pillow and passed to an OCR engine, Tesseract or EasyOCR, which extracts the text, including regional languages. A keyword-based module classifies the document type, and a multi-date detection algorithm labels each date as an issue date, expiry date or date of birth using keyword proximity, with a fallback rule for unlabelled dates. The document type sets a priority level (high, medium or low), which drives a reminder scheduler that sends email or SMS notifications as the expiry date approaches. Before anything is saved, the user confirms or corrects the detected details.

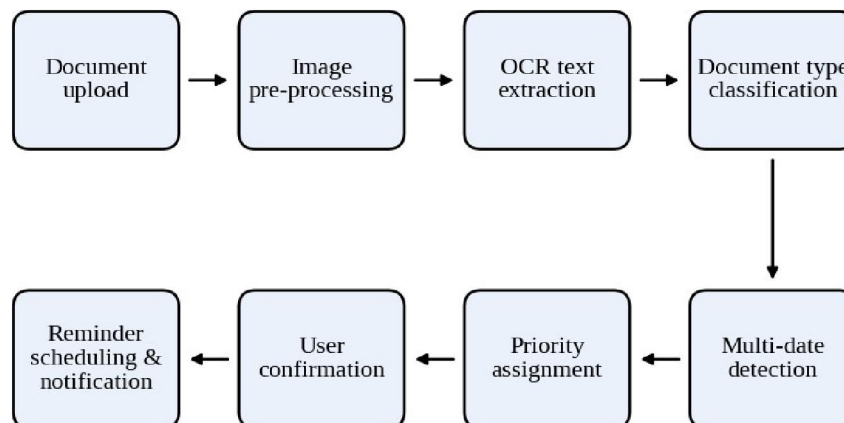


Fig. 1. Proposed SmartDoc workflow derived from the surveyed literature.

Each stage answers one of the gaps: keyword-proximity date labelling addresses gap 1, the rule-based classifier gap 2, multi-language OCR with pre-processing gap 3, priority-driven scheduling gap 4, and the confirmation step gap 5. We plan to test the system on sample documents of different types and languages and to measure OCR accuracy, classification accuracy and reminder reliability. Because this paper is a survey, we report no performance results for SmartDoc.



V. CONCLUSION AND FUTURE SCOPE

In this paper we surveyed the literature behind automatic document expiry tracking, organised it into five themes and compared nine representative studies. OCR and document-understanding methods are mature, but they trade accuracy against cost, computation and data. Extraction research targets business documents rather than identity-type documents, date interpretation lacks role assignment, reminder research sits outside the document domain, and user correction is recommended but rarely built in. These findings define the gaps SmartDoc is meant to address.

Future work includes (i) implementing SmartDoc and reporting its measured accuracy and reliability; (ii) a deep-learning-based document classifier, compared against the rule-based one once labelled data is available; (iii) document forgery and tamper detection; (iv) a mobile application with camera auto-capture, building on smartphone capture research [3]; (v) WhatsApp-based reminders; and (vi) a shared, privacy-preserving evaluation set of expiry-bearing documents.

REFERENCES

- [1] R. Smith, "An overview of the Tesseract OCR engine," in Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR), 2007, pp. 629–633, doi: 10.1109/ICDAR.2007.4376991.
- [2] JaiedAI, "EasyOCR: Ready-to-use OCR with 80+ supported languages," GitHub repository. [Online]. Available: <https://github.com/JaiedAI/EasyOCR>
- [3] J.-C. Burie et al., "ICDAR2015 competition on smartphone document capture and OCR (SmartDoc)," in Proc. 13th Int. Conf. Document Analysis and Recognition (ICDAR), 2015, doi: 10.1109/ICDAR.2015.7333943.
- [4] T. Hegghammer, "OCR with Tesseract, Amazon Textract, and Google Document AI: A benchmarking experiment," J. Comput. Soc. Sci., vol. 5, pp. 861–882, 2022, doi: 10.1007/s42001-021-00149-1.
- [5] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2015, arXiv:1502.07058.
- [6] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining (KDD), 2020, pp. 1192–1200, arXiv:1912.13318.
- [7] J. Strötgen and M. Gertz, "HeidelTime: High quality rule-based extraction and normalization of temporal expressions," in Proc. 5th Int. Workshop Semantic Evaluation (SemEval), 2010, pp. 321–324.
- [8] R. B. Palm, O. Winther, and F. Laws, "CloudScan—A configuration-free invoice analysis system using recurrent neural networks," in Proc. 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR), vol. 1, 2017, pp. 406–413, arXiv:1708.07403.
- [9] J. P. Fry and R. A. Neff, "Periodic prompts and reminders in health promotion and health behavior interventions: Systematic review," J. Med. Internet Res., vol. 11, no. 2, p. e16, 2009, doi: 10.2196/jmir.1138.
- [10] S. Amershi et al., "Guidelines for human-AI interaction," in Proc. CHI Conf. Human Factors in Computing Systems, 2019, doi: 10.1145/3290605.3300233.

