https://doi.org/10.2352/EI.2026.38.7.IMAGE-270
©2026 Society for Imaging Science and Technology

A Configurable Multi-Agent System for Feature Extraction from Multimedia Documents

Wangda Zhang, Anthony Absher, Zhen Li, Yujian Xu, Bin Shen; Celonis, United States

Abstract

This paper presents an experimental multi-agent system developed for robust feature extraction from diverse multimedia documents, including images, PDFs, and technical drawings. Addressing the enterprise demand for structuring unstructured data, the system employs a flexible architecture that intelligently orchestrates specialized agents—ranging from (Optical Character Recognition) OCR and image processing to Large Language Models (LLMs)—to achieve high-fidelity extraction. A key innovation is the system's high configurability, which keeps human experts in the loop to refine extraction logic via prompt engineering. Furthermore, the architecture supports hybrid edge-cloud deployment, allowing raw documents to be processed locally to satisfy strict data sovereignty requirements, with only non-sensitive data ingested centrally. The experimental system has shown scalability and efficiency in real-world use cases.

Introduction

Extracting structured data from multimedia documents is a critical challenge for modern enterprises. These documents vary significantly, ranging from text-heavy reports and table-rich invoices to complex, image-centric engineering drawings. Traditionally, converting these formats (e.g., PDF, JPG) into electronic records relies on manual processes that are time-consuming and error-prone. Furthermore, no single technology suffices for all document types; OCR excels at text, but fails at semantic reasoning, while LLMs offer summarization but often struggle with precise spatial layout analysis. Consequently, a robust solution must effectively combine these distinct technologies.

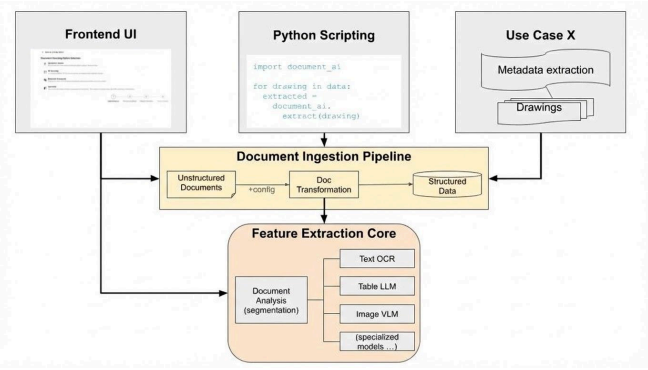


Figure 1. *Architecture overview of the Document Feature Extraction system.*

To address these challenges, we developed a configurable multi-agent system designed for end-to-end multimedia understanding. The system architecture, shown in Figure 1, consists of three core layers: (a) a Document Feature Extraction Core built on a multi-agent framework; (b) a Document Ingestion Pipeline that handles data transformation; and (c) a suite of frontend applications that allow users to control and experiment with extraction logic. By integrating specialized agents for segmentation, recognition, and reasoning, the platform supports complex enterprise environments and diverse use cases.

We made the following novel contributions:

1. *Configurable Multi-Agent System with Human-in-the-Loop*: We introduce a system where human experts actively define extraction logic via prompt engineering, rather than relying on black-box solutions. This approach allows users to craft precise instructions for specialized agents (e.g., OCR, LLMs), enabling rapid adaptation to new use cases and enhancing accuracy through domain-specific guidance.
2. *Agentic Orchestration of Specialized Agents***.** We transition from a fixed, linear pipeline to a dynamic Agentic Orchestration model. In this framework, an Orchestrator Agent first analyzes a document's layout 'overview' and the user's target schema to formulate a specific extraction strategy. This strategy-first approach ensures that the most effective specialized tools are applied to different document regions, optimizing both cost and accuracy.
3. *Hybrid Edge-Cloud Extensibility for Data Sovereignty*: We address privacy concerns through a novel architecture supporting hybrid edge-cloud deployment. Processing agents can be deployed at the edge to handle raw, sensitive documents locally; only the extracted, non-sensitive structured data is ingested centrally. This ensures compliance with data sovereignty regulations while maintaining the scalability of cloud-based analytics.

System Overview

This section presents the problem statement, defines the system requirements, and outlines the proposed solution: a configurable multi-agent system designed for feature extraction.

Objectives

The primary objective of this project is to develop a robust and adaptable system capable of automatically and accurately extracting structured data from unstructured multimedia documents (e.g., images and PDF files). This capability is intended to facilitate continuous feature extraction, and support a diverse range of business applications.

Beyond accuracy and adaptability, a core objective of the system is to ensure Reliable and Auditable Extraction. This is achieved through Provenance (Traceability), which allows every extracted value to be traced back to its specific source region (page number and bounding box coordinates) within the original document. This feature is essential for enterprise validation and regulatory compliance.

In general, the system needs to handle at least three types of multimedia documents: text-rich, table-rich, and image-rich. However, it is not ideal to only classify them as one certain type and documents usually contain more than one form of data, sometimes combining text, table, and images together. Therefore, the system has to correctly segment and detect different types of data in the document, and then for each type, the system should intelligently apply corresponding technologies and tools.

An additional requirement is addressing privacy and security concerns. To enable processing closer to the data source to mitigate customer concerns regarding uploading raw, sensitive documents to a central cloud system, the system needs to ingest only extracted, non-sensitive data. This requires distributed processing in a hybrid environment where certain features can be extracted locally to the customer's data source, rather than running centralized feature extraction for all documents.

Multi-Agent Architecture

The system's core methodology revolves around a configurable multi-agent architecture designed for intelligent orchestration of specialized agents. This architecture is built to accommodate various data types, storage mechanisms and frontend locations, ensuring modularity and flexibility. The key components and steps involved are:

1. Document Pre-processing Agent: This agent, exemplified by the Document class, handles the initial ingestion and pre-processing of multimedia documents. This includes converting images/PDFs into a format suitable for analysis and, crucially, abstracting away details of artifact storage. For complex documents like technical drawings, this agent can perform information localization (e.g., automatically cropping relevant sections) to improve the efficiency and accuracy of subsequent processing by focusing on high-value areas.
2. Prompt Engineering and Human-in-the-Loop: The system is highly configurable, allowing human experts to directly influence the extraction process. Users define features to extract and provide general prompts (e.g., "you are an expert in xyz, your goal is to accurately extract requested information from the provided document..."). This human input guides the specialized agents (table understanding agent, LLM-based text summarization agent, OCR text extraction agent, etc.), enabling tailored extraction logic for diverse use cases.
3. Specialized Agentic Orchestration workflow: The system acknowledges that different types of agents—such as OCR for text extraction, image processing for visual analysis, LLMs for semantic understanding, and table understanding algorithms—each have unique strengths and weaknesses. The multi-agent architecture is designed to intelligently orchestrate these specialized agents, combining their capabilities to overcome individual limitations and achieve comprehensive and accurate feature extraction from complex multimedia documents. The workflow can be divided into 4 stages:
    - Stage 1: Document Overview (Layout Analysis): A vision-based model identifies and categorizes document regions (e.g., tables, forms, signatures, barcodes) and assigns stable IDs to each segment.
    - Stage 2: Agent Planning (Strategy Only): An LLM-based Orchestrator Agent receives the layout overview and user requirements to plan the extraction—deciding which segments to process and which specialized extractor to use for each.
    - Stage 3: Deterministic Extraction: The system executes the agent's plan using specialized tools (e.g., table transformers, OCR engines) to extract data from the targeted regions.
    - Stage 4: Result Merging & Provenance: The system reconciles results into a single payload, providing full spatial provenance by mapping each field to its source coordinates.
4. Data Storage and Downstream Use: Extracted features are pushed into structured databases (e.g., extending existing tables with new columns) for easy access and reuse in various downstream applications via SQL queries.
5. Decoupled Computational Framework. To ensure the system can handle high-resolution visual data (such as complex engineering schematics) without sacrificing accuracy, we propose a Decoupled Computational Framework. This logic shifts away from linear processing toward a more efficient, stateful approach:
6. persist
    - Shared Visual Context: The system ingests the document into a centralized data environment once. This allows multiple specialized agents to interact with the same visual 'source of truth' simultaneously or sequentially, preventing the redundant feature extraction often found in traditional pipelines.
    - Spatial Provenance and Traceability: A core feature of this framework is the automated mapping of extracted data to its original spatial coordinates. Every extracted feature is indexed to a specific page and bounding box, providing a deterministic audit trail that links structured outputs back to the raw image pixels.
    - Resource-Optimized Orchestration: The computational load is managed by separating the Reasoning Phase (where the Orchestrator analyzes layout and intent) from the Extraction Phase (where specialized models execute the plan). This separation allows the system to allocate more intensive processing power only to the document regions identified as high-value or high-complexity.
7. Persistent Document Context. A Decoupled Extraction Paradigm, or 'Upload Once, Extract Many.' methodology has been adopted to persist the document context. In traditional imaging pipelines, the extraction request is tightly coupled with the image upload, leading to redundant pre-processing. Our framework treats the ingested document as a Persistent Visual Context. Once a document is analyzed during the Overview phase, its layout and segment data are stored as a digital twin. This architecture allows users to execute multiple, diverse extraction schemas—targeting different data points or using different agentic strategies—against the same source image without re-initiating the costly layout analysis or OCR stages.

Core Components

In this section, we describe the details of the components of the document feature extraction system.

Configurable Multi-Agent Orchestration

The multi-agent framework is highly configurable, offering a dual-mode orchestration engine that supports both deterministic and probabilistic workflows. At its core, the system utilizes a pipeline management architecture that executes a sequence of processing units. This architecture decouples the execution logic -

such as error handling and asynchronous concurrency - from the business logic, enabling two distinct operational modes:

Fixed Workflow (Deterministic): For high-volume, standardized use cases where the document layout is known, the orchestration follows a pre-defined linear path. For example, in a table extraction scenario, the pipeline is explicitly configured to execute a rigid sequence: Ingestion > Table Detection > Cropping > Structural Extraction > Formatting. This ensures maximum predictability and low latency for stable document types.

Agentic Orchestration (Probabilistic): For heterogeneous unstructured data, the system employs an agent-led orchestration model that decouples strategy from execution. This involves two distinct phases:

- **Agent Planning (Strategy Only):** An LLM-based Orchestrator Agent receives the document overview and user-defined hints. It formulates a specific plan detailing which segments to process, which specialized extractor to apply to each (e.g., table vs. form extractor), and whether to use overlapping crops for high-resolution regions. Crucially, the agent itself does not perform the extraction, which minimizes hallucinations and ensures the logic remains auditable.
- **Deterministic Extraction:** Following the agent's plan, a core pipeline executes the extraction using targeted tools on specific document segments to collect structured results and provenance data.

To maintain precision within these workflows, the system implements a "Human-in-the-Loop" configuration via a rigorous Prompt Engineering Framework. Rather than relying on raw text strings, prompts are treated as versioned configuration objects. This allows for Prompt Inheritance, where specific user extraction rules can be "overlaid" onto base system templates. Furthermore, a Request Enhancer agent acts as a meta-cognitive layer, refining high-level user instructions into strict JSON Schemas. This ensures that even in fully autonomous modes, the downstream agents utilize constrained decoding to produce machine-readable, structured outputs.

Document Understanding Agent

The Document Understanding component serves as the perceptual foundation of the system, responsible for ingesting, normalizing, and segmenting raw document streams into distinct, semantically meaningful artifacts. This agent acts as an abstraction layer over diverse input sources—whether cloud object storage, local file systems, or API uploads—ensuring a unified interface for downstream processing.

Once ingested, the agent performs Multi-Modal Decomposition, a process of separating the document into optimized representations for different types of intelligence agents:

PDF Decomposition: For PDF documents, the system employs advanced parsing algorithms (such as Docling [7]) to rasterize pages into high-resolution images for visual analysis while simultaneously extracting a structural representation (e.g., Markdown). This dual-modality allows the system to preserve the semantic structure of text (headers, lists, bolding) while retaining the visual context required for understanding layout-dependent elements like forms or charts.

Visual Normalization: For raster inputs, the system utilizes image processing pipelines to standardize inputs into a canonical format (e.g., greyscale tensors). This normalization reduces dimensionality for neural network inference and ensures consistency across diverse input formats (TIFF, PNG, JPEG).

Crucially, this agent performs the initial **Document Overview** phase required for the agentic workflow. Using a vision-based layout model, it identifies and isolates distinct regions—including **tables, text, forms, signatures, stamps, and barcodes**—and assigns a stable ID to each segment. This overview produces a structured map of the document (including page count, dimensions, and a list of segments with bounding boxes), providing the necessary context for the Orchestrator Agent to formulate an extraction strategy.

The Document Understanding Agent has been enhanced to recognize a broader taxonomy of visual entities beyond standard text blocks and tables. In the initial Document Overview phase, the vision-based segmentation model identifies and classifies segments into specific categories, including:

- Functional Elements: Tables, forms, and headers.
- Authentication Features: Handwritten signatures and official stamps/seals.
- Data Encodings: Barcodes and QR codes.

Each identified segment is assigned a unique, stable identifier and spatial bounding box. This allows the subsequent Orchestrator Agent to select the most appropriate specialized tool for each specific feature—for example, invoking a signature verification model only for segments categorized as 'Signatures'—thereby optimizing computational efficiency and accuracy.

Document Extraction Agents

Once the document has been segmented, the system routes specific artifacts to specialized extraction agents. This domain-specific approach mitigates the "Jack of all trades, master of none" problem inherent in generic models.

Text Data

For standard text extraction, the system employs a tiered strategy balancing cost and accuracy:

1. OCR Engines: For high-quality scans and standardized fonts, we utilize optimized open-source engines like Tesseract and EasyOCR. These agents are computationally lightweight and effective for dense blocks of text.
2. Vision-Language Parsing: For complex layouts, handwriting, or degraded scans where standard OCR fails, the system routes data to Large Vision-Language Models (VLMs). These models transcribe text while preserving spatial relationships and reading order, which is critical for understanding forms and non-linear documents.

Image Data

For visual tasks, such as analyzing screenshots for Task Mining or classifying photographic evidence, the system utilizes Visual Encoding Agents:

1. Visual Grounding: We employ VLMs to perform visual grounding, allowing users to query the image naturally (e.g., "Find the 'Order ID' located in the top-right header"). This agent can return bounding box coordinates, allowing downstream tools to crop and extract the specific region of interest.

2. Semantic Embeddings: For search and classification, agents utilize embedding models (e.g., CLIP) to map visual content into a semantic vector space, enabling natural language search over image archives (e.g., "Show me all invoices with a red stamp").

Table Data

Tabular data presents a unique challenge, as standard OCR often flattens the 2D grid structure into a 1D stream of text, destroying row-column relationships. To address this, we integrate specialized Table Understanding Agents:

1. Docling [7]: Developed by IBM, this library utilizes layout analysis models to robustly identify table boundaries and convert them into structure-preserving formats like Markdown or JSON.
2. GMFT (Give Me Formatted Tables) [8]: This lightweight toolkit leverages Microsoft's Table Transformer (TATR) to detect and extract tables with high throughput. It converts visual tables directly into Pandas DataFrames, ensuring that financial statements and invoices are parsed as relational data rather than unstructured text.

Engineering Drawings

For the highly specialized domain of engineering drawings (e.g., for the Scania use case), standard text extraction is insufficient due to the complex spatial arrangement of dimensions, tolerances, and geometric symbols. We employ the eDOCr2 framework, a specialized multi-stage pipeline :

1. Heuristic Segmentation: The agent uses computer vision techniques (e.g., OpenCV) to segment the drawing into functional regions: "Information Blocks" (title blocks), "Dimensions", and "Feature Control Frames" (FCF).
2. Hybrid Recognition: Text within these segments is extracted via OCR, but critically, it is then verified by a VLM (e.g., Qwen2-VL or GPT-4o). The VLM uses the visual context to correct OCR errors common in technical drawings (e.g., confusing a diameter symbol 'Ø' with the number '0').
3. CAD Parsing: For digital-native assets, the system includes a parser for CAD STEP (.stp) files, allowing for direct extraction of geometric features without the need for rasterization.

Results

The system has successfully validated various use cases and demonstrated strong capabilities in feature extraction from enterprise customer multimedia documents. In this section, we discuss several case studies in production that showcasing various aspects of the document feature extraction system.

Case Study: Engineering Drawings

Engineering Drawings: We applied the Document Feature Extraction system to automate extracting metadata (e.g., Document ID, material specifications) from complex technical drawings. Standard OCR approaches failed due to the heterogeneity of the visual layout, where text is often embedded within geometric structures or rotated.

While engineering drawings are notoriously heterogeneous in their overall content, the metadata required for this use case was located within deterministic regions (e.g., standardized title blocks). Consequently, the system utilized the Fixed Workflow topology. This allowed the pipeline to bypass the agentic planning phase and directly apply specialized extraction models to pre-defined spatial coordinates, ensuring high throughput and 100% consistency in field mapping.

First, the Document Understanding Agent ingested the PDF drawings. Leveraging the edocr2 [1] library , the agent performed Layer Segmentation to decompose the drawing into its constituent parts. Crucially, it localized and cropped the Title Block (typically the bottom-right table containing metadata) from the rest of the geometry.

Extraction: These high-resolution crops were then passed to a Vision-Language Model with a structured prompt defined in the configuration (e.g., instructions_dict). The prompt explicitly instructed the model to “extract all clearly-labeled fields and prominent identifiers and return strict Json".

Outcome: By feeding the VLM only the relevant "Information Block" rather than the entire noisy drawing, the system achieved near-perfect accuracy in metadata extraction. The pipeline automatically generated a JSON payload containing the document_id and other specified features. This validated the architectural decision to decouple "Localization" (via segmentation agents) from "Understanding" (via VLM agents).

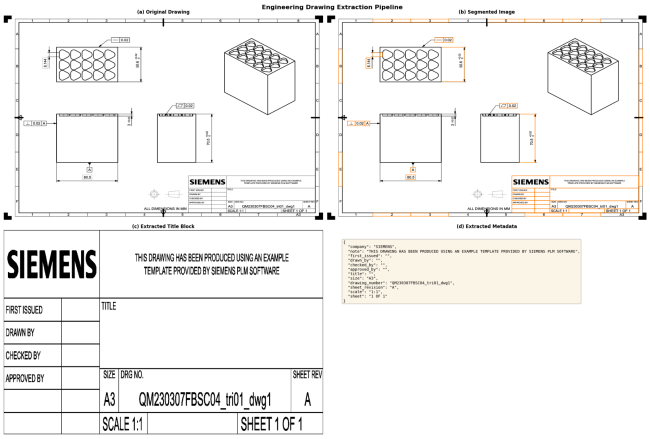


Figure 2. Automated Feature Extraction Workflow for Engineering Drawings.

Experimental Results

To quantitatively evaluate the impact of our architecture, we conducted a controlled experiment extracting the document ID from 77 engineering drawings. We compared two distinct approaches:

1. Naive LLM Extraction: Sending the raw, full-resolution image of the drawing directly to GPT-4o with a standard prompt.
2. Pipeline Extraction: Using our segmentation pipeline to identify and crop the relevant tables, then sending these localized crops (alone or combined with the full image) to the model.

The results, summarized below, demonstrate that the pipeline approach significantly outperforms the naive method:

Extraction Method	Accuracy (Correct/Total)
Naive LLM Extraction (Full Image)	71.43% (55/77)
Pipeline: Crops Only	96.10% (74/77)

Pipeline: Full + Crops	**97.40% (75/77)**

Table 1. Accuracy Comparison: Naive LLM vs. Segmentation Pipeline

The naive approach achieved only 71.43% accuracy, often struggling to locate the specific text field amidst the visual noise of the drawing. In contrast, using the pipeline to feed the VLM localized crops increased accuracy to 97.40%. This confirms our hypothesis that segmentation is critical for enabling LLMs to accurately extract text features from engineering imagery. The few remaining failures in the pipeline approach were due to the segmentation algorithm missing unusually small metadata tables.

Throughout the use cases above, we have demonstrated the following key results:

1. *Improved Extraction Quality via Information Localization*: For projects involving complex visual documents like technical engineering drawings, intelligently cropping relevant sections and feeding only those localized images to multimodal LLMs dramatically improved extraction accuracy. This transformed initially poor results into nearly perfect ones, highlighting the effectiveness of combining image processing agents with LLM agents.
2. *Prompt-Driven Adaptability*: The human-configurable aspect, particularly through prompt engineering, has proven critical. Tailoring prompts to specific document types and desired features allowed for significant performance enhancements, demonstrating the system's adaptability to diverse extraction requirements.
3. *Addressing Customer Privacy Concerns*: The design for edge-cloud extensibility directly addresses critical customer needs concerning data privacy and security. The ability to process raw documents locally and only ingest extracted features is a significant outcome that enhances trust and broadens deployment possibilities, particularly for sensitive data.

These results collectively demonstrate the system's potential to provide a highly accurate, adaptable, and deployable solution for enterprise-level feature extraction from multimedia documents.

Case Study: Invoice Analysis

Invoice analysis, typically a fully manual procedure, constitutes the foundation for a multitude of business applications. Examples include, but are not limited to, spend categorization, fraud detection, data intake, and general financial management. Given the broad applicability of this process and the clearly identifiable manual bottleneck, investigating the efficacy of automated document extraction for this procedure is a logical line of inquiry. This case study demonstrates the efficacy of the Agentic Orchestration mode in handling diverse invoice formats and languages (German) by dynamically adapting the extraction strategy to the unique visual structure of each receipt.

The present case study utilizes a collection of real-world receipts and invoices from various sources. These documents comprise both native digital formats and photographic reproductions of physical invoices. The invoices represent a diversity of purchase types, such as accommodation, physical goods, and event tickets. For the purposes of evaluation and comparison, the results will be differentiated between digital and photographic invoices, as the quality variance between these source types is expected to be considerable. The extraction task targets five specific features from each invoice: Invoice Payer, Invoice Payment Recipient, Invoice Date, Total Invoice Amount, and Currency Denomination. All invoices employed in this experiment are in the German language, and the target denomination is Euros. While some invoices may list dual denominations, the intended target denomination for extraction remains the Euro. The success criteria for our results are categorized as follows: a "success" signifies the correct extraction of all five features; a "partial success" indicates the correct extraction of at least three out of the five features; and a "failure" is defined as the correct extraction of fewer than three features.

Experimental Results

The results of our experiments with extracting these five features from our real-world examples are as follows:

Digital Invoices (8 Total)

Success	5 (62.5%)
Partial Success	3 (37.5%)
Failure	0 (0%)
Total Weighted Success*	**85%**

**Weighted Success = .625 Success + .225 Partial + 0 Failure = .85 Total*

Photos of Invoices (5 Total)

Success	2 (40%)
Partial Success	2 (40%)
Failure	1 (20%)
Total Weighted Success*	**80%**

**Weighted Success = .4 Success + .32 Partial + .08 Failure = .80 Total*

Overall, features were successfully extracted from a total of 13 invoices. The extraction of digital invoice copies proved highly effective, with a perfect success rate in the test set. The three instances of partial success observed in digital copies were exclusively attributed to the document extraction system incorrectly swapping the payer and recipient fields. Conversely, failures in the extraction of invoice photographs were characterized by errors in the extracted data, likely stemming from poor image quality, such as illegible, very small, or ambiguous text on the receipt. The extraction results for invoice photographs included one failure case where only two fields were accurately captured.

Although the limited scope of the test set precludes the drawing of definitive conclusions regarding the overall reliability of document extraction for this application, the initial findings from these real-world examples demonstrate encouraging promise. The success rates of extraction are posited to be significantly improvable through system feedback or specific training focused on the expected document types. This is particularly relevant for digital invoices, where the core data was correctly extracted but confusion arose specifically in differentiating between the payer and recipient values.

The efficacy of automated document extraction for both digital invoices and photographic images of invoices appears promising, exhibiting a success rate of reading information 80%. Further improvements can be achieved through specific model training on these document categories and by ensuring that the input images used for extraction are of high quality to mitigate the misinterpretation of data by the Large Language Model (LLM).

Related Work

The landscape of document intelligence has evolved rapidly from monolithic OCR pipelines to sophisticated agentic frameworks. Our work situates itself at the intersection of Multi-Agent Systems, Visual Document Understanding (VDU), and Human-in-the-Loop (HITL) workflows.

Agentic Frameworks for Document Understanding. Recent research has increasingly adopted agent-based architectures to tackle the complexity of long-context and multi-modal documents. Sun et al. proposed DocAgent [3], a framework that employs a hierarchical "outline-then-read" strategy, mimicking human reading patterns to navigate lengthy documents. Similarly, Colakoglu et al. introduced AgenticIE [4], which utilizes a "planner-executor-responder" loop to adaptively select tools for extracting information from complex regulatory documents. While these academic systems focus on fully autonomous reasoning, our work emphasizes a configurable orchestration layer. We distinguish between "Fixed Workflows" for repetitive, high-volume tasks and "Agentic Workflows" for heterogeneous inputs, a practical distinction often necessary for enterprise Service Level Agreements (SLAs).

Visual Document Understanding (VDU) and Table Extraction. The transition from text-only to multi-modal understanding is marked by the LayoutLM series. LayoutLMv3 [5] unified text and image masking, setting new benchmarks for layout analysis. For OCR-free approaches, Donut [6] (Document Understanding Transformer) demonstrated the efficacy of mapping input images directly to structured JSON outputs, bypassing error-prone bounding box detection. In the domain of table extraction, specialized tools have emerged to preserve structural integrity. Docling, developed by IBM, leverages layout analysis models to accurately reconstruct table structures from PDFs. Microsoft's Table Transformer (TATR) forms the backbone of lightweight extraction toolkits like GMFT, which focus on high-throughput table detection. Our system integrates these specialized libraries (Docling, GMFT) as modular agents, preventing the "hallucination" of table structures often seen in generic Large Language Models (LLMs).

Specialized Algorithms for Engineering Drawings. Engineering drawings represent a distinct challenge due to their non-linear layout and geometric syntax. Generic VDU models often fail to capture the spatial relationships inherent in these documents. Scheibel et al. introduced DigiEDraw [2], which utilizes density-based clustering (DBSCAN) to group text elements based on spatial proximity. More recently, Toro and Tarkian proposed the eDOCr2 [1] framework, which employs a hybrid approach: using heuristic segmentation to identify functional regions (e.g., Feature Control Frames, Title Blocks) before applying VLM-based verification. Our system adopts a similar philosophy for the Scania use case, deploying specialized segmentation agents to isolate "Information Blocks" before invoking costly VLMs, thereby optimizing both accuracy and token usage.

Human-in-the-Loop and Privacy-Preserving Architectures. While fully autonomous agents are the goal, enterprise deployment requires reliability. Methodologies like Active Prompting [9] for Information Extraction (APIE) demonstrate that iterative, uncertainty-aware prompting can significantly improve extraction quality. Our system incorporates this via a rigorous "Prompt Engineering Framework" that treats prompts as versioned configuration objects. Furthermore, addressing data sovereignty, recent surveys on Federated Learning and Edge-Cloud [10] architectures highlight the necessity of processing sensitive data locally. Our hybrid deployment model aligns with this, performing pixel-level processing at the edge and transmitting only structured metadata to the cloud, ensuring compliance with strict data governance regulations.

Conclusions and Future Work

Conclusion. We have presented a robust, configurable Multi-Agent Experimental System that effectively addresses the challenges of extracting structured data from complex multimedia documents. By integrating human-in-the-loop prompt engineering with a flexible orchestration of specialized agents (OCR, VLM, etc.), our system overcomes the limitations of monolithic models. The successful deployment in hybrid edge-cloud environments demonstrates its ability to resolve data sovereignty concerns while maintaining scalability. Future iterations will focus on expanding multi-modal capabilities and embedding deeper AI integration into enterprise data pipelines.

Future work. Looking ahead, we aim to evolve the system in four key areas:

1. Adoption of Foundational Multi-Modal Models: We plan to integrate next-generation foundational models that offer deeper native multi-modal understanding, reducing the reliance on separate segmentation and recognition steps for highly complex visual documents.
2. Closed-Loop Learning from User Feedback: While the current system enables human configuration, future iterations will implement active learning mechanisms to automatically fine-tune agents based on user corrections and feedback loops, progressively reducing the need for manual prompt engineering.
3. Expansion of Document Type Coverage: We aim to extend support to a wider array of unstructured data formats beyond standard PDFs and images, addressing broader enterprise archives such as proprietary CAD formats and legacy digital assets.
4. Deep Integration into Data Pipelines: To improve scalability, we will move beyond standalone extraction tasks by embedding AI capabilities directly into the data ingestion pipelines, enabling continuous, real-time feature extraction for high-volume data streams.

References

[1] J. Villena Toro and M. Tarkian, "Optimizing Text Recognition in Mechanical Drawings: A Comprehensive Approach," Machines, vol. 13, no. 3, 2025.

[2] B. Scheibel, J. Mangler, and S. Rinderle-Ma, "Extraction of dimension requirements from engineering drawings for supporting quality control in production processes," Computers in Industry, vol. 129, 2021.

[3] L. Sun et al., "DocAgent: An Agentic Framework for Multi-Modal Long-Context Document Understanding," in Proceedings of EMNLP, 2025.

[4] G. Colakoglu, G. Solmaz, and J. Fürst, "AgenticIE: An Adaptive Agent for Information Extraction from Complex Regulatory Documents," arXiv preprint arXiv:2509.11773, 2025.

[5] Y. Huang et al., "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," in Proceedings of ACM Multimedia, 2022.

[6] G. Kim et al., "OCR-Free Document Understanding Transformer," in Proceedings of ECCV, 2022.

[7] IBM Research, "Docling: An Open-Source Document Conversion Toolkit," 2025. [Online]. Available:(https://github.com/DS4SD/docling)

[8] Microsoft Research, "PubTables-1M and Table Transformer (TATR)," in Proceedings of CVPR, 2022.

[9] X. Zhao, "Large-Scale Materials Knowledge Extraction Using LLMs and Human-in-the-Loop," Drexel University, 2025.

[10] J. Liu et al., "Edge-Cloud Collaborative Computing on Distributed Intelligence and Model Optimization: A Survey," arXiv preprint, 2025.

Author Biography

Wangda Zhang received his Ph.D. in the department of computer science at Columbia University. His research interests include database, machine learning, and LLM applications.

Anthony Absher received his B.S. in computer science with a minor in mathematics from Seattle University. His work at Celonis focuses on the application of AI and ML models in the company's customer-facing platform.

Zhen Li received his Ph.D. in electrical engineering from Tsinghua University. He is a machine learning engineer at Celonis Inc. His work focuses on machine learning and AI applications.

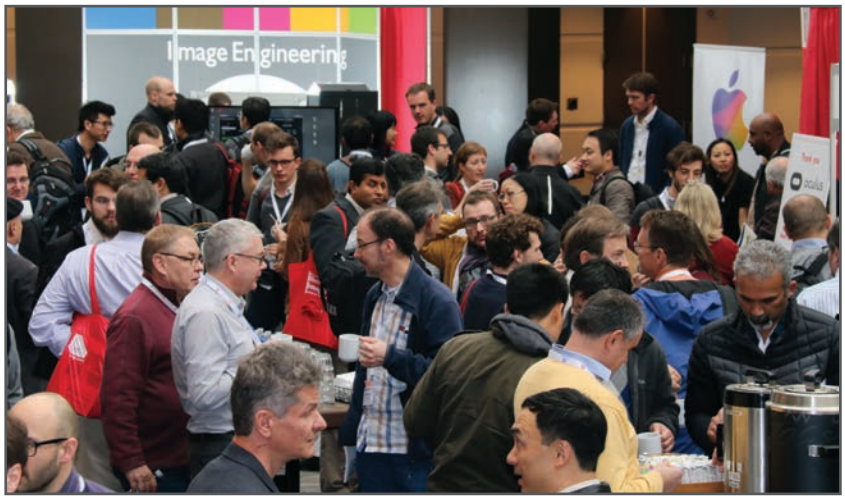
Yujian Xu received his B.S. in mechanical engineering from Shanghai Jiaotong University and his Ph.D. in the department of electrical and computer engineering at Purdue University. He is an AI engineer at Celonis. His work focuses on the detection and alignment of 3D surface properties, and LLM applications.

Bin Shen received his B.S. and M.S. degrees in Electronic Engineering from Tsinghua University (2007, 2009), and M.S. and Ph.D. degrees in Computer Science from Purdue University (2011, 2014). He is Engineering Director (AI/ML) at Celonis. Previously, he was with Pinterest's Advanced Technology Group (2019–2022) and Google New York (2015–2019). His interests include image processing, machine learning, and data mining.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org