Custom OCR with AWS Textract: Handwritten Survey Processing at Scale

A labor union, representing professional employees across the University of California system, was confronted with a challenge. During contract campaigns, the union collected thousands of handwritten member surveys—typically 8,000 to 10,000 per survey—during in-person meetings. They needed a way to digitize these forms and inject responses into a Google Sheet that would plug directly into their existing data workflow.

Previously, the union had worked with a vendor that required them to mail in paper surveys for manual data entry. They did not receive the final dataset until after survey collection ended, which significantly slowed their ability to act on results in real time. This delay hampered their capacity to make timely, data-driven decisions during critical campaign windows.

Union management approached MTech seeking assistance with automating the digitization process. The process needed to ensure careful authentication for two groups of users each of which had distinct access authorizations to perform different tasks.

The first group of union members would be authorized to merely submit the handwritten forms without being able to retrieve any other member's forms; this would ensure privacy of personal identifiable information (PII).

The second group of union members, consisting of union management, would be authorized to view the original forms and their corresponding digital versions (digital twin).

After a review of the initial specifications, MTech evaluated multiple architectures with different capabilities, development costs, and operational costs. After union management selected an architecture, MTech developed a project plan involving multiple milestones that would incrementally deliver ever-increasing functionality. The incremental approach would permit union management to evaluate the progress of the new system, and evaluate whether to proceed to the next milestone without risking payment for the full system up-front.

Requirements and Architecture

User Interface and User Authentication

Union management needed to determine how to serve two distinct user groups spread across multiple locations with varying devices and technical capabilities. After careful consideration, they chose to deliver the automation system as a web application—enabling easy access from any modern web browser on any device.

As the two groups required different functionality, each received a distinct web application tailored to their specific needs.

The front-end of the web application is delivered through AWS CloudFront to provide a low-latency response for the users via its rapid content delivery network. The first line of defense makes use of CloudFront's geolocation filtering to limit access from selected locations where union members were permitted to connect from. The next line of defense makes use of AWS Cognito providing secure, frictionless user identity and access management (IAM).

Uploading of Scanned Forms

The web application for the first group of users permits them to upload scanned forms (i.e. digital scans or digital photographs) to a secure and ephemeral AWS S3 bucket coordinated by a serverless AWS Lambda function.

Securing access to the uploaded documents is paramount: download access from the AWS S3 bucket is technically impossible by design of the architecture and by design of the access restrictions in AWS which are configured to deny by default.

Automation and the User Experience

Each upload of the document triggers, via AWS CloudWatch, its own independent and parallel automation process handled by an AWS Step Functions state machine. As each independent process could take several minutes, it is desirable to provide the user with a responsive feedback regarding the automation's progress. That is accomplished by the automation process posting status updates to AWS DynamoDB which are subsequently relayed to the user's web application via a separate AWS Lambda function.

Text Extraction and Synthesis

AWS Textract is very capable of extracting text content (optical character recognition (OCR)) including handwritten text. Nevertheless, there are several deficiencies that require post-extraction enhancements. The first deficiency pertains to the raw handwritten text which is sometimes illegible. A second deficiency pertains to whether the handwritten text is assigned by AWS Textract to the appropriate question. A third deficiency pertains to whether the handwritten text can be validated against expected information (e.g. name, location, etc.)

Addressing those challenges requires extensive auto-corrections, re-interpretations, and synthesis with other information already know by union management in order to make more meaningful interpretations of the responses. Upon completion, the “digital twin” of the form is recorded within AWS DynamoDB table along with accuracy confidence levels as to each piece of interpreted text.

Integration with Google Sheets

Each "digital twin" response is subsequently directed according to the confidence-level of the text extraction accuracy. Those responses with high-confidence accuracy are copied over to a Google Sheet that plugs directly into the union's existing data workflow. The integration also enables the union to match submissions to their member database using unique identifiers—a significant improvement over their previous process. In contrast, responses with lower-confidence accuracy are flagged for review, allowing union management to quickly scrutinize and correct only the exceptions rather than manually transcribing every submission.

Data Sovereignty and Cost

All of the scanned documents and processing are maintained within the union management’s AWS account. Access to the data is restricted to union management whose access is permitted only by proper authentication with AWS Cognito.

Furthermore, as the various system components are configured to be serverless and consume minimal storage, the system incurs practically zero cost while not in use.

Deployment into AWS

Testing the various versions of the system without affecting the current version in production requires a method of provisioning new and updated infrastructure reliably and quickly. Towards that end, the entire system is defined in AWS CDK which is a framework for defining cloud infrastructure in code and provisioning it through AWS CloudFormation. It further provides the ability to audit the evolution of the infrastructure over time.

Outcome

After internal testing, the automation system was activated and deployed to the field. The benefits were immediate and notable.

- Union members were able to upload paper surveys from anywhere in the field and receive results in minutes—rather than waiting until after survey collection ended.
- The responses were automatically corrected, cross-referenced, matched to the union's member database using unique identifiers, and made available in real time.
- Union management could now act on results during the campaign rather than after it—a critical advantage for time-sensitive organizing efforts.

Across thousands of responses uploaded from the field, the automation system has proven to be fast, reliable, and cost-effective. Perhaps most importantly, the system was designed to be modified and reused for future campaigns and survey projects—providing lasting value beyond the initial engagement.

Alex C, Director of Operations, summarized the engagement: "The customized system dramatically increased our processing speed... We would absolutely recommend [MTech] to any organization looking for a customized OCR solution to improve data collection and processing."

If you'd like help with your next project, reach out to discuss.

Get it done right
Get MTech