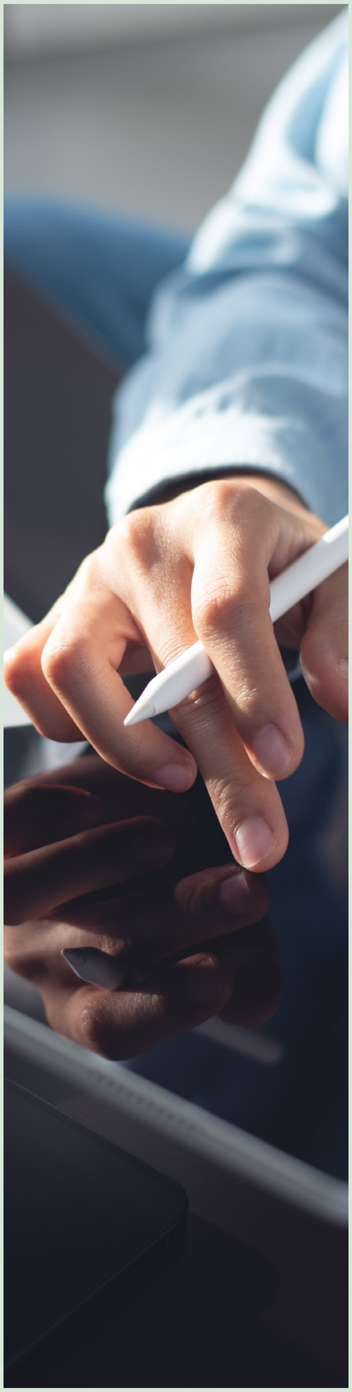
WHITE PAPER



OBJECT-BASED VECTORIZATION OF GEOSPATIAL DATA – AN AI APPROACH

Abstract

Vectorization—the process of converting raster data into vector format—remains a significant challenge in the geospatial industry. Traditional methods of producing a vector data from a raster, especially when working with scanned PDFs, make image interpretation complex and require substantial manual effort to accurately map the individual assets. This white paper presents a proof of concept that leverages an AI-driven approach to automate this conversion process by handling the map data as layered components or objects, such as electrical conductors, roads, and buildings. The focus is on extracting line and polygon features using BOTs, which substantially reduces manual effort and enhances efficiency.

Infosys®
Navigate your next

Business Challenge

The mapping of utilities, assets, and topography using digital technology is key to effective management and better use of spatial data. New tools and innovative approaches are required for utility providers to stay competitive. In this context, converting raster files to vector format—also known as vectorization—plays a crucial role in processing and managing geospatial data with efficiency and accuracy.

However, this vectorization process is traditionally complex and labor-intensive as it involves interpreting the image characteristics, such as lines and polygons, and mapping individual assets manually. The primary challenge lies in finding a concrete, reliable method to effectively convert scanned images into vector representation (i.e. points, line, and areas).

Traditional approaches to interpreting data from the raster images, such as scanned PDFs, hinders and complicates automation efforts. From a legacy modernization perspective, finding an effective conversion process is crucial to minimize operational costs while maximizing accuracy. This white paper proposes a novel method of handling data as individual layers or objects, each with its own characteristic features, allowing for more precise and efficient process of data conversion from raster to vector format.

Approach

This AI-driven PoC approach focuses on extracting and converting the linear and polygonal features into structured layers through a workflow discussed below.

Experimentation with high-level workflow

The high-level workflow depicts the various stages of vectorization process. This workflow has been adopted to derive the automated output and cleanse the deliverables using commercial off-the-shelf (COTs) data reviewing and cleansing tools. The stages of vectorization process are illustrated in the figure below.

High-level workflow

Upload images
- Scanned PDF/Image

Pre-processing
- Image enhancement
- Deskewing & Rotate

Feature Extraction
- Line detection (Ex. Hough Transform)

Segmentation
- Thresholding
- Morphological Operations

Vectorization
- Skeletonization
- Polyline Extraction

Machine Learning
- Training models
- Feature matching

Post processing
- Smoothing & Simplification
- Quality Assurance

Output Generation
- Export to GIS formats (Ex. Shape file, GeoJSON)

Source of PoC Data

Scanned PDFs from electric utilities were used as the primary data source for this white paper. These scanned PDFs contain utility facilities or network components, such as electrical conductors and road network presented as linear features. The scanned images also identify other land base, polygonal features such as buildings.

Pre-processing

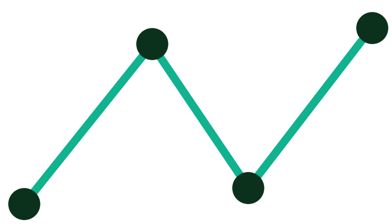
In this stage, the quality of the scanned image is enhanced using techniques such as contrast adjustment, noise reduction, and binarization. This stage involves de-skewing and rotation where any skewness or rotation in the image is automatically corrected to ensure that the features are properly aligned.

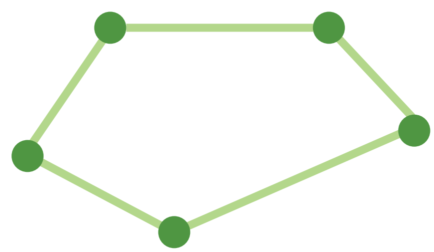
Feature Extraction

This stage focuses on mapping linear features such as electrical conductors and roads with enhanced accuracy. It includes several key processing steps to ensure precise extraction and representation of these line features.

Segmentation

This stage involves *thresholding* where the image is converted into a binary format, effectively distinguishing linear features from the background. This stage also involves *morphological operations*, utilizing techniques, such as dilation and erosion, to refine the extracted features, thereby improving accuracy and clarity.

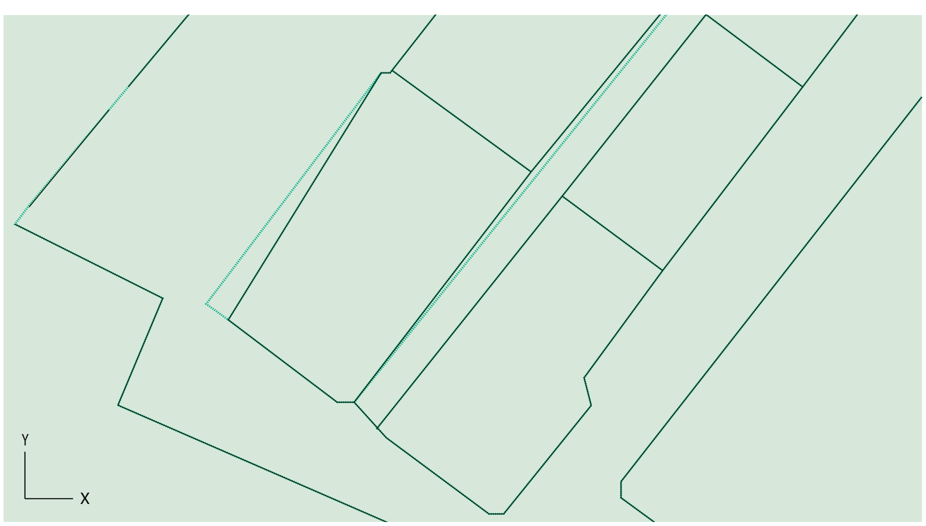




Vectorization

This process involves extracting or identifying lines and polygons from the scanned images. When scanned lines appear several pixels wide, the first step is typically to thin them down to retain only their centerline or skeleton. This process, known as **skeletonization**, reduces the thickness of the lines to a single pixel width, creating a simplified, skeletal representation of the original features.

Next is **polyline extraction**, which involves tracing the skeletal lines and curves to transform them into vector data. In this step, the thin skeletal lines are converted into a series of coordinate pairs or points that define a polyline, enriching them with meaningful data for further analysis or use.

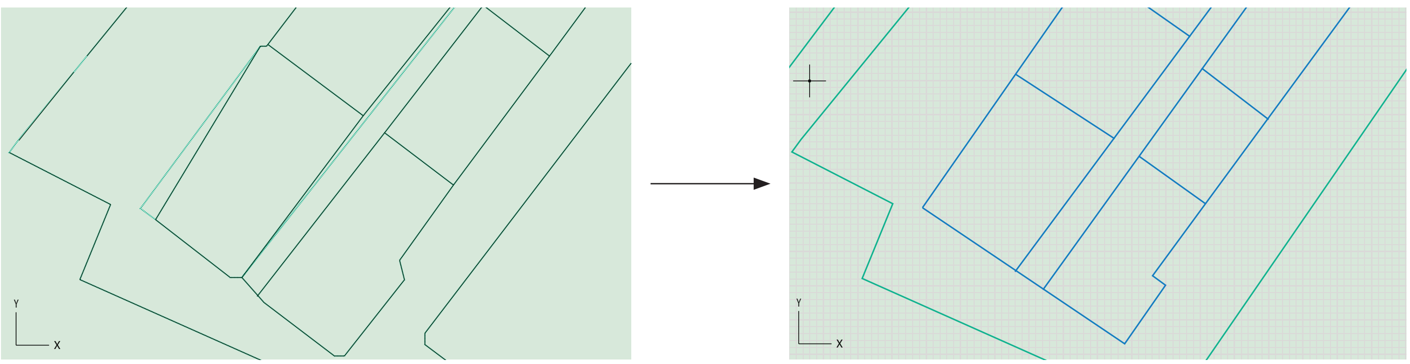


Machine Learning

Machine learning plays a crucial role in automating feature extraction, particularly in recognizing and classifying linear features. Training models, such as *Convolutional Neural Networks* (CNNs), enable accurate identification of different types of linear elements. These trained models are then utilized for feature matching, allowing extracted elements to be categorized into predefined classes, such as roads and electrical conductors.

Post Processing

Post-processing steps further refine the results, including smoothening and simplification techniques to reduce the number of vertices while preserving essential details. Additionally, quality assurance measures, whether manual or automated, are employed to validate and correct any errors in the vectorized data, ensuring high accuracy and reliability.



Output Generation

The final step in the process involves exporting the vectorized data into standard GIS formats for further analysis and practical applications. Converting the processed data into widely used formats such as *Shapefiles* or *GeoJSON* ensures compatibility with various GIS platforms, enabling seamless integration into mapping, spatial analysis, and infrastructure planning workflows. This structured output allows for efficient data sharing, visualization, and advanced geospatial analysis, supporting diverse use cases in urban planning, utility management, and transportation networks.

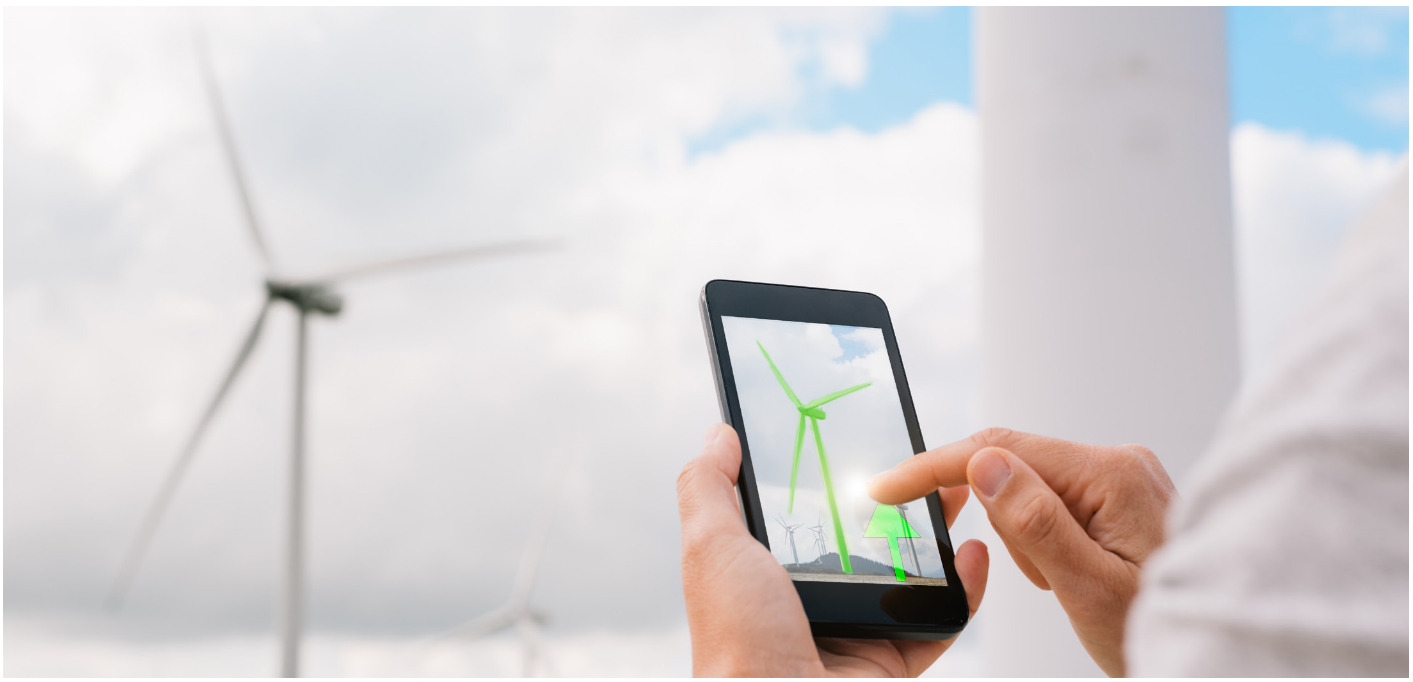
The automated output is cleaned using data reviewing and cleansing tool. For this case, the CAD tools were utilized.

Analyzing the Results

This proof of concept demonstrated significant effort savings across different feature types and complexities. The metrics are summarized below:

Feature Type	Weighted Overlay Density	Unit – Minutes			Savings in TAT
		Manual Effort	Semi-Automatic Effort	Savings	
Line (Electrical conductors and roads)	Medium	105	40	65	62%
	Complex	390	210	180	46%
Polygon (Buildings)	Medium	15	5	10	67%
	Complex	45	15	30	67%

These metrics indicate substantial effort savings based on the proposed object-based geometric handling methods.



Technical Development

To ensure the best deliverables, it is crucial to develop specific AI-powered tools tailored to the project's requirements. The following proposed tools aim to enhance the efficiency and accuracy of mapping geospatial features.

Line Geometry Tool

This tool will focus on mapping linear features, such as electrical conductors and roads. It automates the detection and mapping of lines from raster data, significantly reducing the need for manual intervention.

Shape Geometry Tool

The tool will focus on mapping polygonal features, especially buildings. Its primary functionality is to identify and outline building shapes from raster images with high precision. This ensures accurate mapping while enhancing efficiency. By leveraging this tool, users can achieve improved accuracy in capturing building footprints, significantly reducing turnaround times and minimizing manual effort.

Budget Considerations

The development of these AI-based tools is anticipated to require a modest budget, but initial studies and previous experience suggest they could potentially lead to substantial cost savings.

These savings will be achieved through several key factors. First, reduced manual effort will be made possible by automation, significantly cutting down the time and resources spent on manual mapping tasks. Second, increased efficiency will be driven by AI-powered tools that speed up the mapping process, allowing for quicker project completion. Finally, enhanced accuracy in mapping will minimize errors, reducing the need for costly corrections and rework. Overall, these improvements will optimize workflow, reduce costs, and enhance overall project productivity.

Key Takeaways

This proof of concept highlights the feasibility and efficiency of automating raster image to vector conversion using a layered (object-based) approach. By focusing on individual geometrical objects and leveraging tools such as BOTs, significant reductions in manual effort were achieved. Additionally, this method is scalable, making it suitable for large-scale applications across various geospatial domains, including mapping, infrastructure monitoring, utility management, and more. AI-driven automation enhances efficiency, driving broader adoption in industries that depend on accurate and timely geospatial data processing.

About the Authors

Author

Karthigaiselvan R

Karthi heads GIS and AI Annotation services and has 21+ years of techno-commercial experience in GIS and AI annotation services. He specializes in building innovative solutions and ensuring seamless project delivery. His expertise spans solution design, execution, and driving business growth. He did his M. Tech degree in Remote Sensing from Anna University, Chennai, and his Engineering in Electrical & Electronics from Government College of Technology, Coimbatore.

Co-author

Shyam Rao

Shyam heads the Digital Business Services Unit at Infosys BPO and has over 20 years of experience in global business service delivery with specific focus on Digital Marketing Services (DCX) and GIS Operations. Shyam is responsible for marketing, sales and delivery of Digital Operations for leading Fortune 500 clients and has led several consulting engagements to define the digital operating model and drive innovation in their digital operations. He is an alumnus of London School of Economics and the Digital Marketing Institute of Ireland.