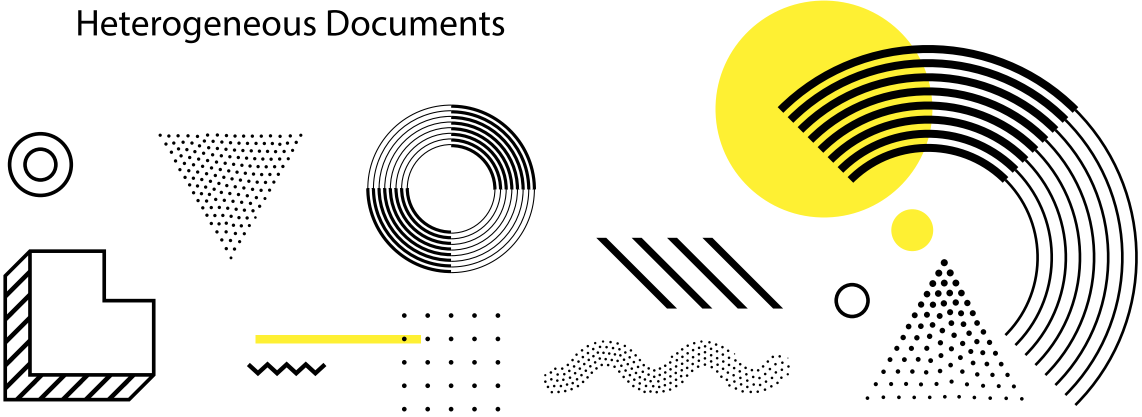
# Vision: Moving the Needle from Sight to Insight

## Creating Insights from Complex Heterogeneous Documents



By Kajari Ghoshdastidar

Senior Product Manager,
Infosys Nia,
EdgeVerve Systems Ltd. (An Infosys Company)

The evolution of AI can simplify and make existing enterprise operation models efficient. But its real value lies in allowing them to tackle greater complexity with confidence. Computer Vision, which deals with automating tasks usually performed by the human visual system, is one of the most exciting areas in this regard. Today, the intersection of several technological advances ranging from the rise in computing power even on the edge devices, to advancements in Machine Learning algorithms, are making Computer Vision a reality beyond hypothetical use cases, its capacity even exceeding human visual computing power. This article seeks to understand how computer vision algorithms can help us mimic human capabilities in using visual cues to find insights in visually rich complex documents.

Enterprises across industries and verticals process a staggering amount of information. The different content and context of this data can make processing and classification a significant challenge. Categories can include operational documents such as contracts or invoices to more complex materials such as infographics.

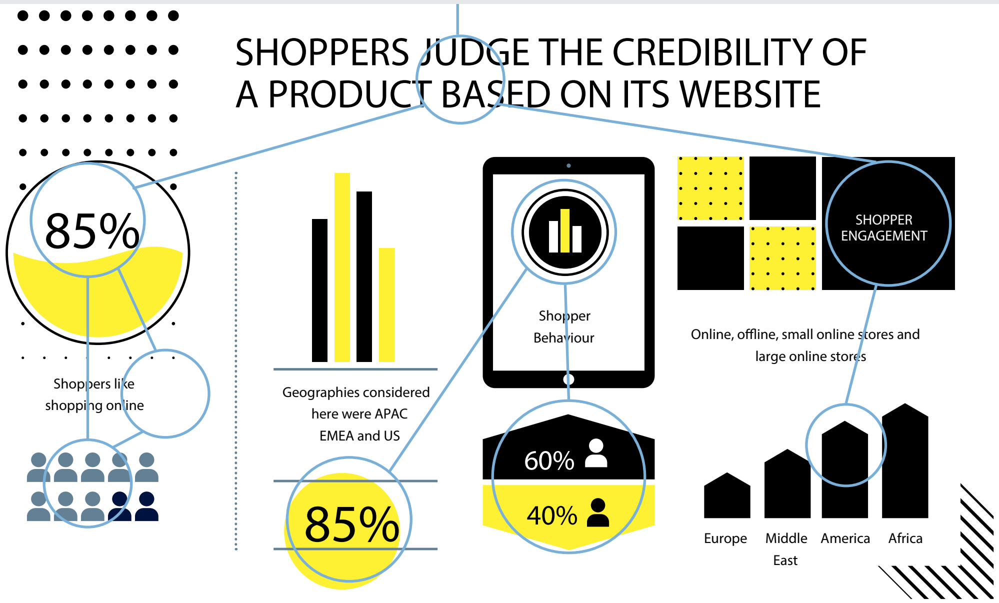
## Visually Rich Documents

Documents can be broadly segmented into four categories:

- Continuous textual flow like books
- Form-based textual documents such as forms, receipts, and invoices
- Visually-rich documents like posters, banners, and infographics
- Images such as pictures, screenshots, and video frames

## Anatomy of a Visually Rich Document

Sparsely placed text and visual segments logically connected using visual cues to create a coherent story.


SHOPPERS JUDGE THE CREDIBILITY OF A PRODUCT BASED ON ITS WEBSITE
85%
Shoppers like shopping online
Geographies considered here were APAC EMEA and US
85%
Shopper Behaviour
60%
40%
SHOPPER ENGAGEMENT
Online, offline, small online stores and large online stores
Europe
Middle East
America
Africa

### Artifacts and Formatting Cues

| | | | |
|---|---|---|---|
| Images | Logos | Word Art | Symbols |
| Chart | Bar | Text | Caption |
| Font style | Color | Background | Whitespace |
| Positioning | Size | Orientation | Symmetry |

Visually rich documents, as you can see, are inherently heterogeneous. The complexity arises not from being abstract but from the fact that humans typically, and instinctually, interpret a combination of seemingly disconnected visual elements and limited text to generate a perfecty coherent message (refer to Fig 1: Anatomy of a Visually Rich Document).

Traditional Optical Character Recognition (OCR) continues to be used at large for the first two categories, but it falls short for the remaining ones. OCR for character or word recognition relies on layout analysis or 'zoning' of a document to establish a baseline for various document elements like character shape, size, orientation, text background, among other factors.

Therefore, consistency of format and the presence of a clear pattern in font styles and backgrounds are critical in ensuring OCR accuracy. The reality for the visually rich documents in the enterprise context, however, is anything but homogenous and consistent.

## Areas of Impact

Unlike structured textual documents that rely on templates and NLP-based analytics for information extraction, the semantic structure of visually rich documents is observed primarily by visual cues interpreted by the human brain. In this context, the ability to automate and augment document recognition for intelligence at scale can deliver a powerful impact.

Consider the following use cases:

1. Ad Tech Companies - These organizations are likely to analyze a wide range of materials such as posters, pamphlets, catalogs, digital ads, and other content assets. The results could be used to inform their conversion strategy, devise promotions, and focus their content creation approach.
2. Marketing - Teams need to analyze marketing assets, competition communication, and even industry research materials to develop cogent approaches for marketing and sales.
3. Research - Large research organizations frequently have to sift through thousands of pages of information in different formats, creating inefficiency, and also running the risk of bias and inaccuracy from human processing.
4. Retail - To extract information from product labels.

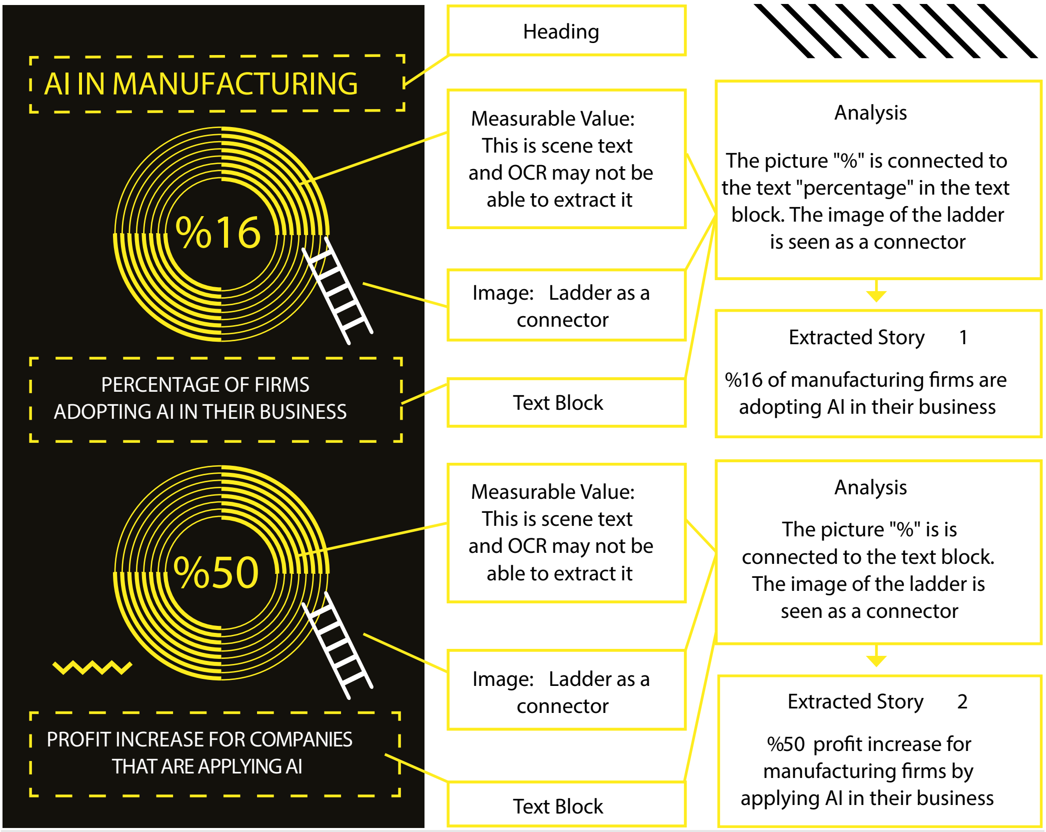
With traditional OCR just not equipped to handle this level of complexity, each of these areas can require a substantial effort of time and investment. It is here that, with accuracy in object identification and image classification doubling over the past decade to 99 percent , the progress in the field of computer vision is proving transformative. So, how does it work?

## Seeing and Understanding: Multimodal Extraction Approach

A cursory output that offers some detail, often incomplete, would still require manual interpretation to make progress, defeating the purpose of the exercise. The actual benefit of computer vision lies in identifying meaning, not mere object recognition and classification . Multimodal image extraction is the technique that refers to the collective analysis of multiple information types in the same document to create a coherent understanding of its content. In this method, the visual cues from the image are used to tie different document segments (or entities) into a cohesive message.

Computer vision augmented with text analytics can extract the semantic structure and content from such documents. To paint a picture, a visually rich document can be represented as a graph, with each node of the graph containing specific information and the edges of the graph connecting the information logically. Computer Vision and OCR are used in conjunction to extract data from each node while the graph's neighborhood knowledge ties up all the information together to build a consumable narrative . We can understand this better with an example.

Computer vision and OCR are used in conjunction to extract data from each node while the graph's neighborhood knowledge ties up all the information together to build a consumable narrative.



AI IN MANUFACTURING

%16

PERCENTAGE OF FIRMS ADOPTING AI IN THEIR BUSINESS

%50

PROFIT INCREASE FOR COMPANIES THAT ARE APPLYING AI

Heading

Measurable Value: This is scene text and OCR may not be able to extract it

Image: Ladder as a connector

Text Block

Analysis

The picture "%" is connected to the text "percentage" in the text block. The image of the ladder is seen as a connector

Extracted Story 1

%16 of manufacturing firms are adopting AI in their business

Measurable Value: This is scene text and OCR may not be able to extract it

Image: Ladder as a connector

Text Block

Analysis

The picture "%" is is connected to the text block. The image of the ladder is seen as a connector

Extracted Story 2

%50 profit increase for manufacturing firms by applying AI in their business

## Final Story

6% Of manufacturing firms are adopting AI in their business and their profits increase by 50% from applying AI in their business

The infographic above has been selected as an example of digitized content with a range of elements that may throw off traditional OCR. It includes text and scene text alongside images in the forms of logos, scene text, and vector illustrations. Now, the heterogeneous structure and combination of element formats in this asset will confuse OCR even from a recognition and classification standpoint, much less the more complex challenge of assimilation and comprehension. Here is how a typical multimodal approach can help:

- The algorithm detects that the marked blocks are connected. The other blocks, if any, are separate and will not be considered integral to the main story output. This ability is achievable using neighborhood information through graph embedding.
- Computer Vision then uses the hierarchy of information to understand that ‘AI in Manufacturing’ is the heading, before making the inference that all mentions of ‘firm’ in the text boxes refer to a manufacturing firm.
- When you look at the numbers highlighted in the graph — 16% and 50% — you can see that they represent scene text, much like the text on a stop sign in a photograph of a busy road. OCR cannot process these text formats, but Computer Vision can understand the difference intelligently and apply separate scene text processing techniques for interpretation.
- All text blocks will be identified accordingly and directed to the OCR module.
- The ‘ladder’ in the image is identified as a connector. The fact that there is a second connector in the image tells the system that the story is incomplete at the first connector, and there is more information to look for.

- It then extracts the information in the first half of the asset, creating a story for that segment before repeating the same process for the second half.
- Once extracted, the technology pieces together all the information to generate a consumable and analyzable output.

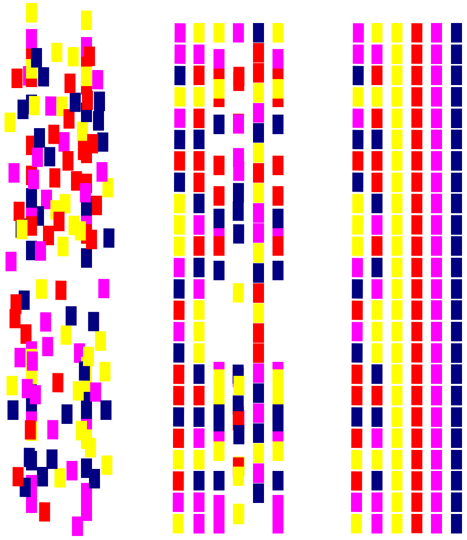
As you can see, the value of the output in this scenario lies in the analysis just as much as the identification. Unlike OCR, Computer Vision can comprehend information and context, much like the human brain and often faster, significantly improving document analysis speed and quality.

While the final choice of tech stack may vary, enterprises must look to augment their image recognition needs with intelligence. The approach is fast becoming a growth driver in fields ranging from healthcare where AI has been found to detect illness from scans faster than physicians and automotive where the decision-making ability of autonomous cars is almost singlehandedly driven by advanced Computer Vision.

## Navigating the Challenges

As with any emerging technology that promises disruption at scale, Computer Vision faces a few challenges. To be able to interpret content-rich data, it is vital to construct deeper and broader vision models that cover a more extensive range of data features and classes. Creating these supervised models require access to massive labeled training datasets, which can be nearly impossible in real-world applications. Furthermore, even if this data is available, it requires a substantial investment of manual effort, resulting in processing delays and chances of inaccuracy. This is why training of models with less training dataset is a popular active research area. Approaches like active learning, semi-supervised learning, unsupervised learning with data reconstruction, noise injection, and N-shot learning are some of the training techniques being explored to create accurate models with less training data set.

Training data set is not the only constraint we have. Computation time and cost for training and inference for such deep learning models are significant. Most of our customers are looking at quick onboarding (fast training), real-time prediction support and frequent retraining of the models with little to no downtime. We can leverage



recent advancements in both hardware and software-based approaches such as decentralized training infrastructure, adaptive incremental learning to help standardize and optimize resource usage.

> For adoption, enterprises must look to work with platform and product partners already equipped with such advanced techniques to address these challenges to improve business values, while ensuring business continuity and optimal resource usage.

Built on the Infosys Nia platform, Nia DocAI offers enterprises the ability to identify information from images and scanned documents using object detectors, OCR, handwriting recognition, and signature tagging. DocAI combines advanced Machine Learning, Computer Vision, and natural language processing to offer a robust layer of intelligence, delivering on-demand services such as intelligent document processing, data enrichment, cognitive search catering to critical business needs like contract analysis.

[1] *https://arxiv.org/abs/1903.11279*
[2] *https://dl.acm.org/doi/10.1145/2682571.2797092*