Computer Vision

Greete Veesalu

2024-06-04

An overview of computer vision, how it works and it's various uses in making digital collections more accessible…

Introduction

Computer Vision is a field of Artificial Intelligence (AI) that uses Deep Learning models to teach machines to see and interpret images and videos similar to how the human visual system does. At its core, Computer Vision involves functions like process, detection, and analysis of visual data. It uses algorithms to extract meaningful information from images and enables machines to identify objects, recognise patterns, and make decisions based on what they see. It is behind a lot of today's technology like facial recognition, object detection, or Video Assistant Referees at football matches.

Deep Learning uses neural networks to model complex patterns within data. These networks, much like the human brain, consist of interconnected layers of artificial neurons. During training, these models process massive datasets of images, enabling them to progressively identify features and refine their predictions with increasing accuracy. For instance, a network might initially learn to detect edges and shapes. As training progresses, it learns to identify more complex features such as curves, textures, and ultimately, the entire object itself. This continual process of feature extraction and refinement enables Deep Learning models to achieve remarkable accuracy.

Convolutional Neural Networks (CNN) have become a cornerstone of image recognition as these neural networks are particularly well-suited for processing grid-like data, such as images. CNNs use convolutional layers that extract local features from the input image (edges, corners, etc.), which are then combined to form more complex representations in subsequent layers, allowing the network to learn hierarchical representations of the image content.

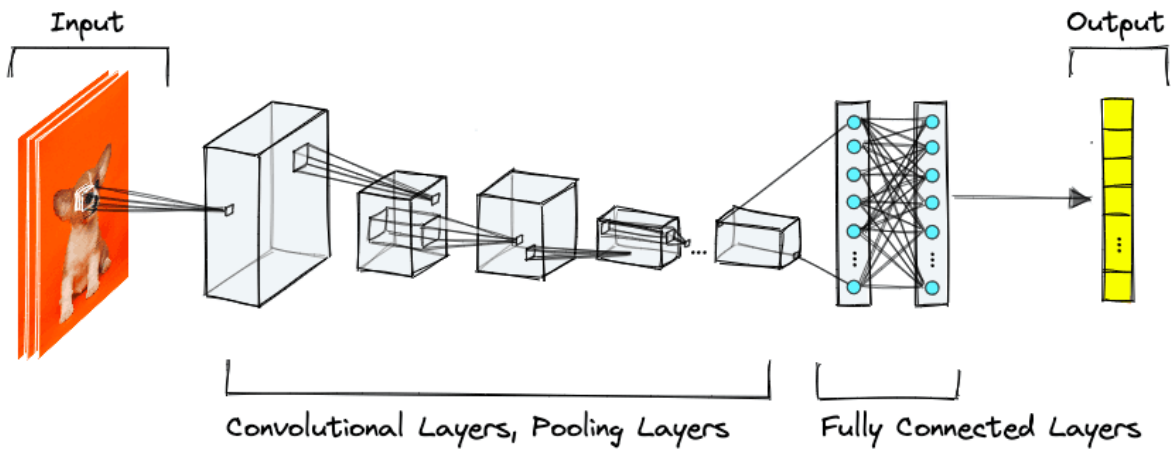


Image retrieved from https://www.pinecone.io/learn/series/image-search/cnn/#What-Makes-a-CNN.

Key functions of Computer Vision include:

- Image processing - Preparation of input images to enhance them, e.g., noise reduction, sharpening, filtering, extraction of colour, etc.
- Object detection - Identifying specific objects within an image or video (e.g. R-CNN, SSD).
- Scene understanding - Analysing the overall context of an image.
- Object classification - Assigning specific objects to a category or class (e.g. AlexNet, VGG).
- Image segmentation - Dividing an image into distinct regions or objects (e.g. U-Net, Mask R-CNN).
- Motion analysis - Tracking the movement of objects within a video (e.g. optical flow).

Key challenges in Computer Vision include:

- Biases in training data - For example, facial recognition systems trained predominantly on lighter-skinned people struggle to accurately recognise darker-skinned people.
- Privacy and ethics - The usage of visual and biometric data raises questions regarding privacy and ethics, especially when dealing with black-box models.
- Generalisation - Models struggle to generalise objects or environments they were not trained on.
- Occlusions - Often objects in an image are partially blocked by other objects, making it difficult for algorithms to correctly identify them. Advanced models, like Mask R-CNN, can help with missing details, but it requires large amounts of training data with occlusion examples.

- Fine-graining - Distinguishing between classes or very similar categories is difficult. For instance, it's relatively easy to differentiate between a cat and a dog, but it is challenging to distinguish dog breeds.
- Lighting - Changes in lighting conditions can alter the appearance of objects in an image.
- Resource needs - Training and running Deep Learning models require a lot of computational resources.

Computer Vision in the Library Sector

Computer Vision addresses the challenges posed by vast and diverse collections, enabling libraries to improve services and access to knowledge.

- Image Classification and Cataloging - Libraries manage large digital libraries containing very different material. Computer Vision and automated cataloging automate the classification of the collections, thus reducing the manual labor required and improving discoverability.
- Object Detection and Description - Models can automatically detect features and items in the material, this is useful for creating detailed metadata in large collections. Libraries can also use object detection to improve accessibility, by generating descriptive metadata for images and videos.
- Automatic Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) - OCR/HTR helps libraries to make printed texts and handwritten documents digital. Thus, libraries use OCR/HTR to create searchable digital collections of books, manuscripts, newspapers, etc. OCR/HTR also supports accessibility, enabling text-to-speech conversions for visually impaired users.
- Visual Search and Content-Based Retrieval - Computer Vision improves search functionality in digital libraries through visual search. This allows users to upload an image and retrieve visually similar items from the collection. Content-based image retrieval also supports researchers by enabling them to find works that are similar.

Relevance to the Library Sector (Case Studies/Use Cases)

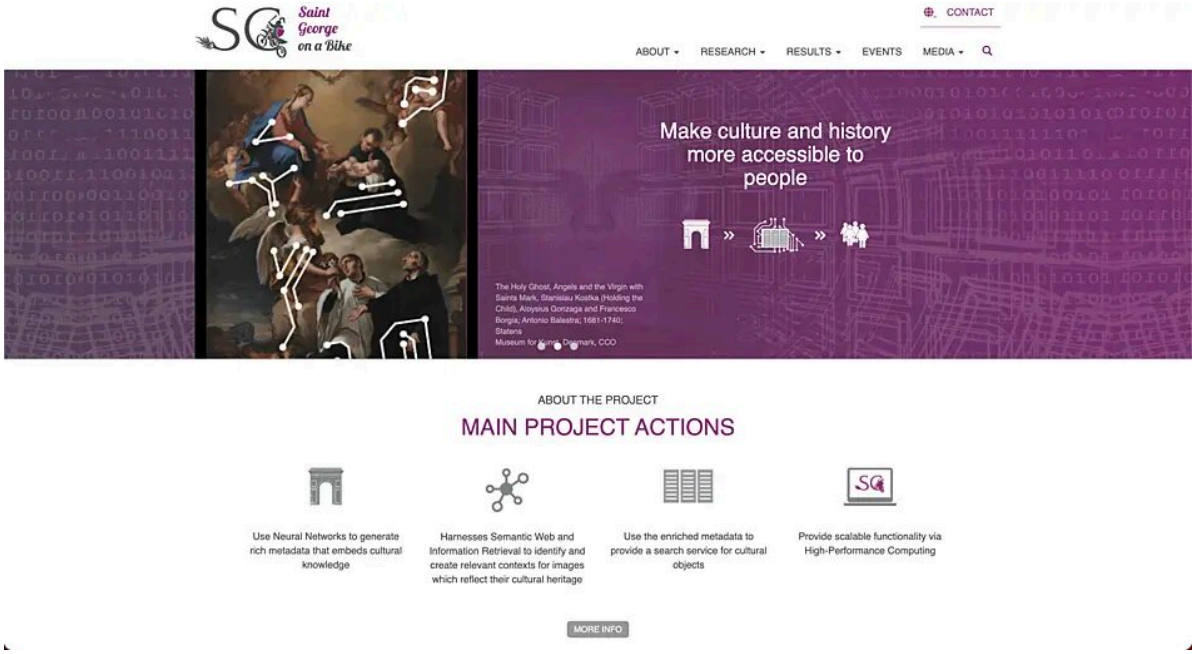
Having explored the foundational concepts of Computer Vision, let's now delve into how it is being applied in the library sector. We will take a look at some key applications within areas like collection enrichment, access and delivery, and accessibility.

Enrichment of collections

Computer Vision is transforming the enrichment of library collections by automating the classification and annotation of visual materials. With the help of object detection and image

recognition, it identifies objects, patterns, and relationships within images, enabling the creation of richer metadata and contextual information. This not only improves discoverability, allowing users to locate materials more easily, but also allows for deeper insights into the historical, artistic, or cultural significance of the materials.

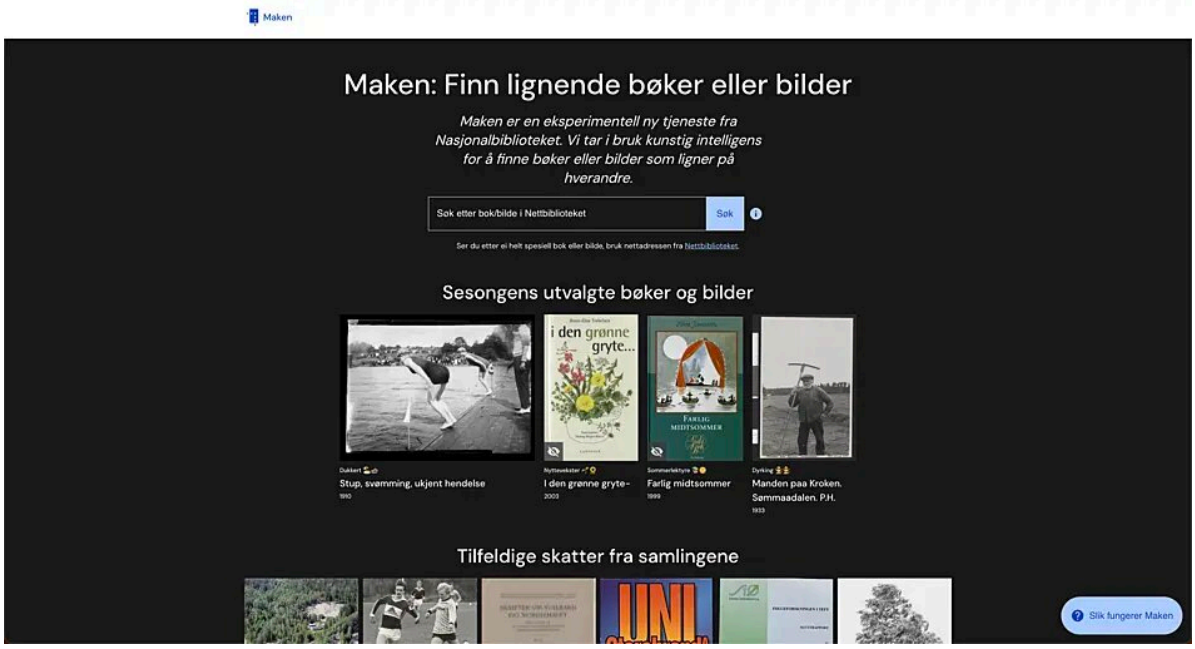
A case study conducted by the University of Calgary's Libraries and Cultural Resources explored the use of Sheeko, a metadata generation software, to create descriptive metadata from images using pre-trained models. The study concluded that, with moderate human intervention, both the descriptions and titles generated by the pre-trained models were usable. Similarly, Saint George on a Bike which was a project led by Europeana, aimed to improve the classification and metadata of cultural heritage objects, by using CNNs for object detection to identify and create semantic context for the images. These examples demonstrate the potential of combining automated AI-powered tools with human expertise to improve metadata creation, while enriching cultural heritage collections and making them more accessible and meaningful for diverse audiences.



Access and delivery

Similarly, Computer Vision is improving access, delivery, and research within library collections by enabling more intuitive and sophisticated ways to explore and analyse visual data. Advances in image recognition and retrieval allow users to discover connections, find similar materials, and engage with collections in entirely new ways, improving both the efficiency and depth of exploration.

For instance, imgs.ai is a text-based visual search engine designed for digital collections. It uses approximate k-NN algorithms and integrates the OpenAI CLIP model to link textual queries with visually similar content, offering new ways to navigate collections. Similarly, MAKEN, a service developed by the National Library of Norway, uses object detection to identify and retrieve similar images from the library's digital collection. Thus providing users with yet another possibility of engaging with visual material. These examples show how Computer Vision is making it easier to access digital collections, helping users to find connections, and discover materials quickly and effortlessly.



PixPlot is a project by the digital humanities lab of Yale University Library, it uses CNNs to cluster similar images.

Accessibility

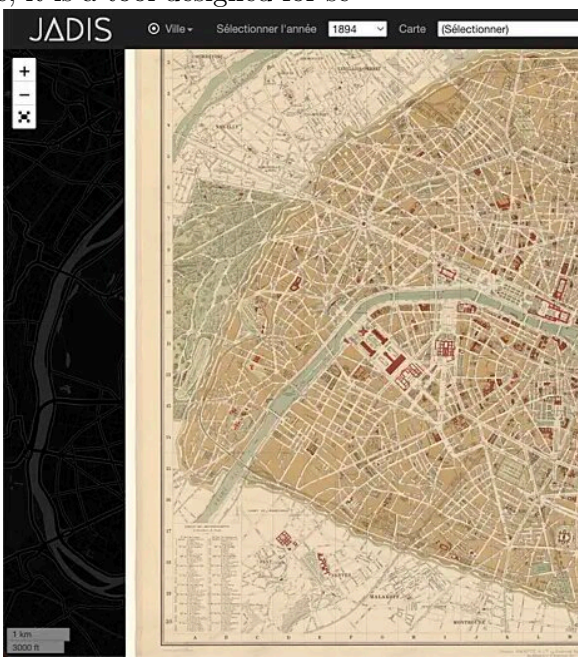
Computer Vision significantly improves accessibility in libraries by working alongside technologies like OCR and HTR. OCR/HTR convert printed and handwritten texts into machine-readable formats, enabling text-to-speech applications for users with visual impairments. Read more about OCR and HTR from Automatic Text Recognition (OCR/HTR) topic guide.

Beyond text, Computer Vision enables creation of descriptive metadata for visual materials, such as photographs, maps, or videos, thus providing richer contextual information that benefits users with cognitive or visual disabilities. For instance, a visually impaired user might hear a description of a painting that includes details about the objects and people depicted. This richer contextual information ensures that libraries remain inclusive spaces where all users can access and engage with collections in a meaningful way.

Semantic segmentation of maps

Semantic segmentation in an exciting frontier where Computer Vision can be of use for identifying and categorising distinct elements within maps, such as roads, buildings, water bodies, and land use regions. By using CNNs and U-Net architectures, models partition maps into meaningful segments, allowing for detailed analysis and improved usability. This supports historical research, geographic analysis, and so on, by extracting structured data from cartographic sources.

For example, JADIS, developed by the National Library of France, it is a tool designed for semantically segmented, georeferenced, and geocoded maps of Paris. Similarly, MapReader, an outcome of the Living with the Machines project, offers a Computer Vision pipeline for analysing large collections of historical maps.



Hands-on activity and other self-guided tutorial(s)

For those new to the field, Google Colab Notebooks offer a good starting point with plenty of notebooks on Computer Vision, one great example being this collection. The TensorFlow Playground allows users to experiment with neural network architectures in an interactive way, offering insight into the understanding of how those work. GLAM Workbench is a hands-on guide that focuses on applying Computer Vision to GLAM collections, mainly from New Zealand and Australia. It explores how digitised images can be analysed to identify patterns and improve the accessibility of digital heritage. For practical coding experience, versatile tutorials are found at OpenCV and RealPython. If a more humanities oriented point of view

is desired, then Computer Vision related tutorials can also be found at the Programming Historian.

Recommended Reading/Viewing

For those looking to get a deeper understanding of Computer Vision, some key resources include:

- Arnold, T., Tilton, L. (2023). *Distant viewing: Computational Exploration of Digital Images.* The MIT Press. https://doi.org/10.7551/mitpress/14046.001.0001
- Krizhevsky, A., Sutskever, I., Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS 2012).* Retrieved from https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
- Szeliski, T. (2022). *Computer Vision: Algorithms and Applications.* Springer. https://doi.org/10.1007/978-3-030-34372-9
- Stanford University's course CS231n Convolutional Neural Networks for Visual Recognition 2016 and 2017 lectures are up on YouTube.

Finding Communities of Practice

There are AI4LAM, CENL network group 'AI in libraries' and IFLA Artificial Intelligence Special Interest Group communities, while in the IIIF community there is AI/ML community group. There are also conferences like Fantastic Futures.

Top universities, for example Harvard, Stanford and Helsinki, offer, sometimes free, online courses. Also sites like Coursera, FutureLearn, and Udemy are worth a look. Open-source tools like OpenCV and TensorFlow mentioned above, but also PyTorch are handy.

One can join Slack workspaces of AI4LAM or IIIF.