أكاديمية طويق
Tuwaiq Academy

Literature Review
Computer Vision

ABSTRACT
This is our first assignment for the course *Introduction to Computer Vision in Python*. We read some research papers to help us understand the basics of computer vision and to learn about the latest advancements in the field. This helped us gain knowledge about how computer vision works and how it is used in real-world applications.

Contents

Introduction :

In recent years, the field of artificial intelligence has witnessed significant advancements, particularly in areas that aim to replicate human cognitive functions. One of the most prominent and rapidly evolving subfields is Computer Vision. This discipline focuses on enabling machines to interpret and understand visual information from the surrounding environment, simulating the way humans perceive and respond to images and videos.

Computer vision combines techniques from image processing, machine learning, and deep learning to develop systems capable of recognizing patterns, detecting objects, and making decisions based on visual input. With the increasing availability of large-scale data and powerful computational resources, the applications of computer vision have expanded across various domains, including healthcare, manufacturing, transportation, and security.

This research aims to explore the fundamental concepts of computer vision, its key technologies, real-world applications, and the challenges that continue to shape its development.

Importance of Computer Vision

The ability of machines to “see” and interpret their surroundings has unlocked a wide range of possibilities across various sectors. In healthcare, computer vision plays a vital role in disease detection through advanced medical imaging analysis. In the field of security, it enables real-time surveillance, facial recognition, and threat detection. In transportation, self-driving vehicles rely heavily on computer vision systems to recognize objects, navigate roads, and avoid obstacles — effectively mimicking human perception with high precision.

Computer Vision Challenges

(Summary)

Computer vision is advancing rapidly, but it still faces key challenges such as ensuring data quality, reducing the effort of data labelling, processing large amounts of data efficiently, and addressing ethical and privacy concerns.

On the technical side, some systems suffer from poor GPU utilization due to:

- Slow data transfer between CPU and GPU.
- Tasks not suited for parallel processing.
- Memory bottlenecks and synchronization delays.
- Uneven distribution of tasks across GPU cores.

Improving memory access, using asynchronous data transfer, and optimizing parallel processing can help solve these issues and boost performance.

Methodology and techniques

Object detection techniques have evolved from traditional methods that relied on manual feature extraction like SIFT and HOG. With the advancement of deep learning techniques, Convolutional Neural Networks (CNNs) emerged, significantly improving performance by automatically extracting features from images. These advancements have led to a substantial improvement in the accuracy and efficiency of object detection systems across various applications.

Traditional object detection methods followed a structured pipeline consisting of three main steps. First, potential object regions were identified using techniques like the Sliding Window method, which involved scanning the image at multiple scales and locations. Second, hand-crafted feature extraction techniques such as SIFT (Scale-Invariant Feature Transform) and HOG (Histogram of Oriented Gradients) were used to represent the image regions. Third, machine learning classifiers like SVM (Support Vector Machine) and AdaBoost were applied to classify the extracted features .While effective in simpler scenarios, these methods struggled in the presence of complex backgrounds, varying object scales and poses, and required intensive computational resources, especially with large-scale datasets.

Some of the most well-known traditional algorithms include the Viola-Jones Detector (VJ) and Histograms of Oriented Gradients (HOG). The VJ Detector was the first widely adopted algorithm for face detection, relying on a cascade classifier to locate faces. It performed well under constrained conditions but failed when exposed to lighting variations or diverse facial expressions. HOG, on the other hand, extracted edge and gradient features from small image patches and was primarily used for pedestrian detection. Despite its initial success, HOG lacked robustness when applied to more complex object categories and real-world environments. These limitations, including high false positive rates and redundant proposals, highlighted the need for more scalable and accurate methods, eventually paving the way for the adoption of deep learning.

With the advancement of deep learning, object detection became more efficient and accurate. The introduction of Convolutional Neural Networks (CNNs) revolutionized object detection by enabling automatic feature extraction and end-to-end learning. Deep learning-based object detectors are divided into two main categories: two-stage detectors, which are highly accurate but slower, and one-stage detectors, which are faster and suitable for real-time applications.

Two-stage detectors operate in two distinct steps. First, they generate region proposals — areas of the image that are likely to contain objects. Then, these proposals are passed through a classifier to determine object categories and refine bounding box coordinates. One of the earliest models in this category is Region-Based Convolutional Neural Network (R-CNN), which is a special type of Convolutional Neural Network. R-CNN works by extracting around 2000 candidate regions from an image using a technique called Selective Search. A huge number of regions are individually warped, and each proposed region is resized and passed through a CNN to extract features. These features are then classified using traditional classifiers such as SVM. Despite its accuracy, R-CNN suffered from slow training and testing, and Selective Search was inefficient in complex scenarios.

To address the inefficiencies in R-CNN, researchers introduced several improvements. Fast R-CNN incorporated a technique called RoI Pooling, which allowed the entire image to be processed in one pass, greatly reducing computational cost. Faster R-CNN further enhanced the architecture by introducing the Region Proposal Network (RPN), which generates region proposals internally within the model itself, eliminating the need for Selective Search. These innovations significantly improved processing speed and accuracy, making Faster R-CNN one of the most widely used object detection models. However, despite its advantages, it still falls short in real-time performance due to its two-stage nature.

In contrast, one-stage detectors eliminate the region proposal step entirely and predict object locations and classes in a single forward pass. This approach makes them much faster and more suitable for real-time applications. One of the most popular one-stage detectors is YOLO (You Only Look Once. YOLO divides the input image into a fixed grid, where each grid cell is responsible for predicting bounding boxes and class probabilities. This method allows YOLO to detect multiple objects in a single evaluation of the network, offering both speed and simplicity. [3]. YOLO excels in scenarios requiring fast inference, such as autonomous driving or real-time surveillance. It reduces errors and improves accuracy and speed compared to traditional methods. However, it has some limitations, particularly in detecting small or overlapping objects, which can lead to lower performance in complex scenes. Another prominent one-stage detector is SSD (Single Shot MultiBox Detector). SSD improves upon YOLO by using multiple feature maps to detect objects at different scales, which enhances the detection of objects of varying sizes. SSD offers a better balance between speed and accuracy, although it also struggles with smaller targets .

The evolution of object detection has shifted from traditional handcrafted techniques to powerful, data-driven deep learning models. While traditional methods laid the foundation, their limitations in scalability, efficiency, and accuracy led to the development of CNN-based approaches. Among these, two-stage models offer superior accuracy and are ideal for high-precision applications, while one-stage models provide the speed necessary for real-time use. Choosing the right technique depends on the specific needs of the application, whether speed, accuracy, or balance is the priority.

Major findings and trends

Vision Transformers :

Imagine you have a big picture to understand. **CNNs** look at this picture through small windows, focusing on tiny details like edges and shapes, then gradually combine these details layer by layer to see the whole image. On the other hand, **Vision Transformers (VTs)** cut the picture into many small patches and let every patch "talk" to every other patch all at once, so the model understands how all parts connect globally right away. Unlike **clustering**, which just groups similar things together, transformers and CNNs learn to recognize and understand images by exploring features and relationships. This way, CNNs are great at spotting local patterns quickly, while transformers excel at grasping the big picture all at once.

Transformers work by first slicing an image into small pieces, turning each piece into a list of numbers, and adding information about where each piece belongs. Then, using a clever process called **self-attention**, every piece compares itself to all the others to find important connections. A special summary piece gathers all this information and helps the model decide what the image shows—whether it's a dog, a car, or anything else. This ability to look at the whole image globally makes transformers very powerful for many computer vision tasks.

Transformers were first made to understand language like reading and writing but now they're used to understand pictures too. They start by cutting a big image into many small square patches, for example, a 224×224 image sliced into 16×16 patches makes 196 pieces. Each patch is then turned into a list of numbers that describes what's inside it. Because transformers don't naturally know where these patches belong in the image, extra information called positional embeddings is added so the model remembers their places. The key magic happens in the self-attention step, where every patch looks at all the other patches and figures out which ones are important to understand the whole picture—like a big group chat where all the puzzle pieces talk to each other. Finally, a special summary piece called the classification token gathers all this information and helps the model decide what the entire image shows, whether it's a dog, a car, or a tree.

Quantum ML :

QIANets uses a blend of quantum-inspired optimization techniques adapted for CNNs. In quantum-inspired pruning, researchers applied a probabilistic algorithm to determine which weights were essential for accurate performance and which could be removed. Tensor decomposition — or breaking down large, complex data grids, or tensors into simpler parts — followed, reducing the size of weight tensors through aforementioned SVD by selecting a limited number of singular values for each layer in the CNN. The annealing-based matrix factorization was then used to factor weight matrices into two lower-dimensional matrices, simulating a low-energy optimization state typical in quantum annealing. This iterative process minimized the difference between the original weights and their simplified forms, ensuring minimal data loss.

Testing was conducted on DenseNet, GoogLeNet, and ResNet-18 across a limited number of trials. The goal was to quantify the benefits of quantum-inspired techniques for CNNs in terms of latency and accuracy. Researchers indicated that the framework's performance was most promising in controlled settings, with low variations in test conditions.

Applications:

Computer vision has broad applications in modern technology and industries. Its use has spread widely across multiple sectors, including healthcare, security, manufacturing, and entertainment. Here, we discuss the main applications of computer vision and explain the key technologies that support its development and practical application.

Healthcare: With the increase in diseases, computer vision has proven its effectiveness in assisting doctors in various medical fields such as dentistry, cardiology, and dermatology.

Dentistry: Automated X-ray analysis to detect cavities, nerve infections, and gum disease.

Dermatology: Identifying skin conditions and moles for early detection of skin cancer.

Remote patient monitoring: Monitoring and analyzing vital signs, movement, and behavior of patients at home or in hospitals using cameras and sensors.

Security: With the increasing need for surveillance to reduce crime, computer vision has become an essential tool in modern security systems. It enhances surveillance accuracy and speeds up criminal detection, such as:

Facial recognition: Automatically identifying individuals through surveillance cameras to detect known criminals.

Intrusion detection: Automatically detecting unusual movements or unauthorized entry into restricted areas.

Crowd monitoring: Analyzing large gatherings such as events to detect abnormal behavior, fights, or potential threats.

Object detection: Identifying weapons or unattended bags in airports, shopping malls, and other public places.

Education: Computer vision is used to improve learning experiences and monitor classroom environments more effectively.

Student monitoring: Monitors student attention, distraction, or absence using facial expression tracking during classes.

Smart attendance systems: Automatically recognize students' faces and record their attendance without the need for manual data entry.

Exam proctoring: Monitors cheating behaviors, such as looking away, using devices, or using multiple devices during remote exams.

Manufacturing : Computer vision plays a vital role in improving efficiency, safety, and accuracy in industrial settings. It is widely used in several key areas, including:

Quality Control: Automatically detecting product defects during the manufacturing process, such as missing parts, surface damage, or assembly errors.

Robotic Automation: Directing industrial robots to perform tasks such as picking and placing items, assembling components, and performing visual inspections.

Predictive Maintenance: Monitoring the health of machinery using visual data to predict potential failures and schedule maintenance before they occur, reducing downtime and costs.

Process Automation: Real-time tracking of machinery and worker movements to ensure smooth and efficient operations.

Conclusion:

In conclusion, computer vision stands as one of the most transformative branches of artificial intelligence, bridging the gap between human perception and machine intelligence. This research has explored the foundational techniques, from traditional feature-based methods to advanced deep learning models such as CNNs, YOLO, and Vision Transformers, which have revolutionized how machines understand and interpret visual data.

The study highlighted the evolution of object detection, the rise of transformer-based models, and emerging trends such as quantum-inspired machine learning, reflecting the rapid and continuous innovation in the field. Furthermore, we examined the diverse real-world applications of computer vision, spanning healthcare, security, education, and manufacturing — each demonstrating the immense potential and practical impact of visual intelligence systems.

Despite its advancements, computer vision still faces challenges related to data processing, ethical concerns, and real-time efficiency. However, ongoing research and technological improvements are steadily addressing these obstacles, paving the way for more robust and scalable solutions.

Ultimately, as computer vision technologies continue to evolve, they promise to reshape the future of industries and enhance everyday life, making machines not only smarter — but more perceptive, responsive, and human-like in understanding the world around them.

Reference

1. IBM. (2021, July 27). *What is computer vision*. IBM. https://www.ibm.com/think/topics/computer-vision
2. OpenCV. (2024, February 14). *Your 2025 guide to the top 6 computer vision problems*. https://opencv.org/blog/computer-vision-problems/#Common-Computer-Vision-Problems
3. SciForce. (2024, March 8). *Top computer vision opportunities and challenges for 2024*. *Medium*. https://medium.com/sciforce/top-computer-vision-opportunities-and-challenges-for-2024-31a238cb9ff2
4. Ambika199820. (2023, September 18). *What is computer vision? (History, applications, challenges)*. *Medium*. https://medium.com/@ambika199820/what-is-computer-vision-history-applications-challenges-13f5759b48a5
5. The Quantum Insider. (2024, October 30). *Quantum-inspired techniques cut latency in computer vision without sacrificing accuracy*. https://thequantuminsider.com/2024/10/30/quantum-inspired-techniques-cut-latency-in-computer-vision-without-sacrificing-accuracy

6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929. https://arxiv.org/abs/2010.11929