King Saud University
College of Computer and Information Sciences
Computer Science Department

AI with Computer Vision: A Literature Review

Prepared By:

Nouf Alsubaie	**441200911**
Fatimah Alhumaidhi	**441200921**
Layan Alarifi	**441201121**

Supervised By:
T. Nouf Alshenaifi

First Semester 1444/2022

I Introduction

Artificial intelligence (AI) is a field of computer science that concentrates on making intelligent agents that are capable of thinking and reasoning rationally to solve different problems [1]. Artificial intelligence is inter-related with other subfields like machine learning and deep learning. Nowadays, AI became a wide area of research, it includes the study of many important topics like natural language processing (NLP), expert systems, speech recognition and computer vision. AI has many useful applications in real life and it provides effective solutions in business, transportation, health care, and other fields.

In 1959, Arthur Samuel created the subject of study known as machine learning (ML), a subfield of computer science that enables computers to learn from experience without being explicitly programmed [2]. This concept can also refer to the procedure of gathering data and algorithmically creating a statistical model based on that data in order to solve a real-world problem [3]. Computers can now complete tasks like (classification, clustering, prediction, pattern recognition, etc.) thanks to the development of ML. ML algorithms can be categorized into four groups based on the learning method: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

In supervised learning, the machine learns how to map the input into output after observing a number of examples that explicitly show what is the correct output for each input [1]. In practice, this means that supervised learning needs labeled datasets in order to train the model. While in unsupervised learning the machine recognizes patterns in the input data and draws conclusions without explicit feedback given to the machine [1]. Thus, in an unsupervised approach, there is no need to provide labeled datasets.

Computer Vision is a field that is concerned with acquiring, analyzing, and processing digital images and videos via computers. Computer vision first emerged in the early 1970s and was thought of as the visual perception part of an ambitious plan to imitate human intelligence and give machines intelligent behavior. The desire to retrieve three-dimensional structure from images with the aim of obtaining a full scene understanding distinguished computer vision from the predominant field of digital image processing [4].

Applications of computer vision in artificial intelligence (AI) and machine learning (ML) have received a great deal of attention in recent years as a result of recent advancements in computer hardware and parallel processing, some of computer vision applications in AI include: Object recognition, Image Generation and Image Classification. In this review, we talk about each of these applications in depth, showing how they are applied and used, and when were they invented. We also talk about the latest technologies used in the field and explain how they work.

II Literature Review

1 Object Recognition

Recognition is a visual ability that includes identification and discrimination between different objects [5]. The term object recognition in computer vision refers to the task of identifying an object and categorizing it as an instance of a certain class [5]. The goal is to find a correspondence between some of the image features with similar features in the object model [6]. The main problem in object recognition is to determine which features to use and how to obtain the correspondence between the image and the model.

Based on the available information about the object, there are two types of object recognition; model-based object recognition and view-based object recognition. In the first type, a 3D model is available representing detailed information about the object parts, shape and how different parts are combined. On the other hand, the view-based recognition lacks a fully described 3D model, instead, it maintains some information about how the object looks from different view angles and uses these data to recognize the object [6].

Before the rise of deep learning, scientists had to apply classical machine learning approaches to extract the image features manually before training the model [7]. This required researchers to have extensive domain knowledge in order to design a detection algorithm that looks for predefined features in an image, yet these models do not achieve the required accuracy and are still vulnerable to mistakes if the image size or resolution has changed [8]. However, the use of deep learning models has led to great improvement in object recognition.

The main advantage of deep learning models over traditional learning models is that in deep learning the algorithm can learn from the data itself without human supervision. This is done by using highly connected parallel networks known as neural networks which are inspired from the human nervous system and brain. The neural network consists of multiple layers. In each layer there are a number of nodes that perform a certain processing or computational operation, each node has dense interconnections with other nodes in the network [6].

Since the deep learning models are data-driven, we no longer need to extract features manually [6]. The model can automatically extract the features by processing the images and recognizing the pixel pattern that made up a certain object. In this way, the model can learn by itself which features to look for to recognize the object. There are many techniques and models in deep learning, but some of them have shown significant improvement when used in computer vision applications. Convolutional Neural Networks (CNNs) and Deep Belief Networks (DBNs) are widely used in the computer vision field [9].

Convolutional neural networks(CNNs) consist of three types of neural layers: (i) convolutional layers, (ii) pooling layers, and (iii) fully connected layers, each type of layers is responsible for different operation [9]; The convolutional layer is what makes CNNs different from typical neural networks, it uses special image filters called kernels to extract image features and generating the feature maps, after the first convolutional layer only low-level features like edges and corners are extracted, then the output is fed again into another convolutional layer to extract higher-level features. The final reasoning is done by the fully connected layers that follow several convolutional and pooling layers [9]. The CNN model is known for its capability in feature learning, However, a major drawback of CNN is that it relies on the existence of a labeled data set.

Many works on object recognition apply CNNs. In a study conducted by Matija Radovic et al. [10], a convolutional neural network model is trained on a set of aerial images including real-time video feed from unmanned aerial vehicles (UAVs). Due to the high speed of the UAV environment, the proposed model uses 26 convolutional layers instead of 24-configuration to increase the accuracy and the speed of the detection. the results show that the CNN was able to recognize “airplane” objects in the data set with 97.5% of accuracy while only 16 instances were incorrectly categorized (Figure 1).

Another research [11] uses convolutional

Figure 1: . (a) Yellow arrow points at the instances where "airplane" object is present but not detected by the CNN, (b) red arrow points to the instance where "airplane" object is wrongly identified.[10].

neural networks in the medical field to recognize anatomical objects in surgical real-time videos. The dataset videos used in the research were recorded during different surgeries carried out at a medical university in Tokyo. The model was trained to recognize different anatomical objects like blood, vessels, uterus, forceps, ports, gauze and clips in the surgical images. After training the model has achieved an accuracy of 83% and precision of 80%. which proves the efficiency of using CNNs in object detection.

Deep Belief Networks (DBNs) is a generative model structured by stacking multiple restricted Boltzmann machines (RBM) and training them in a greedy manner [9]. The RBM is an undirected graphical model containing a set of stochastic visible variables and hidden variables such that each visible variable is connected to each hidden variable with a restriction that the connection between them must form a bipartite graph [9]. The DBNs use RBM to initialize the deep network and then train the model using a greedy strategy. The main advantage of the above structure is that it solves the problem of selecting appropriate parameters and does not require labeled data since the process is unsupervised [9].

As an example of using DBNs for object recognition; There is a study conducted to detect manipulation in satellite images using deep belief networks [12]. The manipulation includes splicing images from different sources to generate manipulated images. The researchers train and test four different DBNs with the MINST data set. And found that the proposed model was able to detect and localize manipulation even with small forgeries [12].

2 Image Generation

Image generation tasks are tasks where the goal is auto-generation of digital images or videos by a deep learning model. For this task, a network called Generative Adversarial Network (GAN) was first introduced in a 2014 paper by Goodfellow et al. [13]. Generative Adversarial Networks are semi-supervised deep learning models that consist of two parts: a generative model and an adversary model, the generative model generates samples by passing random noise through a multilayer network, which then send the output to the adversary model. The adversary model is a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution [13]. this framework was based on the two-player minimax game with the following objective function:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim pdata(x)}[\log D(x)] + \mathbb{E}_{z\sim pz(z)}[\log(1 - D(G(z)))]$$

where the discriminator D tries to maximize the probability of assigning the correct label to both training examples and samples from the generator, and the generator G tries to minimize the correct labels [13].

GANs showed great success in image generation tasks. Resent applications of image generation is concerned with generating images from text description, Zhang et al. [14] used a model called Stacked Generative Ad-

versarial Network (StackGAN) to Synthesize lifelike images from English text description. Even with using deep learning models like GAN, generating images from text description is a hard task for a computer to process. In [14], the problem was divided into 2 smaller stages, where the first-stage GAN model creates the fundamental shape and colors of the object based on the provided written description, producing low-resolution pictures. The output of the first stage with the written description serve as inputs for the second stage of GAN, which creates high-resolution images with realistic looking qualities. With the refinement process, it is able to fix flaws in first-stage results and add interesting details to the image. Figure 2 shows examples of some generated images by GAN models

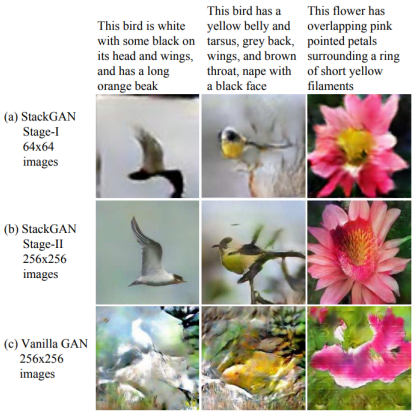

Figure 2: Comparison of StackGAN and a vanillaone-stage GAN for generating images from text [14].

deep generative models can be applied to many real-life tasks, a 2016 paper by Goodfellow [15] mentioned some of them:

- Super-resolution of a single image: The objective of this task is to create a high-resolution equivalent from a low-resolution image.
- Tasks where the goal is to create art.
- Applications of image-to-image translation can transform sketches into graphics or aerial photos into maps.

Other important applications of generative models might include generating dataset for training other deep learning models. Deep generative models can be categorized according to how they can learn. Models that can learn using the concept of maximum likelihood have different representations or approximations of the likelihood. Figure 3 shows a taxonomic tree of generative models that can learn using the concept of maximum likelihood.

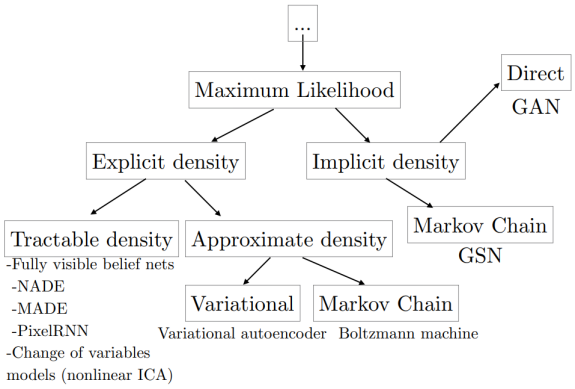

Figure 3: Taxonomic tree of generative models [15].

3 Image classification

Image classification is a key task in computer vision that aims to annotate images with specific labels that describe their semantic information [16]. The semantic information is extracted from the underlying features of the image. This is done by training a classifier to learn from a data set and classify images using a specific classification algorithm [16].

Image classification isolates the characteristic information in the image from color or

shape and does not care, for example, about a flat background or noise. AlexNet, VGG-16, GoogleNet, ResNet, Inception-V3 and DenseNet are all common neural networks models that can be used for image classification [17].

Image classification can be divided into two problems: binary classification and multiclass classification. Binary classification involves designating an input image into one of two classes, while in multiclass classification, two or more classes are included. A classic example of a binary image classification problem is identifying cats or dogs in each input image. Paper [17] also mentioned that deeper networks and highly interconnected networks showed an evolution in the performance of image classification tasks, and with the progress, Image classification became of great importance for the developed dataset such as (SVHN, MNIST, CALTECH-101, CALTECH-256) It is noteworthy that most of the datasets are color where as stated in the paper [17] "These color images are represented in RGB format. Computer, these images are just numbers and have no inherent meaning in meaning. Most modern models developed for classification do not convert the color space to the image and instead use the RGB image directly for classification".

4 Face Recognition

There are many technologies that help in identifying a person in recent years, for example, using fingers and hands for identification, but relying on this technique may not always be useful because it may be exposed to bruises or cracks, which hinders obtaining the desired results, and when using the eye through iris and retina techniques, it requires expensive equipment, and is very sensitive to any movement of the body, while the sound is also not always appropriate because it is subject to noise in public places, but the face images can be obtained easily using a pair of fixed cheap cameras. It doesn't take a lot of people to use the technology, unlike the above techniques [18].

Face recognition can be used for two tasks: 1- Verification, which is matching from one person to another, 2- Identification, and one-to-many matching, i.e. displaying an image of an unknown person and comparing it with a known database until that person is identified. They are applied and used for security, surveillance, public identity verification, image database investigations, and much more. Speaking of the difficulties with face recognition technology, differences in age, differences in lighting, and differences in pose are the main problems it suffers from [18].

Automatic face recognition in its early days was feature-based and used to measure distinctive facial features such as eyes, mouth, nose, etc., and then calculate geometric relationships between facial points, resulting in reduction of the input face image to a vector of geometric features, and then the face is recognized using standard statistical pattern recognition techniques [18].

Modern statistical methods and artificial intelligence methods attempt to identify faces based on the entire image rather than the local features of the face. Statistical methods obtain the results by direct correlation comparisons between the input face and all other faces in the dataset. On the other hand, AI methods use tools such as neural networks and machine learning techniques to recognize faces. The results were satisfactory and good as they were in equal lighting, size and appropriate setting, but it was computationally expensive, and its other drawback is that it is sensitive to changing light and noise and the direction of the face (Figure 4) [18].

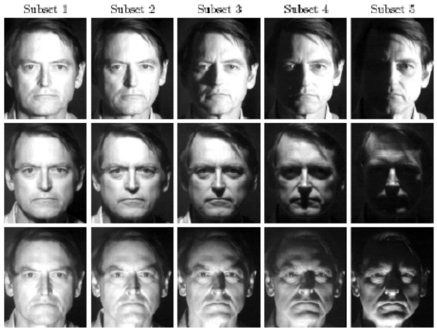

Figure 4: The same person seen under varying light conditions can appear dramatically different [18].

III Discussion

Advances in artificial intelligence over the last few years have greatly improved the field of computer vision. As we reviewed the latest works in the computer vision field, we observed that deep learning models are widely used to perform different computer vision tasks like object recognition, image generation, and image classification.

In the past, scientists had to manually extract the image features before training their models using classical machine learning approaches, These models were susceptible to errors and highly affected by changes in the image size and resolution, So they failed to deliver the required accuracy. until the development of neural network models allows for more accurate performance.

In object recognition, many recent studies use convolutional neural networks (CNNs) and deep belief networks (DBNs) as the main methodology to recognize objects in real-life applications. CNNs are more commonly used when compared to DBNs. CNN approach is preferred for its unique capability in learning features and generating feature maps. However, the CNN model relies on the existence of a labeled data set in contrast to the DBN model which is unsupervised and can work with unlabeled data sets.

It is generally agreed that deep learning models can yield good results and achieve high accuracy, but the main limitation is the time complexity and the computational resources needed to train them. Since all data are interconnected the size and computation increase exponentially. Some researchers suggested removing connections between certain nodes and proposed other variations of the models to reduce the complexity [12].

In addition, we note that most works on object recognition and image classification are usually an improvement on existing models. This improvement is in terms of increasing the accuracy or reducing the complexity of the model. While the image generation field is still inventing new models that perform hard tasks such as text-to-image and text-to-video generation.

IV Conclusion

The invention of deep neural networks has greatly improved our capability to perform computer vision tasks. Convolutional neural networks (CNNs) and deep belief networks (DBNs) are two NNs models that are widely used in the literature and have proven to be useful in many real-life applications. However, The exponential complexity and the high demand for computational resources are still unsolved limitations.

Although computer vision is a well-established field with many successful studies, There are still some hard tasks that need to be studied and improved. Generative Adversarial Network (GAN) is a good example of a newly invented model that shows great success in image to text generation.

References

[1] S. Russell and P. Norvig, *Artificial intelligence*. Upper Saddle River, NJ: Pearson, 4 ed., Nov. 2020.

[2] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Dev.*, vol. 3, pp. 210–229, July 1959.

[3] A. Burkov, *The hundred-page machine learning book*. Andriy Burkov, Jan. 2019.

[4] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer London, second ed., 2011.

[5] J. C. Liter and H. H. Bülthoff, "An introduction to object recognition," *Z. Naturforsch. C*, vol. 53, pp. 610–621, July 1998.

[6] R. Mutihac, "Paradigms in object recognition," 2005.

[7] V. Varadarajan, D. Garg, and K. Kotecha, "An efficient deep convolutional neural network approach for object detection and recognition using a multi-scale anchor box in real-time," *Future internet*, vol. 13, p. 307, Nov. 2021.

[8] S. Goyal and P. Benjamin, "Object recognition using deep neural networks: A survey," Dec. 2014.

[9] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, p. 7068349, Feb. 2018.

[10] M. Radovic, O. Adarkwa, and Q. Wang, "Object recognition in aerial images using convolutional neural networks," *J. Imaging*, vol. 3, p. 21, June 2017.

[11] Y. Bamba, S. Ogawa, M. Itabashi, H. Shindo, S. Kameoka, T. Okamoto, and M. Yamamoto, "Object and anatomical feature recognition in surgical video images based on a convolutional neural network," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, pp. 2045–2054, Nov. 2021.

[12] J. Horváth, D. M. Montserrat, H. Hao, and E. J. Delp, "Manipulation detection in satellite images using deep belief networks," Apr. 2020.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv*, June 2014.

[14] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," *arXiv*, Dec. 2016.

[15] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," 2017.

[16] J. Wu, V. S. Sheng, J. Zhang, H. Li, T. Dadakova, C. L. Swisher, Z. Cui, and P. Zhao, "Multi-label active learning algorithms for image classification: Overview and future promise," *ACM Comput. Surv.*, vol. 53, pp. 1–35, June 2020.

[17] S. N. Gowda and C. Yuan, "ColorNet: Investigating the importance of color spaces for image classification," Feb. 2019.

[18] JafriRabia and A. R., "A survey of face recognition techniques," *Journal of Information Processing Systems*, vol. 5, pp. 41–68, 06 2009.