UDC 004.8+004.93'1 DOI: 10.31673/2412-4338.2025.011767

Anatolii Ivanov
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv
ORCID 0009-0006-5370-445X

A COMPREHENSIVE REVIEW OF THE HISTORY AND METHODS OF COMPUTER VISION

***Abstract**: Extended abstract: The purpose of this article is to explore the history, methods, applications, and usage of computer vision in modern society. It aims to find and describe the potential of computer vision and its pivotal role in shaping the future of technology and human interaction with the visual world. Article describes main methodologies of computer vision, including convolutional neural networks (CNNs), deep belief networks (DBNs), and autoencoders, describing their principles, functionalities, and main usages in image recognition and analysis. Through an in-depth exploration of methodologies and real-world applications, the article highlights the impact of computer vision across diverse domains. It discusses the usage of computer vision technologies through different industries such as autonomous vehicles, healthcare, security, augmented reality, and industrial automation.*

As a result, this article provides a comprehensive analysis of existing computer vision algorithms, describing their benefits and drawbacks in detail. It further explores the primary applications of these algorithms across various domains, illustrating the specific tasks for which each algorithm is optimally suited. In addition, the article delves into case studies and examples that highlight the practical implementations of these algorithms. Looking ahead, the future of computer vision holds the promise of groundbreaking advancements and innovative developments. These technologies are aimed to fundamentally transform methods of perceiving, interacting with, and comprehending the visual world, paving the way for unprecedented changes in numerous fields. The potential impact of these advancements suggests a significant shift in both the technological landscape and our daily lives, creating a new era of visual data processing and interpretation.

***Keywords**: computer vision, image recognition, machine learning, deep learning, artificial intelligence, neural networks.*

Іванов Анатолій Ігорович
Національний технічний університет «Київський Політехнічний Інститут ім. Ігоря Сікорського», м. Київ
ORCID 0009-0006-5370-445X

ВСЕБІЧНИЙ ОГЛЯД ІСТОРІЇ ТА МЕТОДІВ КОМП'ЮТЕРНОГО ЗОРУ

***Анотація.** Метою даної статті є дослідження історії, методів, застосувань і використання комп'ютерного зору в сучасному суспільстві. У ній аналізується потенціал комп'ютерного зору та його ключова роль у формуванні майбутнього технологій, зокрема взаємодії людини з візуальним світом. У статті описано основні методології комп'ютерного зору, зокрема згорткові нейронні мережі (CNN), мережі глибоких вірувань (DBN) та автокодери, пояснюються їхні принципи, функціональність і основні способи використання в розпізнаванні й аналізі зображень.*

Завдяки детальному аналізу методів і реальних прикладів застосування стаття акцентує увагу на впливі комп'ютерного зору у різних сферах. Зокрема, розглядається використання цих технологій у таких галузях, як автономний транспорт, охорона здоров'я, системи безпеки, доповнена реальність та промислова автоматизація. Наприклад, у сфері автономних транспортних засобів комп'ютерний зір сприяє створенню більш безпечних автомобілів, здатних зменшити кількість дорожньо-транспортних пригод. У медицині ці технології дозволяють проводити точний аналіз знімків для раннього виявлення захворювань, що допомагає зберігати життя пацієнтів.

У статті також надається всебічний аналіз існуючих алгоритмів комп'ютерного зору, детально описуються їхні переваги та недоліки. Розглядаються конкретні завдання, для яких кожен алгоритм підходить найкраще, а також наводяться приклади їх практичного застосування.

Майбутнє комп'ютерного зору відкриває нові горизонти, обіцяючи революційні досягнення та інноваційні розробки. Ці технології здатні змінити підходи до сприйняття, взаємодії та розуміння візуального світу, сприяючи значним змінам у багатьох сферах. Очікується, що ці досягнення вплинуть не лише на технологічне середовище, а й на повсякденне життя, відкриваючи нову еру обробки та інтерпретації візуальних даних.

Ключові слова: *комп'ютерний зір, розпізнавання зображень, машинне навчання, глибоке навчання, штучний інтелект, нейронні мережі.*

Introduction. In an increasingly digitized world, the ability to interpret and understand visual information is getting more critical. Computer vision is a main technology that is in the intersection of artificial intelligence, computer science, and cognitive psychology. At its core, computer vision tried to give machines an ability to perceive, analyze, and interpret visual data in a manner like human vision.

At the heart of computer vision lies image recognition, a fundamental task wherein computers find and classify objects, scenes, and patterns within digital photos. This process consists of extracting meaningful features from raw pixel data, building spatial relationships, and making high-level inferences – something that previously was possible only for the human eye.

The growing popularity and real-world relevance of computer vision creates its uses diverse different domains. From enhancing the capabilities of autonomous vehicles to revolutionizing medical diagnostics, the transformative potential of computer vision can be far-reaching and profound. In the realm of autonomous vehicles, computer vision plays a pivotal role in enabling vehicles to navigate complex environments autonomously. By enabling real-time image analysis and object detection algorithms, self-driving cars can find pedestrians, recognize traffic signs, and predict potential hazards – creating a new way of autonomous driving characterized by safety, efficiency, and convenience.

In healthcare, computer vision takes its place as a tool which revolutionized medical diagnostics and patient care. It allows us to find anomalies in medical images and aid in surgical procedures. Computer vision technologies empower healthcare professionals with invaluable insights and diagnostic capabilities, leading to more accurate diagnoses, personalized treatment plans, and improved patient outcomes.

Moreover, integration of computer vision technologies in security and surveillance systems underscores its indispensable role in safeguarding public safety and national security. Facial recognition algorithms enable law enforcement agencies to identify and track suspects, enhance border security, and prevent criminal activities – giving access to new surveillance methods with advanced analytics and real-time monitoring capabilities.

Beyond traditional applications, computer vision is driving innovation in fields such as augmented reality, industrial automation, retail analytics, and more by reshaping industries, transforming business models, and unlocking new opportunities for growth and advancement.

In essence, as computer vision becomes more common and important, it gives new possibilities and can change how we interpret different objects and processes in the digital world. And as progress never stops, as scientists and inventors are always trying new things, computer vision will keep getting better and better. This means it will keep making breakthroughs and reshape the way we see and interact with the world around us.

Problem formulation. Despite significant advancements in computer vision, its potential to improve people's life with visual impairments has yet to be fully realized. Therefore, creating systems that allow such people to interact with certain elements of their environment without relying on others is a significant problem that requires new approaches.

The relevance of this topic is driven by the need for both theoretical justification of existing approaches and the development of new, more accurate, and adaptive image recognition methods. Addressing these challenges opens up possibilities for developing both the fundamental aspects of computer vision and its practical applications in creating innovative assistive systems for people with visual impairments.

This article aims to make a first step to conduct this research and thus collects main information about computer vision to help select the best suited methods to partially solve problem of navigation to help people with visual impairments.

The Aim and Objectives of the Study. This article conducts a comprehensive analysis of the evolution of different methods and approaches in computer vision sphere. Also, it identifies main methods used for different tasks, pointing their advantages, limitations, and main types of specific usages.

Its aim is to give a better overview and understanding of different methods and models, where they can be used and their principles of work. Its main objective is to analyze existing models and methods. Also it aims to help to select the best suited model for conducted research.

Generally, research will contribute to a deeper understanding of the state of the computer vision field and identify directions for its further development, which is important both for the scientific community and for developers of innovative solutions.

History. Computer vision and image recognition. Starting from 1960th, with a development of modern technologies raised a question of system creation, which can make human-like actions. However, at that moment of time, computing power was quite low, and computers weren’t able to process high volumes of data in real time. So, a new sphere of computing was created, which tried to optimize existing theories and create a new one on how image can be processed, which was named computer vision, part of which was image recognition.

So, its history starts from the year 1964, when the first attempt to correct the distortion of photographs, taken by satellite, was made. Till 1970, millions of such photos were processed and exactly that time first attempts to create an image comparison algorithm was made.

In parallel, there were tries to create first algorithms for improving the quality of images, which can change its contrast, remove blurring and highlight certain shapes, which can later greatly simplify the work of people in their analysis and further processing.

Even then it was clear that although these algorithms didn’t give the best results and their speed was still quite low, they can be used in the real world to process real-time data from cameras. And, so, everything started, as usual, from the military sphere. Highlighting specific parts of an image will allow for rockets to find some object on the terrain, such as a tank or a bunker. But, switching from 2D models to 3D models arose a question on how to work with it, as even the principles were the same, but the quality of existing algorithms gradually decreased and was unable to process data in given amount of time with acceptable results.

That leads to 1965 [1], when the first attempts in that sphere were made, especially for automatic objects detection of 3D objects in space. And, therefore, first algorithm was created. It converts images from a camera to linear format, which has only the contours of figures left. After that, for simplicity, only some basis geometrical figures were used ((parallelepiped, cube, triangular and hexagonal prism). During the process of recognition, figures were transformed via rotations, resizing and projections to fit the given photo. This process was repeated for all other figures until suitable ones were found, or all were checked. Even though this algorithm sounds quite simple and easy, assuming existing computer powers in 60s, it was a very important development, from which appeared following ones for two next decades.

At that time, all problems that computer vision was able to solve were limited to 2D space and the first attempts to deal with 3D objects took place. However, it gave an opportunity to create the first robotic mechanism, which was able to precisely determine their own position in space, their scenery and predict what will happen after certain movements [2].

In the 80s, with the development of the medicine sphere and discoveries in how our brain works, the first machine learning algorithms started to appear, which quickly took their place in computer vision sphere. They allowed machines to learn and improve their accuracy over time. This, in its turn, created a possibility to detect and recognize more complex objects than simple geometric shapes and forms.

After rapid development in 80s, computer vision sphere gets some decline in popularity until 2001, when Viols-Jones created his first face recognition algorithm, which created one of the biggest breakthroughs in this field at that time [3].

During the 2000s and 2010s deep learning algorithms were adapted and used in computer vision sphere, which created another revolution in it, enabling robotic machines to learn hierarchical representations of data. The appearance of convolutional neural networks (CNNs) and other deep learning algorithms has made it possible to recognize objects, track motion and perform other complex tasks with greater accuracy than ever before.

Pattern recognition. Going to narrower sphere of pattern recognition on 2D images, meaning finding specific parts of it, it’s worth mentioning what it is. Pattern was at first defined like a categorization or classification problem of input data, which happened via extraction important features from a lot of noisy data [4]. Nowadays it became a complete scientific discipline, whose aim is the classification of objects into a lot of categories or classes. Also, it’s used in many machine intelligence system build for decision making.

It’s history started in 1950s with template matching technologies, which allowed to find occurenct of a specific image (template) within the larger image (or the search image). It’s general work is based on sliding the template across the search image and comparing pixel values at each position. Thus, to compare this similarities and give result score, a number of methods were created, such as Correlation-Based Methods [yyy], Sum of Squared Differences [yyy] and Normalized Cross-Correlation which give different results depending on the image. It’s main limitations are problems with scaling, rotating and changing of view point, at the same time being computationally expensive for large images and templates. But, it still can be used in specific applications where simplicity and speed are prioritised.

During 1960-1970s, such technology as feature extraction was first impletented and used It is one of the cornerstones of computer vision, and has evolved significantly over the decades. From simple edge detection techniques to sophisticated deep learning models, this field has witnessed remarkable progress. So, during these decade, researchers has began exploring ways to extract meaningful information from images. New tecniques like edge detection were developed to identify boundaries and contours within images. These methods, while basic, laid the foundation for more advanced feature extraction techniques.

During 1980s and 1990s the emergence of robust feature descriptors happened like Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) [yyy]. These descriptors could detect and describe local image features, such as corners and edges, in a way that was invariant to various image transformations, including changes in scale, rotation, and illumination, which were main drawbacks of first technologies. This made them highly valuable for tasks like object recognition, image stitching, and 3D reconstruction.

Another path of image pattern recognition techniques goes under the rise of statistical methods during 1970s and 1980s. With the creation of such methods as Statistical Pattern Recognition and Hidden Markov Models (HMMs), which helped to classify image based on feature distribution and find sequentioal patterns in images accordingly. HMMs were used particularly for such tasks as character recognition and object tracking.

Next great step of improvement happened in the 1980s with the creation of artificial neural networks (ANNs), inspired by the human brain's ability to learn and recognize patterns. During this time first Convolutional Neural Networks (CNNs), pioneered by Yann LeCun [yyy], emerged as a powerful tool for pattern recognition tasks. By employing convolutional layers, CNNs could efficiently extract spatial features directly from raw image data. However, due to computational limitations and the lack of large-scale datasets, the potential of CNNs was initially quite restricted.

During the 2010s this sphere witnessed a great shift with the rise of deep learning. Creation of powerful GPUs and massive datasets like ImageNet enabled researchers to train better, deeper and create more complex CNN architectures. Such models as AlexNet, VGG, and ResNet achieved groundbreaking results in image classification, object detection, and other visual tasks. These models could learn hierarchical representations of data, capturing intricate patterns and details that were previously inaccessible to traditional methods.

In conclusion, the field of pattern recognition has undergone a remarkable transformation, from simple template matching techniques to sophisticated deep learning models. While early methods relied on handcrafted features and statistical models, the advent of deep learning has revolutionized the field, enabling the extraction of intricate patterns and features directly from raw image data. As technology continues to advance, we can anticipate even more innovative approaches to image pattern recognition, with applications ranging from autonomous vehicles to medical image analysis.

Nowadays. Today, computer vision is a rapidly growing field with thousands of enthusiasts from different spheres and experts and it is used in different spheres of everyday life, such as autonomous vehicles, medicine, robotics. This technologies, are cornerstones of artificial intelligence, and are rapidly transforming various industries. By enabling machines to "see" and interpret visual information, this technology is unlocking new possibilities and driving innovation.

First of all, image recognition helped to revolutionize healthcare, as it can help analyzing medical images like X-rays, MRIs, and CT scans. Such systems can detect anomalies, assist in diagnosis, and improve treatment planning. Theoretically image recognition technologies has the potential to significantly improve patient outcomes and streamline healthcare processes.

Second sphere is transportation and industries, where it helped to build autonomous vehicles. This way, by equipping cars with advanced cameras and sensors, they can accurately analyze their surroundings, identify obstacles, and make real-time decisions. This way it has the potential to revolutionize transportation, making it safer, more efficient, and accessible. At the same time it allowed to improve quality of manufacturing, retail and security systems, thus increasing productivity and efficiency of them, cutting cost at the same time.

Some more spheres of improvements are gaming and education, where it allowed to incorporate augmented reality (AR) features and seamlessly blend virtual and real words allowing to play, learn and create world around people more efficiently.

As image recognition technology continues to advance, its impact on our lives will only grow. By unlocking the power of visual data, this technology is shaping the future of countless industries and driving innovation across the globe.

Existing methods and means of computer vision

First, talking about computer vision, it’s worth clearly defining what it is and in which areas it can be used. According to [5], computer vision is a combination of image processing methods, pattern recognition and artificial intelligence, which is focused on computer analysis of one or more images. These images can be taken either by one or multiple sensors in a moment or some period. This analysis recognizes and locates the position, orientation, type of an object and results in a detailed description of its location in space. This processing typically uses geometric modelling and complex mappings for pattern recognition and objects search. Although, different strategies can be used for this purpose.

That means that computer vision models include quite a complex network of elements for image and video analysis and locating all or only some specific objects. In some ways, this works like the human eye and in general tries to imitate its behavior.

The primary areas of computer vision usage are autonomous vehicles, face recognition technologies, AR and VR and healthcare.

First of all, they have emerged as powerful tools in the field of medical image analysis, revolutionizing the way healthcare professionals diagnose and treat diseases by automating the analysis of complex medical images, such as X-rays, MRIs, and CT scans, these technologies enable more accurate and efficient diagnoses. At the same time CNNs and deep learning algorithms have significantly improved the accuracy of medical image analysis. They allowerd to detect subtle abnormalities, such as tumors, fractures, and infections, that may be difficult for human radiologists to identify. At the same time, they excluded human factor and error, speeding up the diagnostic process, these technologies aids in early detection and timely intervention, leading to better patient outcomes.

Computer vision also plays a crucial role in enabling personalized medicine and telemedicine. Through image segmentation, physicians can precisely assess the size, shape, and development of lesions, tailoring treatment plans to individual patient needs. Additionally, these technologies can be used to track disease progression, monitor treatment response, and plan surgical procedures. Furthermore it enabled remote diagnosis and consultation. This happens by analyzing high-quality medical images, by which specialists can provide accurate diagnoses and treatment recommendations to patients located in remote areas. This technology expands access to healthcare, especially for underserved populations.

Another sphere of computer vision usage is self-driving cars where it plays the main role by enabling vehicles to "see" and interpret their surroundings. This way, it allows autonomous vehicles to navigate roads safely and efficiently. For this technology to work correctly, it uses three main stages. First one is perception, which happens when computer vision systems, equipped with cameras, lidar, and radar, capture visual data from the environment. This data is then processed to detect and classify objects such as pedestrians, cyclists, other vehicles, road signs, and traffic signals. During this step computer understands its surroundings. Second stage is decision making. During it advanced algorithms analyze visual input in real-time, allowing the vehicle to make decisions. It includes object detection, semantic segmentation, and depth estimation techniques, which help car to understand its surroundings, anticipate potential hazards, and respond appropriately. And the last, but not least is navigational intelligence, which enables self-driving cars to interpret road patterns, lanes, and traffic signals. It also allows to analyze complex scenarios, such as a pedestrian crossing the road or a cyclist turning, where these systems can navigate vehicle safely and efficiently in dynamic environments.

One more sphere is facial recognition, which is a powerful application of computer vision, that has changed the way of identification and authentication of individuals. This technology creates a digital "faceprint" that can be used for a variety of purposes by analyzing unique facial features such as eye spacing, nose structure, and jawline. One of the most prominent applications of facial recognition lies in the field of security and access control. This technology enables secure, password-free verification, making it ideal for unlocking smartphones, accessing secure facilities, and streamlining airport security procedures. With advancements in deep learning, facial recognition systems have achieved remarkable accuracy, even in challenging conditions like varying lighting, angles, or facial expressions.

Beyond security sphere, facial recognition has found applications in various industries like retailing, where it happens to identify customers, track their behavior, and personalize marketing strategies, social media, where it is used to tag individuals photos and healthcare, where it can be used for patient identification, especially in situations where traditional methods may be difficult, such as in intensive care units. At the same time, while facial recognition offers numerous benefits, it also raises significant ethical and privacy concerns. The potential for misuse, such as mass surveillance and unauthorized tracking, has led to debates about the regulation and ethical implications of this technology. As computer vision continues to evolve, it is crucial to find a balance between innovation and privacy rights.

Another sphere, which was shaped manually to image recognition techniques is augmented and mixed reality (AR), where it enables devices to perceive and interpret real world information, allowing people to seamlessly blend digital elements with the physical environment. AR systems utilize computer vision to overlay digital information onto the real world, creating immersive and interactive experiences. It happens by recognizing and tracking objects in the user's environment, thus allowing to accurately place virtual objects, such as 3D models or text, on top of real-world surfaces.

AR technology has a wide range of applications, including retailing, where it allows customers to visualize products in their own homes before purchasing, reducing return rates and enhancing the shopping experience, gaming, where it transforms gaming into immersive experiences, blending the virtual and real worlds; education, where AR can bring abstract concepts to life, making learning more engaging and effective and mixed reality, which uses AR to make one step further by creating a seamless blend of the physical and digital worlds. By combining elements of both AR and virtual reality (VR), MR enables users to interact with virtual objects in real-world environments.

As computer vision continues to advance, we can expect to see even more innovative and immersive creations in all of these fields, which can revolutionize this industries, from gaming and entertainment to healthcare and education, shaping the future of human-computer interaction.

At the same time, as already mentioned, computer vision technology performs several main tasks, namely:

–Object classification – it's a process of automatically recognizing and assigning categories or labels to graphic data objects. Its main function is to allow the computer system to identify some specific objects and patterns in images and give them labels or split by classes based on predetermined categories, defined during the learning process.

–Object detection – the process of finding and identifying a specific object or pattern on graphical data. The main goal of it is to accurately identify or recognize objects among others or remove the background of the image.

–Object identification – the process of finding the location of an object on graphic data in relation to other elements, which allows to detect its position and is heavily used in autonomous vehicles.

–Object segmentation – the process of revealing clear boundaries of the image or some specific parts of it and its shapes.

–Object verification – the process of detecting whether required object exists in the image for further processing.

–Object recognition – the process of recognizing objects in the image and their location. Differs from classification in that during the classification process the class of the entire image is determined, while recognition looks for some specific elements of it.

–Detection of object landmarks – the process of identifying and locating key points on the image, while determining their geometric and structural features. These features can be represented by corners, edges, some key points and parts of an object.

As was already mentioned, object and image recognition are one of the fields of computer vision, which recognizes some specific parts of an image. For this purpose, only a few basic methods can be highlighted, all other are mostly some of their modifications and optimized versions. On low level, all of them are using deep learning models.

Methods of traning computer vision models

Convolution neural networks. The first one is convolutional neural networks (CNNs). They were created based on the structure of the visual system and its models [6]. First models based on it, specifically local connections between neurons, which uses hierarchically organized transformations were described in the 80s [7] of the last century. At the same time, an error gradient, which allows to significantly improve quality of such models by adding some error coefficients was developed much later. This model usually consists of three main types of layers convolutional, pooling and fully connected (FC) layers.

General principle of how CNN works can be seen in Fig. 1 [8]. So, the basic functionality of CNN can be broken into four main steps. At first, after loading an image input, the first layer of this type of network will hold the pixel values of a photo. After that, this data is passed into the convolution layer, which determines the output of neurons by calculating some scalar products between weight and the region connected to the input value. After that, the rectified linear unit (commonly shortened to ReLu) is applied by using sigmoid function to the output, produced by the previous layer.

After that, data gets into pooling layer, which simply performs downsampling along spatial dimensionality of the given input by reducing the number of parameters within that activation. And, at the end everything is passed to fully connected layers (can be more than one), which will attempt to produce class scores from the activation being used for classification, thus estimating each image or parts of an image whether some of searched objects are presented there. Also, usually this operation contains a ReLu, that may be used between layers to improve performance and quality of the result.

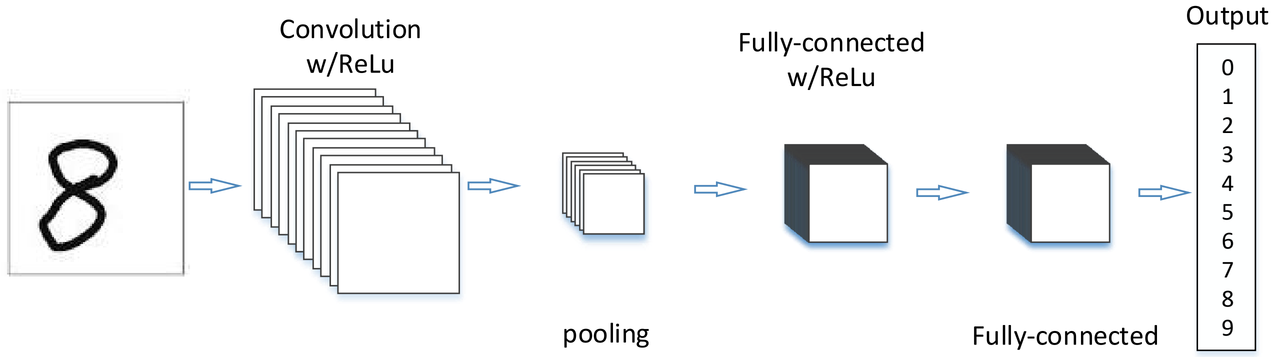

Fig. 1. Simple CNN architecture comprised of just five layers

More complex CNNs can contain more than one convolution and pooling layers, which are executed in synchronous way, one after another, creating a long process. The structure of such networks is shown in Fig. 2 and contains three pairs of convolutional/pooling layers.

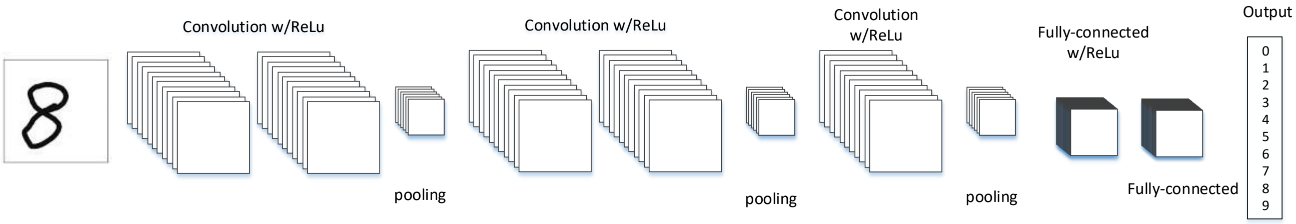

Fig. 2. CNN architecture with multiple convolutional and pooling layers

Going deeper into each of the described three layers, it's worth taking a closer look into convolutional layer and what differs it from usual well known neural network layers. So, let's assume that the input of neural network is 32x32 pixels color image [9], which has 3 RGB channels. Therefore, to connect the input layer to only one neuron in hidden layer 32x32x3 weighted connections will be needed. Adding just one neutron to the hidden layer will double this number of parameters and adding the next one will add around 6000 more parameters. It's clear that two, three and even ten hidden neurons won't be enough for any real classification application. Making the number of neurons equal to the number of pixels that the image has will require 32x32x3 by 32x32 connections, which in total creates 3145728 weights [10].

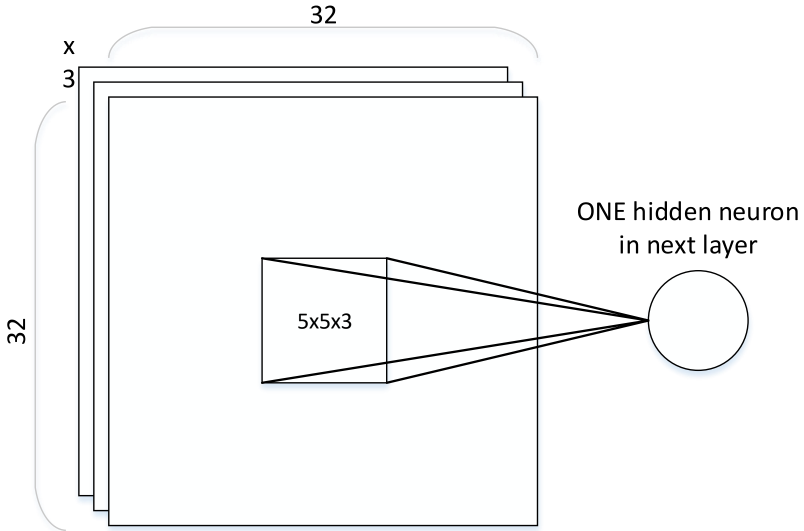

Fig. 3. Convolution as an alternative for fully connected network

So, instead of looking into the whole image, convolutional layers check it by splitting into local regions as shown on Fig. 3. In other words, hidden neurons from the next layer only get values from the corresponding part of the previous one. Thus, making such regions as 5x5 pixels will reduce the number of connections from over 3 million to 76800 weights [11].

Even if the number of weights dropped by 50 times, it can be still a lot of parameters to calculate and because of this, learning such network will require a lot of prepared data. So, more simplifications were created. It was proposed to decrease the number of weights by keeping local connection weights fixed for the entire neurons of the next layer. This will connect the neurons to the next layer with the same weight as the local region of the previous one. It will help to reduce the number of connections to just 5x5x3 or 75 in total [11].

In fact, even more specific optimization was found to decrease the number of parameters even more but still save quality of results and reduce some of the side effects.

The first one is stride method, which assumes that the next layer node has lots of overlaps with its neighbors by looking at the regions [9]. So, that way instead of moving filter just by one node, it will make 5x5 from 7x7 image. But, by increasing stride by 2, the output will be only 3x3. That way, apart from simplicity and removing of overlap, size of the output will be reduced too. One more optimization is padding, when zero-padding method is used. It increases image by 1 pixel into each side, improving capturing of corners and their better recognition, as they will be added not only to one filter part, but processed the same way as any pixels in the center of an image. These are two of the most used methods, even though, much more exist and actively used.

So, summing up, convolutional networks are one of the most popular methods used in image processing and computer vision, having lots of different modifications for different spheres of usage. They can be used but not limited to in:

–Mining sphere to find mixes in some ores [12].

–Face mask detection, which were heavily used during COVID [13].

–Smart refrigerator object detection [14].

–Image texture representation [15].

–Food quality estimation [16].

Deeper Dive into CNN Architectures. As CNN is the most advanced type of image recognition, let’s dive a bit into some of well known architectures and get an understanding of mathematical foundation of it’s main steps.

At first, let’s have a look into some of the most well known models, that were created based on CNN and now are defined as some sort of “standart” of computer vision.

First one is AlexNet model [yyy], which was introduced in 2012 and created some sort of revolution in the history of deep learning. This architecture, which consists of just eight layers, demonstrated the power of deep neural networks for image classification. Main features of this network, that were just firstly used for performing such tasks included:

- ReLU Activation function, as usage of ReLU as the activation function addressed the vanishing gradient problem, giving ability to train deeper networks.
- Dropout, this regularization technique randomly drops out neurons during training, preventing overfitting and improving generalization.
- Data Augmentation. For this feature such techniques like flipping, cropping, and color jittering were used to artificially increase the size of the training dataset.

Next huge step was a creation of VGGNet [yyy] built after the success of AlexNet by exploring the benefits of increasing network depth. In its core it used multiple 3x3 convolutional layers stacked together, achieving a larger receptive field while using fewer parameters than larger filter sizes. Such architecture, known for its simplicity and effectiveness, allowed to create deeper network architectures in the future.

After that, Residual Networks (ResNet) [yyy] introduced residual connections to address the vanishing gradient problem in the core of deep networks. These connections allow the gradient signal to flow directly through multiple layers, enabling the training of extremely deep networks. Adding a skip connection that bypasses one or more layers, ResNet effectively learns the residual mapping, making it easier to optimize deep networks.

One more network that is worth mentioning is InceptionNet [yyy], also known as GoogLeNet, which at first introduced as the inception module. It is a building block that combines multiple convolutional filters of different sizes to capture features at various scales. Such approach allowed the network to extract more information from the input image without creating significantly increasing the number of parameters. Additionally, InceptionNet employs 1x1 convolutional filters to reduce the dimensions of feature maps, making it computationally efficient.

The last model, this article mentions is EfficientNet [yyy], which was introduced as a scaling method that simultaneously scales the depth, width, and resolution of the network. By carefully balancing these dimensions, EfficientNet achieves state-of-the-art performance with fewer parameters and computational cost. Such scaling technique allowed its efficient deployment on various hardware platforms, from mobile devices to high-performance servers.

All of these architectures, along with their countless innovations, have significantly advanced the field of computer vision. They have created the way for more complex and powerful models that continue to push the boundaries of what is possible in artificial intelligence.

And now let’s have a look into mathematical foundation of Convolutional Neural Networks, having a closer look at such parts as convolution operations, pooling layers and operation of backpropagation.

Starting with convolution operation, which lies in the core of the convolution operation, and is a fundamental technique for feature extraction in image processing. This operation involves applying a filter, or kernel, to an input image to create a feature map that highlights specific patterns such as edges, textures, or gradients. It’s mathematical interpretation can be expressed as the next formula:

$$(I * K)(x,y) = \sum_{\mathrm{i}} \sum_{\mathrm{j}} I(x+i, y+j) * K(i,j)$$

Where:

- I is the input image;
- K is the kernel or filter;
- (x, y) are the coordinates of the output feature map.

The kernel, a matrix of weights that slides across the image, performing element-wise multiplication with the underlying pixel values. The sum of these products at each position constitutes a single value in the output feature map.

Kernel size significantly impacts the spatial extent of the patterns captured by the convolution operation. Common kernel sizes include 3x3, 5x5, and 7x7. Larger kernels can capture broader patterns but often require more computational resources. Smaller kernels, such as 3x3, are more efficient and can be stacked to create deeper networks that learn complex hierarchical features.

One more operation is the stride and padding parameters that influence the size and content of the output feature map. The stride determines the number of pixels the kernel moves across the image in each step. A larger stride reduces the size of the output feature map, leading to a loss of spatial resolution. Padding, on the other hand, adds extra pixels to the input image's boundaries, ensuring that the output feature map has the same spatial dimensions as the input.

Correct selection of kernel sizes, strides, and padding can effectively extract features from images, enabling them to excel in various tasks such as image classification, object detection, and image segmentation.

Next main part of CNN are pooling layer that are a fundamental component of CNNs. Their purpose is to reduce the spatial dimensions of feature maps, thereby decreasing computational cost and the number of parameters in the network. At the same time downsampling the feature maps helps to manage overfitting, which is a common problem in deep learning models.

The most common type of pooling is max pooling. This operation involves dividing the input feature map into non-overlapping rectangular regions and selecting the maximum value from each region. By selecting the most prominent feature values, max pooling effectively summarizes the information within each region, leading to a more compact representation. One of the key benefits of pooling is its ability to introduce invariance to small translations and distortions in the input image. By selecting the maximum value within a local region, pooling reduces sensitivity to slight shifts or rotations in the input. This property makes CNNs more robust to variations in image data, improving their generalization performance.

At the same time pooling layers contribute to the hierarchical feature learning process in CNNs. As the network progresses deeper, pooling layers progressively reduce the spatial dimensions of the feature maps. This creates a funnel-like structure, where low-level features such as edges and textures are captured in earlier layers, and higher-level, more abstract features are represented in later layers. By the final layers, the network has a compact, high-level understanding of the input image, enabling accurate classification or object/pattern detection.

While max pooling is the most commonly used pooling technique, average pooling can also be employed. Instead of selecting the maximum value, average pooling calculates the mean value within each region. This method can be useful for preserving average intensity information and reducing noise. However, it is less commonly used than the max pooling due to its tendency to blur the feature maps.

Last operation that is worth mentioning is backpropagation, which is the cornerstone algorithm that lies in the core of the training of deep neural networks. It is a specific method for adjusting the weights and biases in a network to minimize the difference between its predicted output and the true target values. It **consists of** two primary phases forward and backward pass.

Forward Pass is a first one, during which input data is fed into the network. After that it propagates through the layers, with each layer applying its weights and activation function to produce an output.The final output is compared to the true target value, and a loss function is calculated to quantify the error.

Backward Pass, as a counterpart, propagates error signal backward through the network, layer by layer, at each layer, the gradient of the loss function with respect to the weights and biases of that layer is calculated using the chain rule. These gradients indicate the direction and magnitude of the weight adjustments needed to reduce the error. At the same time weights and biases are updated using an optimization algorithm like stochastic gradient descent (SGD) [yyy] or Adam, which adjusts the parameters in the direction of the negative gradient.

Mathematically, the gradient of the loss function L with respect to a weight W can be expressed as:

$$\partial L/\partial W = \delta * a$$

Where:

- δ is the error signal at the current layer.
- a is the activation of the previous layer.

By calculating these gradients for each weight in the network, backpropagation allows the model to learn from its mistakes and improve its predictions over time.

Backpropagation has been instrumental in the success of deep learning, enabling the training of complex models with millions of parameters. By efficiently propagating error signals through multiple layers, it allows deep networks to learn hierarchical representations of data, leading to state-of-the-art performance in various tasks such as image and speech recognition, natural language processing, and more.

Following the foundational components of CNNs, it's essential to understand their training dynamics and practical applications in greater depth. Training a CNN involves iterative adjustments to the network's parameters, such as weights and biases to minimize the error between predicted outputs and true values. Through multiple epochs, or passes over the training data, CNNs progressively improve in accuracy. Techniques such as learning rate scheduling, batch normalization, and data augmentation help stabilize and accelerate training. Learning rate scheduling adjusts the rate at which parameters are updated, typically decreasing it over time to fine-tune the model. Batch normalization normalizes layer inputs to maintain stable distributions, accelerating training by reducing internal covariate shifts. Together, these strategies address various training challenges, ensuring efficient convergence.

Evolution of CNNs get it to a place, where there are widely applied in areas beyond image classification. For example, object detection networks to detect and locate objects within images, advancing applications like autonomous driving and security surveillance. In medical imaging, CNNs have enabled advancements in disease detection by automating complex diagnostic tasks, analyzing features in X-rays, MRIs, and CT scans with high accuracy. Additionally, semantic segmentation models such as U-Net excel in pixel-wise classification tasks, where each pixel in an image is labeled, assisting in fields like satellite imagery analysis and robotics. These applications demonstrate the versatility and transformative impact of CNNs across diverse fields.

Deep belief networks. The second method consists of deep belief networks (DBN) and deep Boltzmann machines. Both belong to the group of models called "Boltzmann family" in a way that they use Restricted Boltzmann Machine (RBM) as a learning module. The restricted Boltzmann machine is a generative stochastic neural network, the structure of which is shown in Fig. 4. In recent years, this model has shown its capabilities in image processing and feature extraction. In its structure it has deep neural network, which uses an idea of hierarchical learning based on neural connections between the human-brain system. The DBN model, in its low level, consists of many constrained RBMs working one after the other and stacking one after the other, meaning that the output of first RBM is the input of the second one and so on. This allows to model any specific dynamic system with any given accuracy, with the condition of having enough number of hidden neurons.

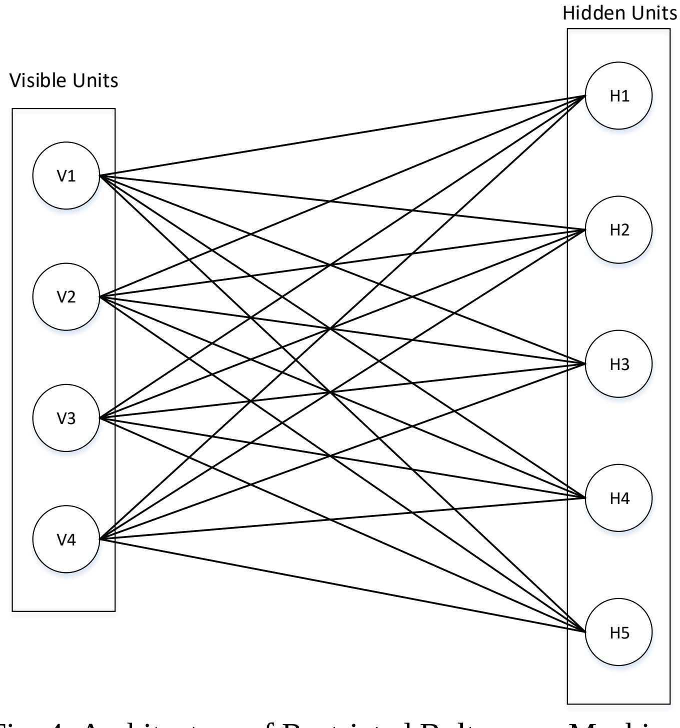

Fig. 4. Architecture of Restricted Boltzmann Machines

In its core deep belief networks entails training the network layer by layer, offering several advantages:

–Optimization of parameter selection for generating an appropriate initialization of the network, which helps avoid getting stuck in poor local optima up to a certain level.

–As it falls under unsupervised learning, it draws conclusions based on clustered data without requiring labeled data for training. However, it should be noted that this approach can be computationally expensive.

Autoencoders. Last, but not least, the method is autoencoders. They use an autocoder as the basic minimal unit or building block, the same as deep belief network uses Restricted Boltzmann Machine. Because of this, before the process of learning is started, it’s necessary to create an autoencoder and its structure. The general structure of such Autoencoders is shown in Fig. 5 [17]. This specific structure was used in [18] to recognize human facial expressions using Stacked Progressive Autoencoders.

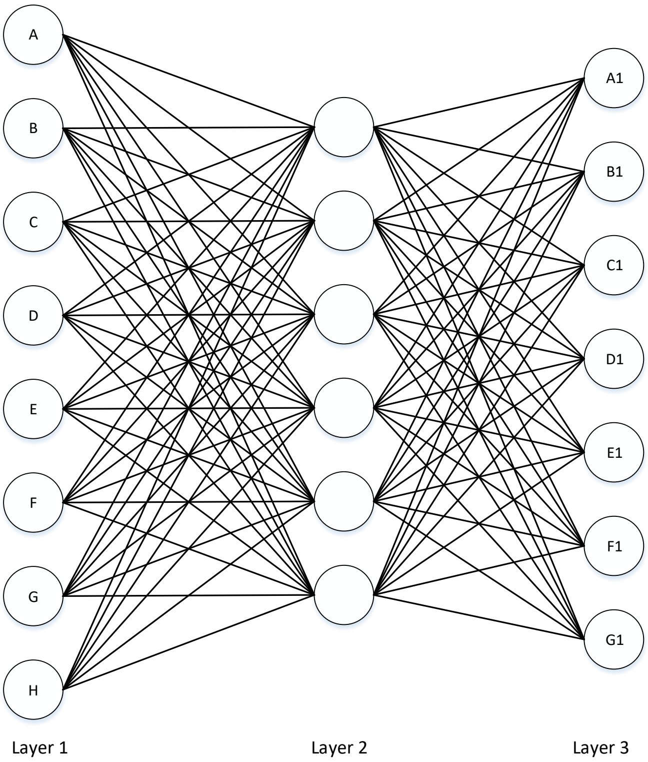

Fig. 5. Architecture of autoencoders

At its basics autoencoder is employed for training logical representations. Instead of predicting a specific target value for a given input, the autoencoder learns to recreate its own inputs through training adjustments. This process involves optimizing the reconstruction error using a back propagation method with the same dimensionality as the input. Diverse types of autoencoders include:

–Sparse autoencoder – is utilized to identify scattered features within input data. It facilitates easy categorization of high-dimensional representations into different classes based on their likelihood, similar to Support Vector Machines (SVM). It also enables interpretation of complex input data by recognizing various classes based on their likelihood. Sparse encoders have potential applications in biological vision.

–De-noising autoencoder (DAE) – enhances the robustness of models by operating effectively even in the presence of random noise.

–Contractive autoencoder (CAE) – represents an advanced version of DAE. It achieves durability by incorporating a contractive penalty to optimize reconstruction error, making it suitable for addressing unsupervised and transfer learning challenges.

Results

When analyzing differences of existing methods, it’s worth mentioning some specifics of each of them. They can be described as table 1 [19]. Under the conditions of testing in this article, CNNs performed much better in recognizing certain parts of objects, even under the conditions of certain transformation of them, than the other two models. At the same time, DBNs and autoencoders had more self-learning and did not need as much control as convolutional neural networks, although they showed slightly worse results on real data.

Table 1

Comparison of described models.

Model property	CNNs	DBNs	Autoencoders
Unsupervised learning	-	+	+
Training efficiency	-	-	+
Feature learning	+	-	-
Scale/rotation/translation invariance	+	-	-
Generalization	+	+	+

Also, considering that there are quite a lot of different variations and modifications of each of the models, such as adding additional elements [20], layers [18], etc., each of them has its own place of application. So, when solving some specific problem, it’s worth considering all of them.

Concerning usage of each of the described methods, even though somewhere they can be interchanged, each of them is better suited for some specific tasks. So, starting with CNN, which shows the best results in:

–Feature extraction – it automatically learns hierarchical features from input images through convolutional layers, capturing patterns and textures at various levels of abstraction.

–Spatial hierarchies – it utilizes pooling layers to down sample feature maps, capturing spatial hierarchies and reducing computational complexity.

–Classification – it often ends with fully connected layers followed by SoftMax activation, giving ability to classify objects within images based on the learned features.

Deep Belief Networks, having a bit unpredictability due to unsupervised learning are usually used for:

–Unsupervised pre-training, where they can be used for unsupervised feature learning to represent high-dimensional data such as images in a lower-dimensional latent space.

–Feature extraction, where DBNs can extract meaningful features from input data by capturing complex patterns and correlations present in images.

–Fine-tuning, where DBNs can be fine-tuned using supervised learning techniques to adapt the learned features for specific computer vision tasks such as classification or object recognition.

Getting to autoencoders, they are better suited to reconstruct input data, so their main tasks may include:

–Dimensionality reduction – they can be used to reduce the dimensionality of image data while preserving key features, enabling efficient storage and processing.

–De-noising – by reconstructing clean images from noisy inputs, de-noising autoencoders can effectively remove noise and enhance image quality.

–Feature learning – they can learn meaningful representations of input images, capturing key features and patterns that can be useful for downstream tasks such as classification or generation.

In brief, the history of computer vision spans nearly six decades, marked by a journey filled with numerous challenges and breakthroughs. Over the years, it has explored various technologies, adapted them to its needs, and even forged new ones to tackle a wide array of tasks. This ongoing saga has witnessed the evolution of computer vision from its inception to its current state, showcasing its resilience and adaptability in overcoming obstacles. From humble beginnings to complex applications, the story of computer vision is a testament to human ingenuity and the relentless pursuit of innovation.

Conclusion

In conclusion, this article describes computer vision and image recognition methodologies and their main usage in applications. From its inception in the 1960s to the present day, computer vision has traversed a remarkable evolutionary path, propelled by advancements in computing power and algorithmic sophistication.

At the core of computer vision lie several key methodologies, each with its unique strengths and applications. Convolutional neural networks (CNNs) have emerged as the cornerstone of image processing, excelling in tasks such as feature extraction and spatial hierarchies. Their ability to automatically learn hierarchical features from images has revolutionized fields like object recognition and scene understanding. Deep belief networks (DBNs) offer a pathway to unsupervised learning, facilitating the extraction of complex patterns and correlations from data. While initially characterized by their unpredictability, DBNs have found utility in tasks such as feature extraction and fine-tuning, particularly in scenarios where labeled data is scarce or expensive to obtain. Autoencoders, with their capacity for dimensionality reduction and noise mitigation, have emerged as invaluable tools in enhancing data fidelity and extracting meaningful representations from input data. Their applications span diverse domains, from denoising images to generating efficient feature representations for downstream tasks like classification and clustering.

The practical implications of computer vision are equally profound, permeating through various sectors of society. In the realm of autonomous vehicles, computer vision enables precise environment perception and navigation, enhancing safety and efficiency on roads. In healthcare, computer vision facilitates accurate medical diagnostics and disease detection, empowering clinicians with invaluable insights from medical imaging data. Moreover, computer vision technologies find applications in security and surveillance, augmented reality, industrial automation, and beyond, showcasing their versatility and ubiquity in addressing real-world challenges.

As a result, the integration of computer vision methodologies into practical applications has yielded tangible outcomes, from improved safety on roads to enhanced medical diagnostics. While challenges remain, including ethical considerations and algorithmic robustness, the trajectory of computer vision continues to point towards a future defined by innovation and societal impact.

In a matter of research, this article provided main spheres of computer vision and gives an understanding that feature extraction and classifications models should be used. The best of existing models are convolutional neural network, which are mainly used for this purpose and this way can be further researched and tested with close to real data for their productivity and possible modifications.

References

1. Roberts L. G. Machine perception of three-dimensional solids. New York: Garland Pub., 1980. 197 p.

2. Shirai Y. Three-Dimensional Computer Vision. Berlin, Heidelberg: Springer Berlin Heidelberg, 1987. 313 p.

3. Szeliski R. Computer Vision: Algorithms and Applications. Springer International Publishing AG. 2023.

4. Gonzalez R.C., Thomas M.G Pattern Recognition: an Introduction, Addison Wesley, Reading, 1978. MA.

5. Haralick R. M., Shapiro L. G. Glossary of computer vision terms. Pattern Recognition. Vol. 24, no. 1. 1991. P. 69–93.

6. Hubel D. H. Wiesel T. N., Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology. Vol. 160, no. 1. 1962. P. 106–154.

7. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics. Vol. 36, no. 4. 1980. P. 193–202.

8. O’Shea K., Nash R. An Introduction to Convolutional Neural Networks. ArXiv.org. 2015.

9. Understanding of Convolutional Neural Network (CNN): A Review / P. Purwono et al. International Journal of Robotics and Control Systems. Vol. 2, no. 4. 2023. P. 739–748.

10. Milosevic N. Convolutions and Convolutional Neural Networks. Introduction to Convolutional Neural Networks. Berkeley, CA. 2020.

11. C. Szegedy. Going deeper with convolutions, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.

12. K. Zhang. Computer vision detection of foreign objects in coal processing using attention CNN, Engineering Applications of Artificial Intelligence. Vol. 102. 2021.

13. S R. N., N M. Computer-Vision based Face Mask Detection using CNN. 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 8–10 July 2021.

14. Jain P. Automated Identification Algorithm Using CNN for Computer Vision in Smart Refrigerators, Computers, Materials & Continua. Vol. 71, no. 2. 2022. P. 3337–3353.

15. Liu L. From BoW to CNN: Two Decades of Texture Representation for Texture Classification / International Journal of Computer Vision. Vol. 127, no. 1. 2018. P. 74–109.

16. Al Haque A. S. M. F., Hakim M. A., Hafiz R. CNN Based Automatic Computer Vision System for Strain Detection and Quality Identification of Banana. 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), Rajshahi, Bangladesh, 8–9 July 2021.

17. Ksheera R Shetty Deep Learning for Computer Vision: A Brief Review, International Journal of Advanced Research in Science, Communication and Technology. 2022. P. 450–463.

18. M. Kan Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses, 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.

19. A. Voulodimos Deep Learning for Computer Vision: A Brief Review, Computational Intelligence and Neuroscience. Vol. 2018. P. 1–13.

20. H. Bouzidi Performances Modeling of Computer Vision-based CNN on Edge GPUs, ACM Transactions on Embedded Computing Systems. 2022.