Advancements in Computer Vision: A Comprehensive Review

Poorva Agrawal
Symbiosis Institute of Technology, Nagpur Campus
Symbiosis International (Deemed University)
Pune, India

Reyon Bose
Symbiosis Institute of Technology, Nagpur Campus
Symbiosis International (Deemed University)
Pune, India

Gopal Kumar Gupta
Symbiosis Institute of Technology, Nagpur Campus
Symbiosis International (Deemed University)
Pune, India

Gagandeep Kaur
Symbiosis Institute of Technology, Nagpur Campus
Symbiosis International (Deemed University)
Pune, India

Samiksha Paliwal
Symbiosis Institute of Technology, Nagpur Campus
Symbiosis International (Deemed University)
Pune, India

Aarya Raut
Symbiosis Institute of Technology, Nagpur Campus
Symbiosis International (Deemed University)
Pune, India

2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET) | 979-8-3503-1901-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICICET59348.2024.10616321

Abstract:
Computer vision is a versatile area that allows a computer to understand and analyze images from the environment. This paper focuses on a comprehensive discussion of where computer vision is today in light of its past, most recent accomplishments, and predictions of where it could go next. The evolution of computer vision over the past few decades has been marked by remarkable progress, driven by advancements in three key areas: utilize deep learning, hardware, and image-processing methods. Artificial neural networks are now commonly used in deep learning, which has improved the abilities of machines to recognize, classify, and understand visual content.
Computationally speaking, parallel with deep learning, other kinds of hardware, like GPU chips, can be used to perform such functions better. This has made it easy to train and deploy complex deep learning algorithms for real-time computer vision applications with expanded reach into many facets.
In addition, image processing has been dramatically improved, making it possible to process and enhance incoming visual data. Traditional and most recent techniques for Computational Photography have enabled the dependable and precise computer-vision system.

Keywords- Deep Learning, Pooling, Image Segmentation, CNN, Semantic Segmentation, Semantic Water shield, Semantic RSST, Instance Segmentation.

I. INTRODUCTION

One such subfield of artificial intelligence includes computer vision, which enables a machine trained in artificial intelligence to see, analyse, interpret, and comprehend extremely complicated images or images related to its surrounding environment. It has been put into practice in medicine and other fields like self-driven vehicles, safety systems, and advanced realities. The growth of computer vision, current difficulties, and prospects are topics this paper will address.

With computer vision, there is an age of artificial intelligence where the machine can understand and make decisions using visual information. This paper will trace the development of computer vision and its importance to the existing technological environment.

However, among various branches within the domain of artificial intelligence exists a very important sub-area called computer vision, which endows computer systems with the ability to understand complicated visual information about their surrounding environment. It implies that artificially intelligent systems can learn to understand and decode multifaceted visual data as sophisticated as the human vision system. The last years have seen dramatic progress within the computer vision environment, impacting everything from the medical setting to autonomous automobiles, security sensors, and virtual reality enhancements.

These advanced learning methods, such as transfer learning, will likely provide a refined and flexible artificial intelligence of the coming age. Notwithstanding this, however, there are still numerous challenges. Researchers and other practitioners still focus on data quality, its interpretability and flexibility in the field.

II. HISTORICAL DEVELOPMENT:

EARLY STAGES OF COMPUTER VISION

The evolution of computer vision can be traced back to the 60's as an interdisciplinary area that makes it possible for the machines to interpret visual data and consequently help in making some decisions. Here is a brief overview of key milestones and contributions from pioneers in the field:

979-8-3503-1901-9/24/$31.00 ©2024 IEEE

EARLY BEGINNINGS (1960S):

Computer vision has its origin in the sixties, where researchers looked for ways to make machine understand images.
Initial attempt involved creation of elementary algorithms for interpreting and understanding geometrical figures within pictures

III. TRADITIONAL COMPUTER VISION

The traditional approach used in computer vision is what we refer to as traditional computer vision. However, over the last couple of years, deep learning has contributed so much in computer vision development but it would not be anything if there was nothing that led up to this revolution. Here are some key aspects of traditional computer vision:

A. IMAGE PROCESSING:

Computer vision used to depend on classic image processing techniques like filters, edge detection, and image enhancement.
Pre-processing involved methods like convolution, Gaussian smoothing, and thresholding so that only significant information could be extracted from the pictures.

B. FEATURE EXTRACTION:

On the subject of computers, features are essential to identify pertinent information. Various image processes precede image feature extraction pre-processing techniques like normalisation, thresholding,
This performs the tasks like binarisation, resizing and so on on the constituent image [12].
Traditional computer vision relies on feature extraction as its pivotal step. This entails outlining specific designs or features of an image, which can provide more information. Object recognition and image matching primarily employed handcrafted features that included Harris corners, SIFT (Scale-Invariant Feature Transform), and HOG (Histogram of Oriented Gradients).

C. OBJECT RECOGNITION:

Traditional computer vision approaches to recognizing objects typically employed template matching algorithms where a portion of the whole input image was compared against other parts of that input image against a known template pattern to match that. Fig. 1 shows the intensity of the human face surface through object recognition.

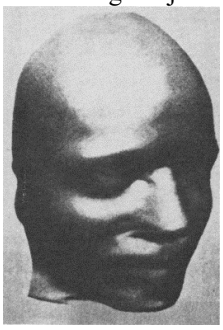

Fig. 1. Intensity image of human face surface [13].

IV. DEEP LEARNING IN COMPUTER VISION:

A. CONVOLUTIONAL NEURAL NETWORKS (CNNS):

Convolutional Neural:
The architecture of CNNs was modelled after the structure of the visual system and the different theories put forward. The first computation model is based on these local neighbourhood linkings among brain nerve cells and for the hierarchical organisation of signals in the nervous system [1].
This is what Neocognitron looks for in transformations of the image.
That is when neurons have identical parameters applied on different patches of the preceding layer and acquire a form of translational invariance at locations.
They later developed convolutional neural network architecture by LeCun and colleagues. CNN takes an image and uses a 3 × 3 or 5 × 5 filter as an input, as shown in Fig. 2.

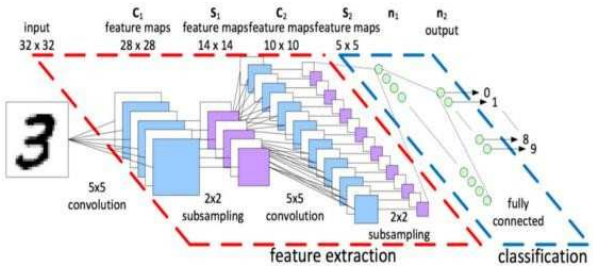

Fig. 2. Internal Convolutional Layer [9].

Three types of neural layers constitute a CNN:

(i) pooling
(ii) convolutional,
(iii) fully connected

Each has its job, but their work differs in each layer.
The input goes through every layer of a CNN.
Finally, a neuron's output volume is converted to an input volume of neuron activation.
Ultimately ending up in the last fully connected layer. Transforming the input data onto a one-dimensional attribute vector. CNNs have achieved much success in the field of computer vision.
For example, facial recognition, object detection, putting vision within power robotics, and self-driving cars.

(i) Pooling Layers:

These pooling layers are responsible for eliminating all redundancy.
The next convolutional layer is the input volume's width x height spatial dimensions. It does not affect the teacher pooling layer. Operations are in cross-section, and the depth dimension of the volume, subsampling, or so to speak, by this layer is also known as down sampling.

This is a result of diminution, leading to a decrease in information. However, this is a valuable loss for the network [1].
The reason is that the decrease in size leads to less computation overhead for subsequent network layers, and its performance is also good against overfitting; average pooling and max pooling are the most commonly used strategies. We provide max-pooling by giving input in a detailed theoretical performance of max and average pooling, resulting in quicker convergence, better invariant characteristics, and improved generalisation. Additionally, it boasts quite a large. The other variations of the pooling layer in the literature, such as stochastic pooling, spatial pyramid pooling, and def-pooling, are driven by different motivations and fulfil unique needs [2].

(ii) Convolutional Layers:
Using different kernel feature maps, a CNN in the convolutional layers convolutes the image and intermediate feature maps. This is because of the superiority of the convolution.
Several works have put forward it as an operational tool.
Replacing fully connected layers to achieve faster learning times [1].

(iii) Fully Connected Layers:
The neural network is made up of layers involving convolutional and pooling for high-level reasoning, where connected levels finally conduct the networks. This was previously indicated as connections neurons in a fully-connected layer are complete to all activations. Their activation, hence, this calculation can be done by multiplying a matrix.
By a bias set. Fully connected layers convert the 2D feature maps into a 1D feature vector.
Classification categories can even be termed the feature vector for further processing [2]. As in fully connected layers Fig 3. shows different CNN components in image processing.

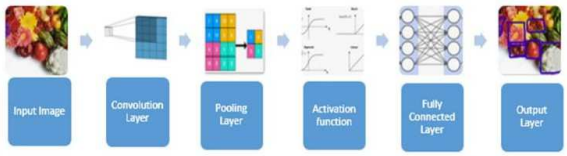

Fig. 3. CNN components [2].

B. Object Detection and Localization

FSOD is based on state-of-the-art CNNs that are currently employed for visual recognition as well as releases of large-scale datasets. Two fundamental constraints affect the fully-supervised object detection task:
1) large-scale instance annotation is time-consuming and expensive.
2) During instance-level labelling, they may unintentionally incorporate vague and uncertain manual annotations that can harm FSOD [3].

The community has begun working on weakly supervised object detection (WSOD), in order to overcome the aforementioned issues. Unlike the fully managed environment, WSOD is designed to find cases with single-level labels (e.g. categories of images as in their totality). On the other hand, who can access larger datasets available online, like Facebook/Twitter? Another related task to WSOD is WSOL, which can recognize just one object in a picture.

WSOD and WSOL differ in detection since they cover multiple and singular occurrences [3]. The object detection and localisation of image is done through (FSOD) and (WSOD) as shown in fig. 4.

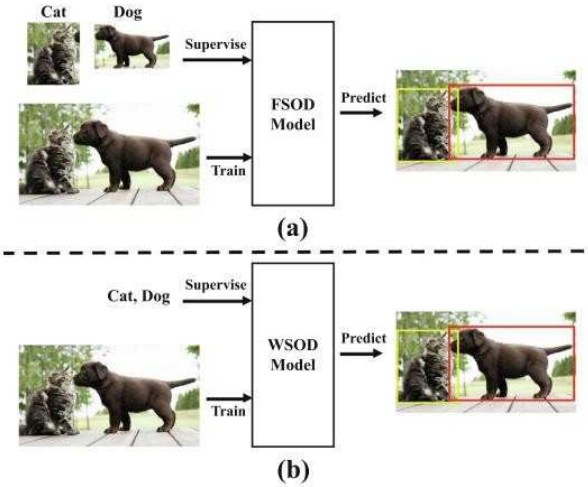

Fig. 4. (a) Fully-Supervised Object Detection (FSOD) uses the instance-level annotations as supervision. (b) Weakly-Supervised Object Detection (WSOD) uses the image-level annotations as supervision [3].

V. Image Segmentation

Image segregation is one of the very critical functions involved in computer vision and imaging. The system finds application in scene interpretation, robotic perception, analysis of medical images, video surveillance, augmented reality, and picture reduction. There is an abundance of image segmentation algorithms existing in the literature. This has made some DL-based image segmentation algorithms a reality, which is attributable to the significant success of deep learning for a backdrop [4].

A. Semantic Segmentation

This work aims at improving both image and perception. Conversely, separating marking materials and essential items would benefit local issues of image understanding. This is meant to underline that this method is not dependent on the age or gender of the individual concerned.
There are two traditional segmentation methods within the family of regions' growing algorithms for our selection of the segmentation algorithm. The first is the watershed.
The first is about segmentation and RSST [5] .

B. Semantic Watershed

This fact inspires the name of the watershed algorithm. How are local areas separated into catchments?

It can be said that the catchment basin comprises the local minima of an elevation function (usually the magnitude of the gradient image). After identifying these minima, the surrounding areas are where different flood regions are progressively inundated. The touch will delimit the regions. Unfortunately, this segmentation occurs due to the inappropriateness of the strategy employed. Marker-controlled segmentation is typically used. The flooding process within markers is suppressed just for features. Therefore, the final number of counties remains identical to the catchment basin [5].

C. Semantic RSST

This starts from the pixel level and continues with the gradual amalgamation of the equivalent regions and breaks into neighbouring regions until some predefined termination conditions are achieved. Traditional RSST will also re-calculate the weights of associated edges and then sort them so that the border with the minor consequence will be picked up each time. This process remains recurrent until a termination condition is satisfied. These could include the number of users who sign up or make purchases. The scales for measuring such criteria may differ, yet typically, they refer to the number of users who subscribe. It is based on region or a criterion on spatial distance [5]. The second one of the bottom-up segmentation approaches is the TRSST. It operates on pixels to perform similar processing on neighbouring homogeneous regions, iterating until a given termination criterion has been achieved. For example, they use a graph for internal image-based region representation –RSST [11].

D. Instance Segmentation

One of the relatively essential and complex, instance segmentation has become the latter. One issue confronting machine vision researcher is obtaining a class-label prediction of the target object; it locates different object class instances at a pixel level. Autonomous is primarily robots. It involves instance segmentation to assist in driving, surveillance, etc. Thanks to deep learning, especially to Convolutional Neural Networks. This led to a rapid rise in the instances of several instance segmentation frameworks such as (CNNs).
Mask R-CNN is a simple and fast instance segmentation algorithm. With an FCN, R-CNN provided high segmentation masks, box regression and object classification using the stage-wise network features extraction via FPN. Features were derived from a path across the top-down network involving lateral connections that are semantically cohesive.[6] .

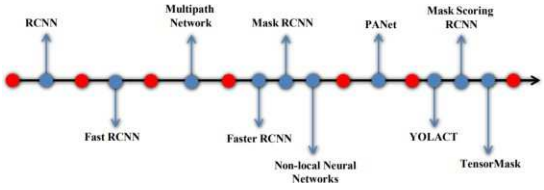

Fig. 5. Timeline for significant instance segmentation techniques [6].

VI. 3D Computer Vision

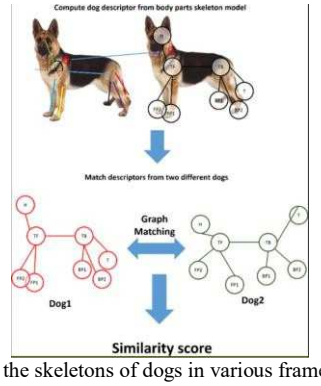

Fig. 6 Comparing the skeletons of dogs in various frames [10].

Recent developments of 3D sensing technologies like LIDAR or UAV sensors and the availability of inexpensive devices, such as Microsoft Kinekt allow for easy acquisition., which has been simplified for researchers. It has made processing data cheaper and easier to access than ever. Because of the type of scanning devices used in various raw data, they are obtained to capture the 3D scene or object of interest. Range images for a given direction come from different camera points in the case of UAV scanners. Then, these images are usually achieved through a registration method or structure-from-motion (SFM). As shown in fig. 6, a 3D computer vision uses different technologies to compare between different frames of dog skeleton.
To eliminate noise data, correlate them with each other, and at last, combine several of these into a single three-dimensional point cloud that can be used for further processing. LIDAR scanners convert the captured scene into a three-dimensional point cloud model. Additionally, acquiring some can improve digital images and the final product quality during charging.
Point cloud, conversely, has Kinect-like devices that capture RGB-D-type images.
An RGB colour and a depth image for one viewpoint of each camera can be included in the point cloud model or treated as two different input channels. Lastly, their concepts originate in the 3D hypercubes of a given section within cancer spectral band. [7].

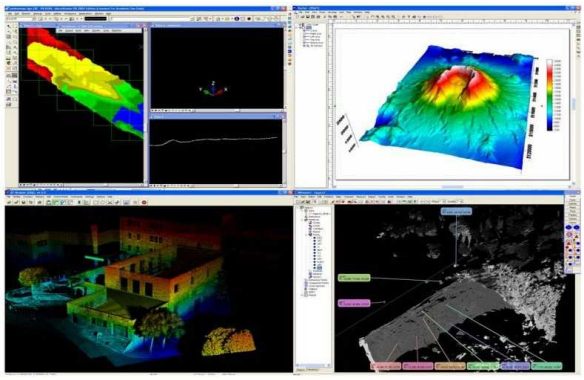

Fig. 7. LIDAR Cloud Processing Software [8].

VII. Future Directions

As AI continues to entrench itself into our everyday life, so does the evolution of computer vision technology continue. Due to advances of cloud computation, Auto ML pipelines, transformer, mobile centered DL libraries, and mobile computers. As this technology scale out, it will be appropriate for various vision applications [17].

A. Scarcity Of Data

A condition arising where there is not sufficient labeled training data or an insufficient amount of data for specific labels as compared with other labels within the dataset is called data scarcity.

This problem could be solved with learning by transfer, few-shot, and zero-shot approaches to datasets. Few-shot or Zero-shot learning succeeds in minimal or no labelled data and works perfectly. This involves a situation where one uses the knowledge from a model derived from another dataset to develop a new model, efficiently resolving this problem [17, 18].

B. Imbalanced Data

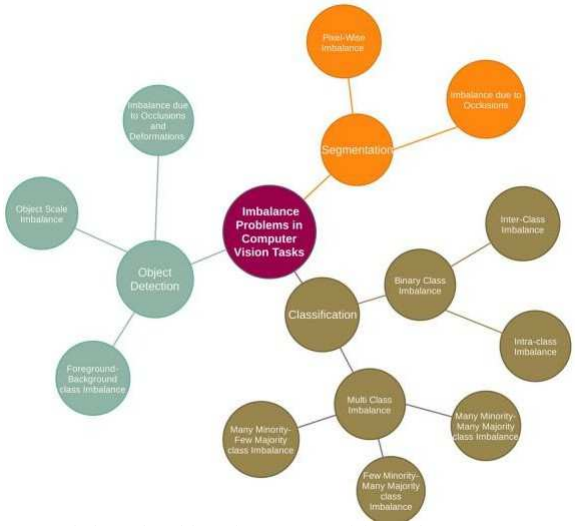

Fig. 8. Imbalanced problems in computer vision tasks [19].

Intraclass and interclass imbalanced could cause class imbalance among classes within the same data set. This phenomenon that occurs when there are larger samples from the majority class than from the minority class is termed as inter class imbalance [19, 20].

VIII. Conclusion

Deep learning and hardware have significantly improved, leading to massive developments in computer vision. Overall, this paper has tried to give a general review from the earliest starting point until now about different methods for the area of computer vision and possible prospects. The impact of such progress is far-reaching, and it brings the concept into a new understanding of society's multidiscipline field that promises to transform our life in the industry and others as our ability to see and understand a modern approach.

Deep learning is one of the most important developments that have changed the face of computer vision. Of note is convolutional neural networks (CNNs) that are capable of unmatched learning of complex patterns and representations directly from visual inputs. A significant advancement have helped to increase the precision and speed of several computer vision procedures such as image classification, object detection, generation or division. Advanced algorithms, in combination with computing capacities like GPUs or other accelerators, have enabled the development of various highly complex models for research and practice.

This paper traces the evolution of computer vision from the rule-based approach and primitive image processing techniques to modern-day data-driven strategies. The development reflects the step-by-step formulation of innovation whereby its results are rooted in the realities and failures of the preceding phase. Improvement on algorithms and algorithms over time has contributed to the increase in the ability of computer vision systems to handle complex situations in reality.

IX. References

[1] Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review1. Computational Intelligence and Neuroscience, 2018, 7068349.

[2] Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K., & Ghayvat, H. (2021)1. CNN variants for computer vision: History, architecture, application, challenges and future scope23. Electronics, 10(20), 247045. https://doi.org/10.3390/electronics102024706

[3] Shao, F., Chen, L., Shao, J., Ji, W., Xiao, S., Ye, L., Zhuang, Y., & Xiao, J. (2022). Deep Learning for Weakly-Supervised Object Detection and Localization: A Survey. Neurocomputing, 496, 192-207. https://doi.org/10.1016/j.neucom.2022.01.095

[4] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523-3542, 1 July 2022, doi: 10.1109/TPAMI.2021.3059968.

[5] T. Athanasiadis, P. Mylonas, Y. Avrithis and S. Kollias, "Semantic Image Segmentation and Object Labeling," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 3, pp. 298-312, March 2007, doi: 10.1109/TCSVT.2007.890636.

[6] Hafiz, Abdul Mueed, and Ghulam Mohiuddin Bhat. "A survey on instance segmentation: State of the art."1 International Journal of Advanced Computer Science and Applications 11.2 (2020): 475-488.

[7] loannidou, A., Chatzilari, E., Nikolopoulos, S., & Kompatsiaris, I. (2017). Deep learning advances in computer vision with 3D data: A survey12. ACM Computing Surveys (CSUR), 50(2), 1-38.

[8] Fernandez-Diaz, J.C., Singhania, A., Caceres, J., Slatton, K.C., Starek, M. and Kumar, R., 2007. An overview of lidar point cloud processing software. GEM Center Report No. Rep. 2007-12-001. University of Florida1.

[9] Dulari Bhatt, Chirag Patel, hardik Talsania, Jigar Patel, Rasmika Vaghela, Sharnil Pandya, Kirit Modi and Hemant Ghayvat. (2021). CNN Variants for Computer Vision: History, Achitecture, Application, Challenges and Future Scope.

[10] Barnard, S., Calderara, S., Pistocchi, S., Cucchiara, R., Podaliri-Vulpiani, M., Messori, S., & Ferri, N. (2016). Quick, Accurate, Smart: 3D Computer Vision Technology Helps Assessing Confined Animals' Behaviour. PLOS ONE, 11(7), e0158748.

https://doi.org/10.1371/journal.pone.0158748

[11] hanos Athanasiadis, Phivos Mylonas and Yannis Avrithis. (2007). A Context-based Region Labeling Approach for Semantic Image Segmentation.

[12] Salau, A. O. and Jain, S. (2019). Feature Extraction: A Survey of the Types, Techniques, Applications1. In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCA),

pp. 158-164

[13] Besl, P. J., and Jain, R. C. 1985. Three-dimensional object recognition. ACM Comput. Surv. 17, 1 (Mar. 1985), 75-145. DOI:https://doi.org/10.1145/289.291

[14] Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. Artif. Intell. Rev. 2020, 53, 5455–5516. [CrossRef]

[15] LeCun, Y. Convolutional networks and applications. ISCAS IEEE 2010, 253–256. [CrossRef]

[16] X. He, Y. Yang, B. Shi, X. Bai, Vd-san: Visual- densely semantic attention network for image caption generation, Neurocomputing.

[17] Supriya v. Mahadevkar, bharti khemani, shruti patil, ketan kotecha, deepali r. Vora, ajith abraham, (senior member, ieee), and lubna abdelkareim gabralla. (2022). A Review on Machine Learning Styles in Computer Vision—Techniques and Future Directions.

[18] J. Tavares and N. Jorge, Lecture Notes in Computational Vision and Biomechanics. Springer, 2012.

[19] V. Sampath, I. Maurtua, J. J. Aguilar Martín and A. Gutierrez, "A survey on generative adversarial networks for imbalance problems in computer vision tasks," Journal of Big Data, vol. 8, no. 27, 2021.

[20] Ali A, Shamsuddin SM, Ralescu AL. Classifcation with class imbalance problem: a review. Int J Adv Soft Comput Applicat. 2015;7:176–204.