YOUNG RESEARCHER
ISSN: 2277-7911Impact Factor: 5.077 (SJIF 2021)
Peer-Reviewed, Refereed, Indexed Journal
Volume - 12, Issue - II, Apr-May-June 2023

A REVIEW ON OBJECT RECOGNITION USING DEEP LEARNING

Mr. Prashant Bhat[1] & Dr. Anil Kumar[2]

Ph.D. Research Scholar, Department of Mechanical Engineering
Shri J.J.T. University, Rajasthan, India
Professor & Research Guide, Department of Mechanical Engineering
Shri J.J.T. University, Rajasthan, India
***Corresponding Author:** Mr. Prashant Bhat*

DOI - 10.5281/zenodo.10029205

ABSTRACT:

Researchers in the field of object recognition have turned a lot of their attention to deep learning techniques because of the inherent power these approaches have in overcoming the flaws of traditional approaches that are reliant on hand-crafted features. Over the course of the last several years, there have been tremendous advancements achieved in object identification made possible by deep learning methods. In this study, various contemporary and effective frameworks for deep learning are presented for the purpose of object recognition. The most current paper on newly developed algorithms for object identification using deep neural networks is offered here in its entirety. It is also covered the many different benchmark datasets that are used in the process of performance assessment. In addition, the applications of the object recognition technique for certain categories of things (such as faces, buildings, plants, and so on) are emphasised here. In conclusion, we will discuss the positives and negatives of the already available methodologies, as well as the potential for future expansion in this field.

Keywords: *Convolutional Neural Network (CNN), Faster R-CNN, Network on Convolution Feature Map (NoC), Deep Expectation (DEX), Deep Residual Conv-Deconv Network, A-ConvNet.*

INTRODUCTION:

Because of the fast advancement in computer technology, the computer has the potential to play an essential role in the completion of normal daily chores of everyday life [1]. Visual object classification and identification are common and automatic processes carried out by the biological visual system of a human being; nevertheless, these tasks are difficult for a computer to do. to mimic because there is a significant degree of mutability in the pictures of objects belonging to the same class when seen under various settings. The field of computer vision faces a significant challenge in the form of object recognition. In light of the fact that the work of implementing object identification on computers is a difficult

one, it is necessary to develop object recognition that is both powerful and less complicated [2]. The digital database containing visual information is always expanding, and in order to effectively manage and analyse the vast quantities of visual information that are now available, methods of image analysis are necessary that can automatically obtain the semantic contexts of the data. One of the most important aspects of the object identification challenge is the context provided by the photos themselves, namely the things that are visible in them. The ability to accurately describe the features of a picture is essential to developing an effective object identification system [3].

During the last decade, a thorough research of high resolution picture classification has been conducted using handmade features from both the spatial and spectral domains. For the purpose of providing the spectral variation information necessary for effective picture classification, the gray-level co-occurrence matrix, abbreviated as GLCM [4], is used as a texture-based descriptor. The enlarged morphological profiles that were suggested by Benediktsson and colleagues [5]. to collect geographical characteristics for the purpose of high-resolution urban picture classification. In addition, the Gabor filter [6] and the wavelet transform [7] were used in order to extract spatial features from pictures with a high resolution. The intra-class variation of the building database, which requires the feature to be handmade, is not an effective option. Therefore, features that were handmade are substituted with features that were retrieved using a sparse coding approach that was given by Chenyadat [8]. Another feature learning model that was proposed by Tuia et al. [9] is referred to as the sparse- constrained support vector machine (SVM). When it comes to scene classification, picture classification, and face recognition, deep features [10] are far more powerful and efficient than low level features.

Techniques that are based on super-pixel grouping (such as MCG [11], CPMC [12, and Selective Search [13]) and approaches that are based on sliding window are the most common types of object proposal approaches (e.g., edge boxes [14], objectness in windows [15]). In addition to this, there are various ways for object proposal that are used as detector-independent external components (e.g., selective

search object detection). The R-CNN [16] approach is used in the role of an object detector in order to divide the proposal area into several item categories or backgrounds. The work that Viola and Jones did that started everything off makes use of the Haar [17] characteristics and the enhanced classifies on sliding window. In order to construct deformable graphical models, the HOG features [18] are merged with linear SVMs [19], a sliding window classifier, and DPM. For effective detection and classification, the overfeat technique combines each sliding window of a convolutional feature map with a fully linked layer. This creates a system that can identify and categorise things more quickly. In the SPP-based detection approach the features are merged from the suggested area on the convolution features map, and then they are initialised to the fully connected layer so that they may be classified.

A decision tree a random forest and a support vector machine (SVM) are some examples of the standard supervised classification models. During the training phase of a random forest technique, numerous decision trees are built, and then the combined predictions from all of the trees are utilised to make a classification. To deal with high-dimensional data, SVM employs the usage of finite training samples. The random forest and the support vector machine (SVM) are examples of shallow models; in comparison to deep networks, their capacity to deal with nonlinear input is restricted. Chen et al. suggested a stacked auto encoder for the purpose of image classification. This auto encoder would predict the hierarchal characteristic of a hyperspectral picture in the spectral domain. The spectral-based properties necessary for the classification of hyperspectral data are represented by a deep belief network (DBN). The authors of this study, Mou et al. presented a recurrent neural network for the classification of hyper-spectral pictures. All of the aforementioned techniques, such as auto encoders, RBN, and DBM, fall within the category of 1-D deep architectures. Processing in a one-dimensional design might result in the loss of structural information about hyperspectral data. The CNN has the capability to automatically uncover the contextual 2-D spatial information for the purpose of picture classification. For the purpose of spectral-spatial classification of hyperspectral remote sensing pictures, researchers have developed a variety of CNN-based

models that are supervised. A supervised, L2-regularized, three-dimensional CNN-based feature extraction model that may be utilised for classification purposes was developed by Chen et al. [29]. Ghamisi et al. presented a CNN model that automatically improves itself. Spectral and spatial characteristics were used to build a classification framework that was presented by Zhao and Du et al. Romero et al. propose an unsupervised convolutional network for spatial-spectral feature extraction using sparse learning to predict the network weights. This is part of the transition from supervised CNN to unsupervised CNN.

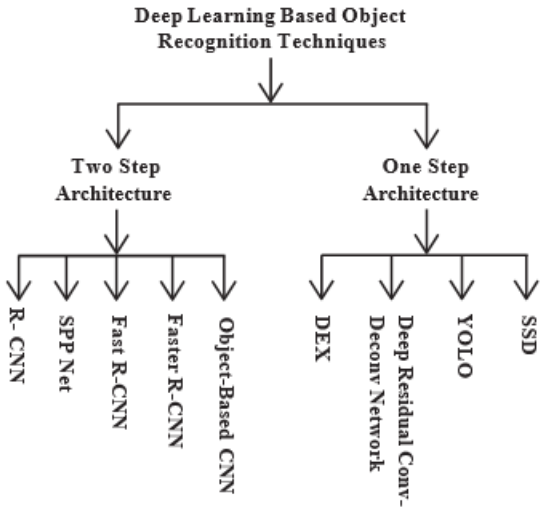

Fig. 1. Overview of deep learning based object recognition techniques

TWO-STEP ARCHITECTURE:

R-CNN was first suggested by Ross Girshick [16] in 2014 with the purpose of improving the candidate bounding box's quality and extracting high-level information via the use of deep architecture. Using the PASCAL VOC. 2012 dataset, R-CNN showed a much improved performance compared to the previously best result. The R-CNN two phases are as follows: During the step known as "Generation of Region Proposal" in R-CNN, around 2,000 region proposals are generated by the use of the selective search approach. By applying selective search, the precise bounding boxes of the arbitrary sizes may be constructed extremely quickly and with a reduction in the amount of searching space required. This is because selective search makes use of a bottom-up grouping and saliency signals to determine the results it returns. Deep CNN was used to extract the deep features from the clipped or distorted area proposed for the deep feature extraction method. The tremendous learning capacity and robust expressive capability of the CNN's structures allowed for the successful extraction of these final and robust 4096-dimensional features. With the assistance of category specific linear SVMs, the area recommendations are recorded as either positive or negative (background) regions. These areas are then subjected to a greedy non-

maximum suppression (NMS) filter after being adjusted by the bounding box regression, which results in the production of final bounding boxes for secured object locations. Despite the fact that R-CNN provides a number of benefits over more conventional methods, there are still certain disadvantages associated with using it. The need for an input picture of a fixed size causes the re-computation of the CNN to take much more time during the testing phase. The training that you get from R-CNN is completed in stages. Because the characteristics of many region proposals are retrieved and saved on the disc during the training process of R-CNN, more time and storage space are needed. because of the pyramid's three distinct levels. Due to the fact that calculation costs were shared prior to the SPP layer, SPPnet achieves superior results in terms of proper region proposal estimates. Additionally, it improves the efficiency with which detection is carried out during the testing phase.

SPPnet has shown great increases in both accuracy and efficiency in comparison to R-CNN, despite the fact that it does have a few disadvantages. The higher storage capacity is a cost for SPPnet as a result of its multi-stage pipeline design, which is analogous to that of R-CNN. It is not possible for the fine tuning method to update the convolutional layer that comes before the SPP layer. As a result of this, it should not come as a surprise that the accuracy of the deep network decreases. In order to solve these issues, Girshick developed a whole new architecture for CNN that he called Fast R-CNN. In the same way as SPPnet does it, the whole picture is processed by the conv layer in the Fast R-CN in order to produce the feature map. The feature vectors of a predetermined length are extracted from each proposed area by the ROI pooling layer. In order to get to the two output layers, each feature vector goes through a number of feature convolution layers (fc layers). The probability of C 1 categories is produced by the first layer, and the location of the bounding box is produced by the second layer using four real-value integers. Pipelining in Fast R-CNN is sped up by sampling the mini-batches hierarchically and by employing truncated singular value decomposition in the layer fc layer. This combination of techniques allows for faster processing (SVD).

The region proposal methods [13] are used in Faster R-CNN to

anticipate the placement of the item in a number of different object detection networks. Fast R-CNN [65] and SPPnet are used as detection networks with a decreased amount of operating time; nonetheless, the issue lies with the calculation of area recommendations. The proposed RPN makes a contribution to the detection network in the form of convolutional features extracted from the complete picture. This fully convolutional network also known as RPN, makes predictions about the object borders and the objectness score. The detection network that is used here is the Fast R-CNN. After that, the Fast R-CNN and the RPN are combined into a single, unified network by making use of the convolutional feature in conjunction with the attention mechanism. In the end, RPN informs the detection network where to observe, and the detection network finds items in that specific location based on what it sees there. The deep VGG-16 is the network that is employed for detection. The generation of region proposals is accomplished by gradually moving a tiny network across the feature map of the final convolutional layer. The tiny network is responsible for the production of the n n spatial window of the input convolutional feature map. Each sliding window is mapped by the feature with the lowest dimension (256-d for ZF and 512-d for VGG). This feature supplies the input for the two sliding layers, which consist of a box regression layer (reg) and a box classification layer (cls). Both networks (Region Proposal and Object Detection) have shown the presence of a shared convolution layer. Work is done using this strategy on the ZF net, which has a total of 5 layers, and the VGG-16 network, which has a total of 13 sharable convolutional layers. For high resolution photographs, accuracy in classification and interpretation as well as speed, play a very important role. This is especially true in the fields of disaster relief and urban planning. When the resolution of the photographs is increased, it is more difficult to recognise intricate patterns because of the increased detail. Deep learning offers a powerful and effective method for minimising the semantic gap that is present in object-based CNNs by simplifying the complicated patterning. Deep learning algorithms are not effective in capturing the boundaries between the various objects because of this. It has been recommended that the deep feature learning approach and the object-based classification strategy should be combined into one in order to

solve this issue. The accuracy of the classification of the high-resolution picture is improved by the approach that was presented. The process is broken down into two stages: first, the extraction of deep features by CNN, and subsequently, using deep features to aid with object-based classification.

APPLICATIONS:

Plant Identification:

The plant identification system is a subfield of computer vision that provides assistance to botanists in the process of fast and readily identifying previously undiscovered plant species. A number of investigations have been carried out in order to broaden the use of leaf data to the classification of plant species. The convolutional neural network is used in this approach to extract the relevant features of the leaves, and the deconvolutional neural network is used to determine how the yield perception of the retrieved characteristics should be calculated. Different orders of venation may be obtained using this approach, which is an improvement over the information about shape. According to the species class, the multilevel representation of leaf data (going from a lower level to a higher level) has been observed. The performance of the plant classification system is improved as a result of this study, which is beneficial in the creation of a hybrid feature extraction model.

Age Estimation:

Age is one of the most significant aspects of both a person's identity and their ability to engage socially. The ability to estimate a person's age may be determined by considering a number of criteria, including posture, facial wrinkles, vocabulary, and information. Age estimate is utilised in the development of a wide variety of applications, such as intelligent human-machine interfaces, as well as safety and protection measures in a variety of fields, including transportation, medical, and security. The advancements made in the field of artificial intelligence (AI) make it possible to more accurately determine a person's age via the use of the deep learning approach. When compared to more conventional methods of age estimate, the deep learning algorithms demonstrate both the efficacy and the resilience of the age estimation process.

Target Classification for SAR Images:

The SAR-ATR method, which stands for synthetic aperture radar automated target recognition, is made up of two components: a feature

extractor and a trainable classifier. The hand-designed characteristics are often removed, which has an effect on the system's degree of precision. By automatically learning features from the vast amounts of data provided, the deep convolutional networks were able to obtain the most advanced results possible in a number of computer vision and voice recognition tasks. Because of the major overfitting problem that arises when using convolutional networks for SAR-ATR. In order to solve this problem, an all-convolutional network, or A-ConvNet, has been suggested as a solution. The A-ConvNet does not use solely completely linked layers; rather, it uses layers that have sparse connections between them. When it came to the classification of targets in the SAR image dataset, the A-ConvNet displayed performance that was superior to that of standard ConvNet.

Face Recognition:

The process of recognising a person based on their face, either from a picture or a database [85] is referred to as facial recognition. In order to solve the challenge of face recognition brought on by the ever-increasing size of the dataset, machine learning methods such as deep neural networks are being used. When applied to huge datasets, the deep learning algorithms perform very well. Particularly convolutional neural networks (CNN) are able to achieve an exceptionally high recognition rate for the face recognition challenge.

COMPARATIVE ANALYSIS:

The table below provides a summary of some of the research that has been done in the field of object recognition using deep learning. The table provides an overview of the several methodology approaches that were used by various studies. The findings that were released by those researchers are really promising, although they were computed for a certain kind of database. The crucial issue is how well these strategies will function when applied to different types of databases. Therefore, it would be beneficial to do a comparison study comparing the various approaches indicated in the literature.

CONCLUSION:

Deep neural network-based object identification methods offer outstanding performance in modern approaches to object recognition because of their tremendous learning

capacity. These techniques are used to recognise objects. The current advances of a deep neural network-based object recognition framework have been discussed in this article. The topic of discussion is object recognition. After that, one-step frameworks like YOLO and SSD, among others, are analysed and discussed. In addition, a discussion is had on the several benchmark datasets as well as the numerous application fields of object identification. In conclusion, we draw a hopeful outlook for the future scope of object recognition in order to have an extensive view on the topic. This work offers valuable insight and direction for future developments in the field of deep learning-based object identification. A survey of the relevant literature reveals that there is need for further development. There is no contextual information available on a global level when using the object-based CNN for high-resolution picture classification approach. In the future, the primary emphasis will be placed on the contextual information in order to further increase performance. This is due to the fact that information about the connection between picture objects influences the effectiveness of classification. Estimation models may be designed in Segnet to calculate the degree of uncertainty associated with predictions made using deep segmentation networks. For the age estimate method DEX, the size of the training dataset may be expanded in the near or far future. Better face alignment may be achieved with the use of more dependable landmark detectors. Exploring the capabilities of Deep Residual Conv-Deconv Network for Hyperspectral Image Classification technique employing APs and estimate profiles that extract spatial information in a manner that is both resilient and adaptable is one of the potential future works that might be done.

REFERENCES:

[1]. Shokoufandeh, A., Keselman, Y., Demirci, M. F., Macrini, D. and Dickinson, S.J., 2012. Many to many feature matching in object recognition: A Review of three approaches. IET Computer Vision, 6(6), pp.500–513.

[2]. Lillywhite, K. and Archibald, J., 2013. A feature construc- tion method for general object recognition. Pattern Recognition, New York, NY, USA, Elsevier Science Inc. Vol. 46, pp.3300–3314.

[3]. Martin, L., Tuysuzojlu, A., Karl, W. C. and Ishwa, P., 2015. Learning based object identification and segmentation using dual energy CT images for security. IEEE Transaction on Image Processing, 24(11), pp.4069–4081.

[4]. Puissant, A., Hirsch, J. and Weber, C., 2005. The utility of texture analysis to improve perpixel classification for high to very high spa- tial resolution imagery. International Journal Remote Sensing, 26(4), pp.733–745.

[5]. Benediktsson, J.A., Palmason, J.A. and Sveinsson, J.R., 2005. Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Transactions on Geoscience and Remote Sensing, 43(3), pp.480–491.

[6]. Bau, M.T.C., Sarkar, S. and Healey, G., 2010. Hyperspectral region classification using a three-dimensional gabor filterbank. IEEE Transactions on Geoscience and Remote Sensing, 48(9), pp.3457–346.

[7]. Huang, X., Zhang, L. and Li, P., 2008. A multiscale feature fusion approach for classification of very high resolution satellite imagery based on wavelet transform. International Journal Remote Sensing, 29(20), pp.5923–5941.

[8]. Cheriyadat, A.M., 2014. Unsupervised feature learning for aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 52(1), pp.439–451.

[9]. Volpi, D.M., Mura, M.D., Rakotomamonjy, A. and Flamary, R., 2014. Automatic feature learning for spatio-spectral image classification with sparse svm. IEEE Transactions on Geoscience and Remote Sensing, 52(10), pp.6062–6074.

[10]. Hinton, G.E., Osindero, S. and Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Computation, 18(7), pp.1527– 1554.

[11]. Arbelaez, P., Pont-Tuset, J., Barron, J.T., Marques, F. and Malik, J., 2014. Multiscale combinatorial grouping. Computer Vision and Pat- tern Recognition (CVPR), pp.328–335.

[12]. Carreira, J. and Sminchisescu, C., 2012. CPMC: Automatic object segmentation using constrained

parametric min-cuts. IEEE Transac- tions on Pattern Analysis and Machine Intelligence, 34(7), pp.1312–1328.

[13]. Uijlings, J.R., van de Sande, K.E., Gevers, T. and Smeulders, A.W., 2013. Selective search for object recognition. International Journal of Computer Vision, 104(2), pp.154–171.

[14]. Zitnick, C.L. and Dollar, P., 2014. Edge Boxes: Locating Object Proposals from Edges. European Conference on Computer Vision, pp.391–405.

[15]. Alexe, B., Deselaers, T. and Ferrari, V., 2012. Measuring the object- ness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11), pp.2189–2202.

[16]. Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Seg- mentation. IEEE Conference on Computer Vision and Pattern Recog- nition, pp.580–587.

[17]. Viola, P. and Jones, M., 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features. Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.I-511–I-518.

[18]. Dalal, N. and Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. Proceeding of IEEE Computer Society Con- ference on Computer Vision and Pattern Recognition, pp.886–893.

[19]. Uijlings, J.R., van de Sande, K.E.T., Gevers and Smeulders, A.W., 2013. Selective search for object recognition. International Journal of Computer Vision, 104, pp.154–171.