An AI & ML Based Detection & Identification in Remote Imagery: State-of-the-Art

Submitted: 8th October 2021; accepted: 8th February 2022

Hina Hashmi, Rakesh Dwivedi, Anil Kumar

DOI: 10.14313/JAMRIS/4-2021/22

Abstract: Remotely sensed images and their allied areas of application have been the charm for a long time among researchers. Remote imagery has a vast area in which it is serving and achieving milestones. From the past, after the advent of AI, ML, and DL-based computing, remote imagery is related techniques for processing and analyzing are continuously growing and offering countless services like traffic surveillance, earth observation, land surveying, and other agricultural areas. As Artificial intelligence has become the charm of researchers, machine learning and deep learning have been proven as the most commonly used and highly effective techniques for object detection. AI & ML-based object segmentation & detection makes this area hot and fond to the researchers again with the opportunities of enhanced accuracy in the same. Several researchers have been proposed their works in the form of research papers to highlight the effectiveness of using remotely sensed imagery for commercial purposes. In this article, we have discussed the concept of remote imagery with some preprocessing techniques to extract hidden and fruitful information from them. Deep learning techniques applied by various researchers along with object detection, object recognition are also discussed here. This literature survey is also included a chronological review of work done related to detection and recognition using deep learning techniques.

Keywords: Convolutional Neural Network, Remote Sensed Imagery, Object Detection, Artificial Intelligence, Feature Extraction, Deep Learning, Machine Learning

1. Introduction

In this continuously growing world in the direction and advancement of technology, remote imagery is serving the world in multiple facets. Data nowadays is collected and stored in digital formats that make interpretation and analysis possible on that. Remote sensing images, images are collected through various satellites, Aerial photography, Lidar, Landsat, Spy Satellites, and sentinel images. Remote imagery is nowadays recorded in digital form so that analysis can be done on them for extracting some hidden information from them. But the limitation with the remote imagery is that the images collected from the satellites are not always up to the mark. It usually contains some kind of noise, unclear visuals, and mixed color channels. So, it becomes necessary to process that data before applying any processing on them [31]. Digital image processing involves many procedures and functions to format and correct the data for segmentation and classification. With the help of those processes and techniques, the data can be used refined and it can be used for various commercial purposes like Earth observation, weather forecasting, Forestry, Agricultural usage, Surface changes, and the analysis of Bio-diversity. Remote applications are also useful for interpreting the road conditions in rural areas, identifying crop conditions, etc. For a long time, remote sensing images are being processed and analyzed with the help of deep learning techniques that is part of neural networks. Although, before the development of deep learning techniques, remote sensing imagery was being observed with the help of a support vector machine (SVM) and various ensemble classifiers like Random forest for change detection or image classification. SVM got high demand and attention due to its capability of dealing with high and multi-dimensional data with a limited amount of training data [2].

In the recent era, the emerging trends in DL have renewed the interest of the remote community for neural networks. From the near 2014, the whole remote sensing community has diverted its attention toward DL as DL techniques and algorithms have recorded its achievements in many image analysis tasks like Land cover Land use observations, object detection, scene classification, etc [30] [14] [36] [52] [35] [53] [54]. Going through the vast literature regarding DL is showing that DL is having general approaches related to the development of basic deep learning algorithms [55] and the detailed reviews for various emerging and advanced fields like medical image recognition, speech recognition techniques [56]. Some studies are showing evidence of DL in remote sensing applications [4]. The literature survey by [57] has focused on the applications of DL in remote sensing images classification for major observations. [15] has performed a complete broad review specifically focusing on related uncommon sub-areas of remote data application areas like 3-d modeling. DL algorithms and techniques are having diverse sub-domain within the remote sensing field and the application areas are continuously increasing in getting the more quantitative and systematic analysis of the data [32]. The motivation of this study is to draft a comprehensive review of DL algorithms in various areas and sub-areas of remote sensing applications that include object detection, image segmentation, scene...
classification, image registration, image fusion, etc in images as well as videos [33]. By going through a vast review in the area of object detection in remote images through deep learning techniques, we have summarized the findings in various research articles. At last, a critical summary is included followed by major gaps and challenges found.

1.1 Remote Imagery and Preprocessing

Remotely sensed images are not simple images, they contain multiple formats and resolution challenges. They can be single channel or multi-band images having variations in their resolutions too. The spatial resolution of remote imagery is the most important aspect that is directly related to the accuracy of objects. To generate land cover maps for various reasons like environment planning, change detection, transport, and traffic planning temporal resolution is used. Medium resolution remote sensed imagery is used for data integration, analysis of urban areas, also to differentiate various zones like residential, industrial, and commercial. By reviewing a huge number of databases related to various articles on remote sensing data, features, and parameters, the information is summarized and shown in Table 1.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote Sensing Data</td>
<td>Hyper-spectral, Lidar, SAR, etc.</td>
</tr>
<tr>
<td>Target Study Area</td>
<td>Urban, Agriculture, Rural, Water</td>
</tr>
<tr>
<td>DL Model</td>
<td>CNN, RNN, AE, DBN, other</td>
</tr>
<tr>
<td>Target</td>
<td>Scene Classification, Image Fusion, Object Detection, Segmentation, LULC Classification, and other</td>
</tr>
<tr>
<td>Processing Parameters</td>
<td>Object, Pixel</td>
</tr>
<tr>
<td>Samples for Training</td>
<td>Value</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Value</td>
</tr>
<tr>
<td>Study Site</td>
<td>Value</td>
</tr>
<tr>
<td>Paper Category</td>
<td>Conference, Journal</td>
</tr>
<tr>
<td>Image Resolution</td>
<td>Value (high resolution, coarse, moderate)</td>
</tr>
</tbody>
</table>

Satellite data includes various other resolutions and types of images. Spatial Resolution is to measure the closed lines in an image. Spatial resolution is dependent on the device from which the image is captured. It is not only to measure the capacity of ppi (pixel per unit). The spatial resolution of any image decides its quality in the form of clarity. It generally refers to the count of independent pixels per unit in any image. Spatial resolution is limited by aberrations, diffraction, imperfect focus, and atmospheric distortion. Figure 1 (a), 1(b), 1(c) is showing the difference between multiple range spatial resolutions.

Spectral Resolution is to resolve spectral features as separate components, spectral resolution is used. Color images involve multiple and distinguished light effects on different spectra as Fig. 2 is showing. Multi-band images can resolve finer differences of wavelengths or spectrum by storing or measuring common RGB images.

Satellite data includes various other resolutions and types of images. Spatial Resolution is to measure the closed lines in an image. Spatial resolution is dependent on the device from which the image is captured. It is not only to measure the capacity of ppi (pixel per unit). The spatial resolution of any image decides its quality in the form of clarity. It generally refers to the count of independent pixels per unit in any image. Spatial resolution is limited by aberrations, diffraction, imperfect focus, and atmospheric distortion. Figure 1 (a), 1(b), 1(c) is showing the difference between multiple range spatial resolutions.

Spectral Resolution is to resolve spectral features as separate components, spectral resolution is used. Color images involve multiple and distinguished light effects on different spectra as Fig. 2 is showing. Multi-band images can resolve finer differences of wavelengths or spectrum by storing or measuring common RGB images.

A temporal resolution is a measurement unit of any area concerning time as a movie and high-speed cameras can capture the scenes at many different points in time. Time resolution that a general movie camera captures generally at a rate of 24-48 frames/ second. Although, high-speed cameras can capture the scenes at a very high speed up to 50-300 frames/ second or even more. Radio-metric Resolution is to determine the fine representation and differentiation among intensity, radiometric resolution is used. It is generally represented as the number of bits or levels. For ex- digital image is having 256 levels i.e. 28 bits. The reflected intensity will be better and finer than most as the bits are higher. While working practically, noise levels are used to limit radiometric resolutions instead of bits representation. Multi-Spectral Resolution is performed on the multi-spectral image the image data is captured across the range of electromagnetic spectrum at a specific wavelength. The wavelength can be captured with the help of supported devices that can separate to capture or detect by filters sometimes beyond the visibility range of light like- ultraviolet and infrared. Hyperspectral Resolution is generally applied on hyperspectral images is also a kind of multi-spectral image that captures the image data at several different wavelengths of the electromagnetic spectrum. To extract the data for each pixel in an image for detection of objects, material identification, hyperspectral images are used.

As remote sensing imagery is having various resolutions and dimensions of the data, rectification, and restoration of data became the important aspect for getting desired information or extracting hidden data by analyzing the image. The pre-processing operation is generally applied to correct and refine platform or sensor-specific radiometric data. It also includes the geometric distortion of the data. For eliminating scene illusions, sensor noise, etc. these types of pre-processing are required for remote imagery.
vious pre-processing methods are used to rectify the data collected by the sensor are included further. Each of these methods is different by their working nature or by having a different sensor or platform used for data acquisition.

**Radiometric Corrections.** Correction of data that includes the issues related to unwanted sensor data, irregularities of sensor data is the prime function of radiometric corrections. After the corrections, the data is converted to accurately measure the reflected light by the sensor.

**Contrast Enhancement.** It involves increasing in contrast value among background and target. An image histogram is a basic concept to understand contrast enhancement.

**Geometric Corrections.** For correcting the geometric distortion that occurs because of Earth-sensor geometrical variations and to convert that into real-world earth-surface coordinates, geometric corrections are used. Distortion can arise due to reasons that include altitude variations, the velocity of the sensor platform, earth convexity, atmospheric divergence, length of the displaced object.

**Spatial Filtering.** To reduce the smaller details in an image. Low pass spatial filters are designed to focus more on homogeneous large pixels of the same tone. This makes the smooth appearance of an image. These kinds of spatial filters are very much useful for reducing random noise from the image. Median and Average filters are an example of low pass spatial filters [34]. On contrary, High pass filters work just opposite the operation that a low pass filter does. It operates to sharpen the fine details in the image to fine-tune the appearance. Some filters like edge or directional detection filters are used to identify the field boundaries or roads in an image.

**Band Rationing.** One of the most commonly applied transformations is spectral rationing or band rationing. It serves to highlight and focus the spectral variations of surface covers.

**Piecewise Linear Stretch for contrast enhancement.** To utilize the complete range of value in brightness component, the maximum and minimum parameters of data allocated to new applied data. For e.g: an image is having a minimum brightness value of 45 and a maximum of 205. If that image (Fig. 3) is represented without enhancement, the values from 0 to 44 and the values from 206 up to 255 will not be displayed. By stretching the values from 0 to 45 and up to 205, the important features can be accessed.

![Fig. 3. Piecewise Linear Contrast Stretch](image)

**1.2 Satellite Image Processing Advancements**

Deep learning is a subset of Artificial intelligence that helps in creating automated applications and services for performing analytical physical tasks without human involvement. Now deep learning is behind each everyday product including digital assistants, self-driven cars, credit card fraud detection, and many more emerging techniques. Deep learning involves a neural network with a minimum of three layers. This network is developed to simulate human behavior through machines by learning them with a huge amount of data. Deep learning is an organized structure of multiple hidden layers. A single layer in a neural network is capable of predicting results; still, additional layers can be used for enhancing accuracy and efficiency [35].

Recent work in this field is showing that deep learning has achieved a lot in the field of replicating human behavior either in a simple task or complex operations like object detection, and image classification [36]. If comparing deep learning with other traditional approaches it is performing outstandingly in result predictions with good accuracy. Deep learning was firstly introduced in the 1980s and it has become the most emerging technology for serving the world in various domains. It requires a large amount of labeled data with the highest computational power to train the model to get more accurate results. Deep learning models learn features directly from data other outside feature extraction techniques are not required while working with a deep neural network. Deep learning can be classified into various categories like Supervised and Unsupervised. Feature extraction is one of the most important aspects of deep learning that uses an algorithm to construct meaningful features automatically for training, understanding, or learning.

From 1943 till now deep learning is being used and improving its applications day by day. Image processing through deep learning can be tracked since early 1943. In [58] created a neural network like the human brain by using a combination of algorithms and threshold logic. In [59], researchers have developed a continuous backpropagation model by improvising the basic neural networks. In 1962, a simple neural network model for image classification using the basic chain rule is developed. In 1965, a deep learning model for group data handling is developed. During the 1970s, the very first AI winter came into existence that uses some AI techniques for basic image processing. In 1973, Neocognitron that was an artificial neural network that used multilayered and hierarchical designs is created. The proposed system was able to recognize visual patterns through the computer. In [50] demonstrated backpropagation by combining CNN along with backpropagation for reading handwritten digits through the computer. In 1991, a model is created to identify the problems related to vanishing gradient. In [60] a paper is presented on deep belief networks for learning the images in a faster manner. For speech recognition, deep learning algorithms are developed. In 2009, ImageNet is launched to serve deep learning researchers. In 2012, AlexNet
is developed which is constructed with multiple GPUs concept. In [61] Generative Adversarial Networks is developed to enhance deep learning abilities in science, art, and fashion.

1.3 Advancement in Image Processing Computing Vision

Over recent years, Computer Vision tasks, such as object detection, image classification especially in digital images and videos grasp a lot of attention from researchers. Computer vision is a part of Artificial Intelligence that makes the computer enables everything that a human mind can perform from simple calculations to analyzing or reading minds [3]. Object detection is a technology related to Computer Vision that directly deals with the detection of objects from various classes [37]. Among all the problems and challenges available in image recognition, object detection is the major challenge to solve with the help of the Computer. Object detection is an emerging technology nowadays as it consists of various applications in various domains like Face Recognition, Video Surveillance, Crowd Counting, Transport Management, Image annotation, Video co-segmentation, tracking the locations of a ball during a cricket match, etc. Figure 4 is showing the milestones achieved in the field of object detection through traditional and deep learning methods.

Computer vision in favor of AI is touching the benchmarks in the world of the computer by increasing its computational power and results in calculation abilities with more accuracy and reliabilities. In [37] highlighted and focused on the role of computer vision in his work. Several deep learning methods are adopted to develop deep learning models for remote sensing images; some of them include transfer learning, learning rate decay, training from scratch, and dropout. Nowadays Artificial Intelligence and the various ways of computing of achieving the same are also in practice and evaluation. So, the researchers are looking the AI ways like Machine Learning and Deep Learning to attain new and amazing results in remote sensing imagery. Hence, during this study, we will be exploring various facets of remote sensing imagery, advanced artificial intelligence-based machine learning, and deep learning techniques for remote sensing imagery segmentation, object detection, and classification.

In [38] presented various examples of measuring space clustering processes for a range of three multispectral images captured over Ariz, Phoenix. In [51], NASA proposed the study at the Centre for Research, the University of Kansas with the cooperation and support of government research agencies and other universities have shown the applicability of satellite imagery in various fields within agriculture, oceanography, and earth sciences. This paper shows how characteristics and features of radar are used to get geo-science information.

In [39] [40], the authors described a process for spatial registration of multi-temporal and digital multi-spectral imagery. Experimental results are also defined here as the result of correlation analysis between digital satellite photographs and multi-spectral imagery. The registration process of space photography and multi-spectral airborne line-scanner digital imagery is also described here.

In [41], authors attempted to identify between welters of results to get unbeatable achievements also liked hurdles, also to assess the contribution to get expected image-processing operations in experimental and operational usage of the upcoming torrent of raw data.

Recent trends in deep learning techniques are emerging with powerful methods for automated systems that involve automatic feature learning through raw data. Most particularly, these methods achieved benchmarks in object detection; this area has become the most interesting area for new researchers [24].

Along with computer vision is touching the trends of the computer world in computational power with a tremendous speed, and result producing capacities with more reliability than human minds. In [37] highlighted the role of vision tasks with the help of certain projects to prove the tested approach for various research works like virtual skinning, virtual painting, human-computer interaction, depth recovery, etc.

![Object detection milestones](image)

**Fig. 4.** Object detection milestones [21] [24] DPM, RCNN

**Object Detection & Identification**

In [42] detected motion of objects using cellular neural network. In [43] provided a model to detect and focus objects in an unfocused background. In [45] developed an object-oriented based segmentation algorithm for edge extraction. In [46] developed a model to detect patterns in spectrograms using cellular neural network. In [47] developed a robust model based on deep learning methods to detect pedestrian using low- and high-level features. In [48] developed a model using CNN for logo detection of vehicles. In [53] provided their review for object detection algorithms for optical remote sensing imagery. They focused to review the brief history of available literature for deep learning techniques and the datasets for the detection of objects from remote imagery. Table 2 is representing the work (done in the period 2012-19) by many researchers on various datasets and extracted results for object detection in earth observation [2]. The review is given for dataset with the category of images, image count including width and the detected objects annotation style. A number of methods are developed for detection and recognition of objects. Many detector algorithms are proposed like ViJ (Viola Jones),
HOG through SVM, Feature dependent detectors \[45\] through DL techniques.

**Tab. 2. Differences among various open datasets (from 2012-2019) \[15\]**

<table>
<thead>
<tr>
<th>Year</th>
<th>Datasets</th>
<th>Instances</th>
<th>Categories</th>
<th>Image Width</th>
<th>Images</th>
<th>Annotation Style</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>TAS</td>
<td>1319</td>
<td>1</td>
<td>792</td>
<td>30</td>
<td>Horizontal Box</td>
</tr>
<tr>
<td>2012</td>
<td>SZTAKI- INRIA</td>
<td>665</td>
<td>1</td>
<td>800</td>
<td>9</td>
<td>Oriented Box</td>
</tr>
<tr>
<td>2014</td>
<td>NWPU VHR-10</td>
<td>3775</td>
<td>10</td>
<td>1000</td>
<td>800</td>
<td>Horizontal Box</td>
</tr>
<tr>
<td>2015</td>
<td>VEDAI</td>
<td>3640</td>
<td>9</td>
<td>1024</td>
<td>1210</td>
<td>Oriented Box</td>
</tr>
<tr>
<td>2015</td>
<td>UCAS-AOD</td>
<td>6029</td>
<td>2</td>
<td>1280</td>
<td>910</td>
<td>Horizontal Box</td>
</tr>
<tr>
<td>2015</td>
<td>DLR 3K Vehicle</td>
<td>14235</td>
<td>2</td>
<td>5616</td>
<td>20</td>
<td>Oriented Box</td>
</tr>
<tr>
<td>2016</td>
<td>HRSC2016</td>
<td>2976</td>
<td>1</td>
<td>1000</td>
<td>1070</td>
<td>Oriented Box</td>
</tr>
<tr>
<td>2017</td>
<td>RSOD</td>
<td>6950</td>
<td>4</td>
<td>1000</td>
<td>976</td>
<td>Horizontal Box</td>
</tr>
<tr>
<td>2017</td>
<td>DOTA</td>
<td>188282</td>
<td>15</td>
<td>800-4000</td>
<td>2806</td>
<td>Oriented Box</td>
</tr>
<tr>
<td>2018</td>
<td>DIOR</td>
<td>192472</td>
<td>20</td>
<td>800</td>
<td>23463</td>
<td>Horizontal Box</td>
</tr>
</tbody>
</table>

**Deep CNN Architecture**

Unlike traditional neural networks, CNNs use convolution operation in their layer \[50\]. CNN’s convolution operation includes multiple stages consisting of four main components, kernel, convolution layer, activation function, and a pooling layer. Each stage represents a feature map in the form of an array \[34\]. Figure 5 is showing the detailed workflow of CNN having several convolution layers with one or more fully connected layers, for producing the final output as a classification module. The main components of CNN are described as follows.

Fig. 5. Convolutional Operation with image matrix (4×4) with kernel matrix (3×3)

**Kernels.** At each location, each kernel is aimed to detect a particular characteristic. There exists a bank of m1 filters in every convolutional layer; the output of the lth layer consisting of feature maps of size \(*\). The ith feature map can be computed as follows:

\[ Y_i^{(l)} = B_i^{(l)} + \sum_{j=1}^{m_1^{(l-1)}} K_{ij}^{(l)} * Y_j^{(l-1)} \]  

Here, represents the trainable bias parameter matrix, which is a filter that connects the jth feature map of the \((l-1)\) layer with ith feature map of layer \(l\). \(*\) is a 2-D discrete convolution operator.

**Convolution Layer.** Convolutional layer is the collection of multiple layers that can calculate the convolution of the inputted image by network weight. The first layer consists of the neuron that can view small images and learn some basic and limited features through it. As the network goes deep inside, layers can view a large portion of the image and can learn more expressive and detailed features by merging previous levels. Each layer is characterized by hyper-parameters to train through spatial features, zero padding, and stride values between various windows that work to control the output layer. The expression can be given as follows:

\[ S_{i,j} = (I * K)_{i,j} = \sum_m \sum_n I_{i,n} * K_{i-m,j-n} \]  

In the process of convolution, a small window slides across the image from left to right and top to bottom. At each sliding window location, the sum of the product is calculated with each kernel element with the input element. The process continues with different kernels to form various kinds of feature maps.

The feature map is having a lesser size than the input image. However, we can pad the values to keep the size the same. The stride shows the gap size among two successive positions. Commonly, a stride size of 1 is chosen, but a greater size can be chosen to reduce the resolution of feature maps.

**Nonlinear activation function.** By getting the feature map, a nonlinear activation function is applied in this process. As Eq. 3 is showing, the calculation functions of activation maps carried only the activated features to the next layer. Then, the activation function can be formed as follows

\[ \phi(Y_i^{(l)}) = \left( B_i^{(l)} + \sum_{j=1}^{m_1^{(l-1)}} K_{ij}^{(l)} * Y_j^{(l-1)} \right) \]  

Some activation functions that can be used in the convolution process are as follows:

**Rectified Linear Unit function (ReLU).** Most commonly used activation function. It is called a half-rectified function. Figure 6 (a) is showing ReLU the graphical representation of the ReLU activation function. Mathematically it can be defined as

\[ f(x) = \max(0, x) = \begin{cases} 0 & if \ x < 0 \\ x & if \ x \geq 0 \end{cases} \]  

**Sigmoid Function.** As figure 6 (b) is showing, it can be represented by a curve like “S”. It is generally used to predict the results as the function varies between 0 and 1. It can be defined as

\[ f(x) = \frac{1}{1 + e^{-x}} \]  

**Hyperbolic Tangent (tanh) function.** It is very much similar to the sigmoid function as it can be seen in figure 6 (c). The difference is in range, the range is here is [-1, 1]. The benefit of using the tanh function
is that here negative numbers are mapped as strongly negative and zero values mapped near zero.

\[ f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (6) \]

**Normalization Layers.** To implement inhibition paradigms, the normalization layer is used for the observation of the biological brain [62].

**Pooling Layer.** After each successive convolution layer, the pooling layer is presented that can reduce input layer size via some non-linear functions. They also help in reducing the computational and parametric amount in the network. It also helps to control over-fitting [63]. Figure 7 represents the pooling operation with [2 x 2] filter. Following are some main pooling operations:

- **Max pooling.** The maximum value is calculated for each input patch. It preserves the maximum value of each stride while sliding over the feature map. Its mathematical representation is as follows

\[ f_{\text{max}}(A) = \max_{m,n}(A_{nm}) \quad (7) \]

- **Average pooling.** For each input, it calculates the average value. This layer divides the input various pooling regions by computing their average values.

\[ f_{\text{ave}}(A) = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} \quad (8) \]

**Fig. 6.** Representation of various activation functions: a- ReLU b- Sigmoid, c- tanh

**Fig. 7.** Operation of Pooling using 2 x 2 filters with stride of two

**Dropout Layer.** This layer is used to help in over-fitting issues also improves the network’s performance. The dropout layer can be applied in any of the layers.

**Fully Connected Layer.** At this layer, the final output is represented in the form of a 2-D array by connected with a fully connected layer. By the result of the Convolution process, this layer classifies an image into various classes. The activation function at the last layer computes the probability of results belonging to each class [16]. Commonly, for multi-class classification, the softmax activation function is used having a range of probability between [0, 1] with a sum equal to 1. The fully connected layer is situated at the last layer in a neural network. The fully connected layer is having the neurons that have a full connection to all the neurons of previous layers [63]. The calculations of this connection can be evaluated by matrix multiplication function later followed by bias offset [62].

The remainder of this paper is being organized as section-2 is showing the various existing pre-trained models on ImageNet and CIFAR-10 datasets, section-3 is giving the incentive contributions at beginning of remote sense image processing, section-4 is the detailed literature review as a tabulated formation categorized by various technological benchmarks used, section-5 shows the challenges mapping with reviewed literature and section 6 involves the complete summary.

## 2. Existing Pre-trained Deep Learning Models and Datasets

Deep neural networks, such as Convolutional neural networks [64], Recurrent neural networks [65], Graph neural networks [66], Attention neural networks [67], have been applied for various AI tasks at a wide range. The pre-trained models can be considered as the models that are developed based on any existing techniques (As mentioned various deep learning techniques). The models that have been already got trained on some dataset like ImageNet, CifarNet-10 are known as pre-trained models. Nowadays pre-trained models are widely used for image classification and object detection. Table 3, 4are showing the summary and comparison of various pre-trained models trained on ImageNet and CIFAR-10 dataset respectively.

**Tab. 3.** Various pre-trained models trained on ImageNet dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Depth</th>
<th>Size (MB)</th>
<th>Parameters (Million)</th>
<th>Top-1 Accuracy</th>
<th>Top-5 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xception</td>
<td>81</td>
<td>88</td>
<td>22.9M</td>
<td>79.0%</td>
<td>94.5%</td>
</tr>
<tr>
<td>VGG16</td>
<td>16</td>
<td>528</td>
<td>138.4M</td>
<td>71.3%</td>
<td>90.1%</td>
</tr>
<tr>
<td>VGG19</td>
<td>19</td>
<td>549</td>
<td>143.7M</td>
<td>71.3%</td>
<td>90.0%</td>
</tr>
<tr>
<td>ResNet50</td>
<td>107</td>
<td>98</td>
<td>25.6M</td>
<td>74.9%</td>
<td>92.1%</td>
</tr>
<tr>
<td>Inceptionv3</td>
<td>189</td>
<td>92</td>
<td>23.9M</td>
<td>77.9%</td>
<td>93.7%</td>
</tr>
<tr>
<td>DenseNet201</td>
<td>402</td>
<td>80</td>
<td>20.2M</td>
<td>77.3%</td>
<td>93.6%</td>
</tr>
<tr>
<td>EfficientNetB0</td>
<td>132</td>
<td>29</td>
<td>5.3M</td>
<td>77.1%</td>
<td>93.3%</td>
</tr>
</tbody>
</table>

**Tab. 4.** Various pre-trained models trained on CIFAR-10
### 3. Literature Review

In [6] authors presented a model for the interpretation and evaluation of scenes for image analysis.

In paper [5] authors proposed a blackboard model as a control structure for the detection of objects in aerial images.

In [7] researchers focused their research on feature extraction in SAR images as this process has a higher level of complexity due to the level of noise and quality issues in these images. This paper implemented an automated algorithm for feature extraction of large-scale objects from SAR imagery. For the experiments, images of the Ottawa area are used through SAR imagery.

In [8] authors proposed a model to detect a cover change in the forest on Landsat data.

In [9] authors developed a model for the detection of manmade objects like airports, bridges, industries, etc.

In [10] authors proposed a classification and detection model named CADCM to target hidden objects in hyperspectral imagery. The process is accomplished in three phases. Initially, a band selection process is applied, next band rationing is done and finally, automatic target detection is achieved. Results show the targets hidden by natural background, shades, or objects can be detected finely.

In [11] two matrices namely object space and image space are used to refer and monitor the existing monocular building detection system with the help of 83 images collected by 18 different sites. By the analysis, the effects of image inscrutability along with objects complexity are examined. Edge fragmentation is also shown here in this research. The usage of rigorous photogrammetric space modeling is also demonstrated.

In [12] a review article is presented to reduce cloud impacts by analyzing various existing algorithms.

In [13] an automatic approach for building footprint extraction and its 3-D reconstruction from the imagery of airborne light and ranging (LIDAR) data is represented. Initially, a digital surface model (DSM) is generated to extract objects higher than the ground surface. To separate a building from other objects, geometric characteristics such as size, height, and shape are used. Extracted building footprints are simplified for better quality using an orthogonal algorithm. Roofs are identified by information like ridgelines and slopes. Finally, an accuracy assessment is conducted by comparing the results with manually digitized building reference data.

In [14] authors represented the image analysis method for the extraction of building features. They have used three consecutive steps to accomplish this task. Initially, the supervised neural network is inputted by RGB multi-band images for roof identification. Next, spatial details are extracted through a hybrid approach of edge and region segmentation. Lastly, the extracted information is used to refine the results.

In [15] authors proposed a building extraction method. GIS data is used as an input in this method. A segmentation algorithm is used to extract the features of the building. GIS data is used to provide prior building knowledge. Data pre-processing, Object segmentation, and result post-processing are the three steps used in this method. Experimental results are also included in this to showcase the efficiency of the algorithms.

In [17] a two-step model method for tree detection is implemented including segmentation followed by classification is proposed here. The results presented show the effectiveness of the approach.

In [18] authors detected bridges in multispectral remote images through their developed model. The multi-seed supervised classification technique is used to classify the multispectral image into eight land-cover types. A knowledge-based approach is used that find out the spatial arrangement of the bridge and its surroundings. Testing is done on the IRS-1C/1-D satellite has a spatial resolution of 23.5m.

In [19] authors investigated ship detection in TerraSAR-X (TSX) ScanSAR images (19-m resolution). Kolmogorov-Smirnov test is applied for the verification of the goodness of fit for the K-distribution to TSX images. A target detection algorithm is developed and also verified.

In [20] amorphous-shaped objects are detected by their developed model. The model is showing the results of experiments achieving high accuracy rates.

In [21] an automatic content-based analysis is presented to detect arbitrary objects in aerial imagery. In this, the two-stage training model using a convolutional neural network is implemented also verified over remote imagery. Model is tested for accuracy using UC-Merced data set with an accuracy of 98.6%.

In [22] a deep CNN model is invented with enhanced functionality for feature extraction along with region classification and region proposal. Their method is based on ResNets that consists of multiple sub-networks (Object detection and Object proposal). To enhance feature map resolution, the output generated by multiple-scale layers is combined. VHR-10 dataset is used to train the model in the proposed work.

In [23] two algorithms are studied (Edgeboxes and Selective Search) for object detection. The evaluation is also performed on both algorithms using high-resolution remote imagery. Algorithms are tested and evaluated through the NWPU-VHR-10 class data set. For performance measurement, execution time and recall rate are used as performance parameters. By the statistical results, authors proved that EdgeBoxes algorithm is showing optimal results over Selective Search in recall rate.

In [25] a feature fusion method is proposed for extracting the fine-grained features from multiple layers of the remotely sensed image as those images have dense features concerning intra-class differenc-
es, high inter-class similarities, and multidirectional objects. They have initially used ResNet50 for extracting the features from multiple layers then the channel attention method is used to enhance the features. As a result, cross-layer bilinear pooling and feature connection are used for fusion. After extracting the features through ResNet50, squeezing-and-excitation (SE) method is used in its advanced version. Along with this, the traditional activation functions are replaced with the HandTan function which is simpler and more effective than traditional activation functions. The deep learning framework PyTorch is used to build the training model. The comparison is shown in their research with the existing models.

In [27] a branch regression framework is presented in a dual-mode based on remote image observations in detecting targets. This model can independently predict various variables and orientations effectively. To deal with multi-level features through spatial pooling, an advanced smart feature is added to the research. This is an example of an advanced, accurate model that is able in doing parallel operations of localization and classification.

In 2020, Yang et al. [28] suggested a novel cloud detection neural network with an encoder-decoder structure, called CDnetV2 as a sequence of work in the detection of Cloud. If Channel Attention Fusion Model (CAFM), Spatial Attention Fusion Model (SAFM), and Channel Attention Refinement Model (CARM). Proposed HFFM is used to extract the linguistic high-level information from HLSIGFs. Here, some experimental results on the ZY-3 satellite thumbnail data set show that the proposed CDnetV2 is achieving exact and accurate results and it is outperforming several state-of-the-art methods.

In [29] authors proposed a novel approach to cope with such kinds of variant labels, i.e., class attention module and decomposition-fusion strategy. A class attention module is created to generate multiple class attention modules. Salient detection is proposed which breaks down semantic segmentation into multi-class major detection and then combines them to produce a semantic map. Some experiments have been done on US3D Dataset. The imbalance label problem is also resolved with more accuracy than the previously available approach.

In [26] PTAN is developed including three-stage strategies for object detection in HD images. The model is achieving a mAP value of 0.7958. On the NWPU VHR-10 dataset, PTAN is achieving a mAP value of 0.9187. In [4] authors proposed an up-sampling, down-sampling feature pyramid for obtaining the richer context information by bi-directionally involving shallow and deep features, and skipping connections. Experiments are done on DIOR, NWPUVHR-10, and on the self-assembled datasets SDOTA, SDD to show the excellent performance of the proposed method by comparing it with other detectors. The proposed method is achieving 74.3% mAP on the public DIOR dataset.

In [1] authors proposed EFPN to detect small objects like plants, small buildings, etc. It is created to improve feature extraction capabilities.

In [3] authors recommended a dynamic curriculum procedure that can learn the object detectors with the help of training images.

4. Literature Review Findings

Table-5 is showing the major findings by thoroughly reviewing many articles and research papers on object detection using deep learning and machine learning. From the past till now many researchers have presented their research work and ideas for feature extraction [7] and object detection including various kinds of objects in remotely sensed images [5] like buildings [13], trees, plants [1], roads, commercial areas, water, quality of soil, ships [4], oil slicks, bridges, industries, airports [9], shadow detection [10] and cloud detection [3].

Tab.5. Literature Review Findings

<table>
<thead>
<tr>
<th>Reference</th>
<th>Algorithm/Model Methodology</th>
<th>DataSet</th>
<th>Parameter</th>
<th>Research Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>A feature fusion method based on improvised ResNet50</td>
<td>Open access Remote imagery</td>
<td>Validation of dataset</td>
<td>Need to be more accurate</td>
</tr>
<tr>
<td>[3]</td>
<td>MSCNN</td>
<td>Geospatial VHR Satellite images NWPU VHR-10 Challenging dataset 650 images with a resolution of 0.5–2.0 m.</td>
<td>Accuracy</td>
<td>Scale and rotation dependent</td>
</tr>
<tr>
<td>[26]</td>
<td>CDnetV2</td>
<td>ZY-3 satellite thumbnail</td>
<td>Accuracy with fine grained features</td>
<td>Model is providing various accuracy rations in variant models. Need to create a hybrid model with more accuracy.</td>
</tr>
<tr>
<td>[4]</td>
<td>Rotated Region Proposal Network</td>
<td>HRSC2016 dataset.</td>
<td>work on dense dataset</td>
<td>More classes and sub-classes can be included</td>
</tr>
<tr>
<td>Reference</td>
<td>Algorithm/Model Methodology</td>
<td>DataSet</td>
<td>Parameter</td>
<td>Research Gap</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------</td>
<td>---------</td>
<td>-----------</td>
<td>--------------</td>
</tr>
<tr>
<td>[49]</td>
<td>object relationship reasoning CNN (ORRCNN)</td>
<td>Aerial Image data set (AID), UC Merced Land-Use data set, and WHU-RS19 data set</td>
<td>accuracy for multiband data</td>
<td>Need more prior information and parameters of the geometrical shape for template designing</td>
</tr>
<tr>
<td>[69]</td>
<td>Deep learning algorithms on NVIDIA DGX-1 supercomputer</td>
<td>Pre-trained dataset of SpaceNet fine-tuned on planet database.</td>
<td>Time Complexity &amp; Efficiency</td>
<td>For the training of CNN, a huge set of training data along with more computation powers is needed</td>
</tr>
<tr>
<td>[22]</td>
<td>An enhanced deep CNN based</td>
<td>VHR-10 data set</td>
<td>substantial number of densely packed objects</td>
<td>Need to be more enhanced</td>
</tr>
<tr>
<td>[21]</td>
<td>Two-stage training model using convolutional neural network</td>
<td>UC Merced Dataset</td>
<td>Arbitrary objects</td>
<td>Sometimes outliers are involved in predictions</td>
</tr>
<tr>
<td>[51]</td>
<td>AASM</td>
<td>Open access satellite images</td>
<td>Ability and Efficiency</td>
<td>Sensitive to shape and viewpoint change</td>
</tr>
<tr>
<td>[35]</td>
<td>Convolutional capsule network</td>
<td>Open access Remote imagery</td>
<td>fine grained features from multiple layers</td>
<td>Unable in detection for robust dataset</td>
</tr>
<tr>
<td>[26]</td>
<td>hierarchical bilinear pooling (HBP) with hierarchical attention and bilinear fusion net HABPNet</td>
<td>1. UC Merced dataset released in 2010 2. AID dataset released in 2017 3. NWPU-RESISC45 dataset released in 2017</td>
<td>Accuracy with fine grained features</td>
<td>Limit number images are used for the testing and training.</td>
</tr>
<tr>
<td>[68]</td>
<td>RTANet</td>
<td>Publicly available open dataset</td>
<td>Speed</td>
<td>Prediction accuracy is limited in number of images given for testing</td>
</tr>
<tr>
<td>[24]</td>
<td>optical remote sensing video (ORSV)</td>
<td>ORSV images</td>
<td>motion-drive</td>
<td>Need to be changed in terms of methodological view</td>
</tr>
<tr>
<td>[23]</td>
<td>Selective Search and EdgeBoxes</td>
<td>NWPU VHR-10</td>
<td>Involves high recall rate, faster</td>
<td>Class imbalance issues occurring</td>
</tr>
</tbody>
</table>

**Benchmark Technique: R-CNN Deep Learning**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Algorithm/Model Methodology</th>
<th>DataSet</th>
<th>Parameter</th>
<th>Research Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>[27]</td>
<td>compatibility loss clustering method (CLCM)</td>
<td>1. DOTA 2. U/CAS-AOD 3. NWPU VHR-10 4. RSOD-Dataset</td>
<td>Accuracy &amp; efficiency</td>
<td>More layers can be added to enhance the accuracy of prediction.</td>
</tr>
<tr>
<td>[59]</td>
<td>R-CNN algorithm with dialed convolution</td>
<td>HRSC2016 dataset</td>
<td>Accuracy with respect to its feature extraction</td>
<td>Computationally expensive</td>
</tr>
<tr>
<td>[49]</td>
<td>CFEM, A context-based feature enhancement module</td>
<td>ISPRS Vaihingen data set</td>
<td>Speed &amp; Accuracy</td>
<td>Traditional Neural Network approach is using, need to be more enhance with respect to methodology</td>
</tr>
<tr>
<td>[25]</td>
<td>PTAN (A patch-based three-stage aggregation network)</td>
<td>1. DOTA 2. NWPU VHR-10</td>
<td>Accuracy &amp; efficiency</td>
<td>Need to enhance the performance</td>
</tr>
<tr>
<td>[29]</td>
<td>Class attention module with multi-class segmentation network</td>
<td>US3D Dataset</td>
<td>Accuracy &amp; efficiency</td>
<td>Results need to be improved on various parameters</td>
</tr>
</tbody>
</table>

**Benchmark Technique: Machine Learning**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Algorithm/Model Methodology</th>
<th>DataSet</th>
<th>Parameter</th>
<th>Research Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>[5]</td>
<td>Blackboard model</td>
<td>Expert systems for image processing</td>
<td>knowledge representation</td>
<td>Real time detection is not possible</td>
</tr>
<tr>
<td>[6]</td>
<td>MultiExpert System for Scene Interpretation and Evaluation (MESSIE)</td>
<td>Suburban images</td>
<td>Class of an object from general structure</td>
<td>Multiple spatial scales and aspect ratio</td>
</tr>
<tr>
<td>[7]</td>
<td>automated algorithm for feature extraction</td>
<td>Images of Ottawa area are used through SAR imagery</td>
<td>Detection of homogeneous areas</td>
<td>Limited data is used</td>
</tr>
<tr>
<td>[8]</td>
<td>A correlation mechanism for bi-temporal band pairs</td>
<td>Multi-temporal Landsat TM data</td>
<td>water reflectance</td>
<td>Traditional approach is used</td>
</tr>
</tbody>
</table>
Reference | Algorithm/Model Methodology | DataSet | Parameter | Research Gap
---|---|---|---|---
[9] | A multi-valued recognition system | IRS satellite Imagery | Multiple class selection | Not showing more accuracy for multiple sub classes data
[10] | CADCM | HYDICE Hyperspectral digital imagery | Hidden targets in hyperspectral images | Only working for hidden targets, also including shadows
[11] | Object space and image space-based matrices | 83 images of 18 sites | Performance evaluation | Limited data is tested
[12] | An automatic recognition system | Various open access databases | minimizing the deleterious effects of cloud | Not tested on a valid dataset
[13] | digital surface model (DSM) | Imagery of airborne light and ranging (LIDAR) | building footprint extraction | Not able to fine-grained the results due to limited classes used
[15] | A segmentation algorithm | GIS data | efficiency of the algorithms | Need to enhance the performance
[36] | SE-MGMM | synthetic aperture radar images | change detection | System is not obtaining a good accuracy for large amount of data or on live images.
[16] | Multistage model for road detection | SPOT multispectral images of district near Hongqiao Airport | improved detection probability | Model is showing multiple areas of Class imbalance
[17] | Two-step approach | LIDAR aerial image | Segmentation using weighted features | multiple scales are needed to be applied
[18] | A model to detect bridges over water bodies | IRS-1C/1-D satellite images of 2.3.5. | multispectral imagery | Dual priorities
[19] | Kolmogorov-Smirnov | TerraSAR-X (TSX) ScanSAR images (19-m resolution) | Ship Detection | Various overfitting problems involved
[20] | neighborhood model | hypothesis imagery and DIRSIG | Amorphously shaped objects | Need to be more enhanced in terms of accuracy
[55] | An algorithm for building shadow detection | panchromatic satellite imagery | Shadow detection | More classes and sub-classes can be included

5. Research Gaps & Challenges in Existing Methods
The goal of object detection is to achieve the highest accuracy with efficiency by developing an automated robust detection algorithm. By critically reviewing the existing work done in this field it is analyzed that two major challenges still exists in finding or detection of objects in remote sensing images. We have divided the found research gap in two categories, accuracy and efficiency. Accuracy can affect due to various reasons like class imbalance, captured image condition or environment, image noise, dual priorities, etc. Table 6 is showing the complete mapping of found research gap or challenges with the reviewed literature in this paper. By critically reviewing many articles it is analyzed that there is still a gap exists in finding out the optimized solution in object detection in remote sensed images.

5.1 Challenges Involved in Achieving High Accuracy
A. Internal class imbalance. One of the major challenges that a model faces while dealing with real objects like shapes, sizes, colors, or directions of objects is class imbalance. This issue can occur due to the mentioned reasons a model could not be able to detect the same objects having a different shape, color, or pose in multiple images [3][7][13][21][35][26].

B. Imagery conditions and unconstrained environments. Factors that include lighting, occlusion, weather conditions, viewpoint, object physical location, shadow clutter, blur, motion, etc. [8] [11] [41] [22] [9].

C. Imaging noise. Imaging noise is one of the challenges in achieving high accuracy. Also factors like compression noise, low-resolution images, and filter distortions [47] [35] [17] [43] [20].

5.2 Challenges Involved in Achieving High Efficiency
Low computational devices like mobile have low memory and less computational speed; it is a bottleneck in detecting objects with high efficiency [39]. Millions of unstructured and structured real-world existing object categories for distinguishing are
a challenge for the detector [19] [23] [36]. Sometimes image data especially remote images are having a large size this became a challenging situation for object detectors. To find some unseen objects is also a challenge [44] [33].

**Tab. 6. Challenges Mapping with Reviewed Literature**

<table>
<thead>
<tr>
<th>Research Gaps &amp; Challenges in Existing methods</th>
<th>Internal Class Imbalance</th>
<th>Imagery Conditions and Unconstrained Environments</th>
<th>Imaging Noise</th>
<th>Challenges In Achieving High Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>[47] Takarli et al. [47]</td>
<td></td>
<td>[36] Tollhuoglu et al. [36]</td>
</tr>
</tbody>
</table>

6. Conclusion

This paper is mainly focusing on a thorough, comprehensive, and chronological study of existing work done in the field of AI including machine learning and deep learning. This paper is performing the role of a pathway to all new researchers who are wishing to work in this field. This review paper is not only focusing on the published papers on object detection in remote imagery. It also involves the review of the papers those are having a modern approach of deep learning like CNN [2] [3] [4], R-CNN [6] [9] [11] [12] [15] and CornerNet [32] [35], etc. The paper is proving a complete outlook on existing ML and DL frameworks, and the existing datasets used in object detection. Some problems and challenges in computing vision are also discussed here like crowd detection, color imbalance, live detection, etc. This paper is reviewing some most relatable topics like Ship detection, building detection, geospatial object detection, cloud detection, small-scale objects with the highest resolution in remote images. Some papers which involve transfer learning concepts are also reviewed in this literature review.

**AUTHORS**

Hina Hashmi* – CCSIT, Teerthankaer Mahaveer University, Moradabad, India, e-mail: hinahashmi170@gmail.com,
Rakesh Dwivedi – CCSIT, Teerthankaer Mahaveer University, Moradabad, India,
Anil Kumar – Indian Institute of Remote Sensing, Dehradun, India.
*Corresponding author

**REFERENCES**

[10] Hsuan Ren and Chein-I Chang, “A computer-aided detection and classification method for concealed targets in hyperspectral imagery”. In:


