AIVP ADVANCES IN IMAGE AND VIDEO PROCESSING
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 7, ISSUE 5
ISSN 2054-7412

A MASK-RCNN Based Approach Using Scale Invariant Feature Transform Key points for Object Detection from Uniform Background Scene

Radhamadhab Dalai[1], Kishore Kumar Senapati[2]
[1]*Asssitant Professor, Birla Institute of Technology Mesra , Ranchi , India,*
[2] *Senior Asssitant Professor, Birla Institute of Technology Mesra , Ranchi , India,*
rdalai.teqip@bitmesra.ac.in

ABSTRACT

Object identification using deep learning in known environment gives a new dimension to the research area of computer vision based automation system. As it uses supervised learning technique using Convolution Neural Network (RCNN) it helps automation software tools and machines to detect and identify objects using vision based systems. One of RCNN technique known as Mask-RCNN has been applied in this proposed design and this paper presents a novel approach to object detection problem using Big Data storage for large set of features based data. Earlier work Faster Region-based CNN has led to the development of a state-of-the-art object detector termed as Mask R-CNN. Some samples of solid material objects used in refractory industry have been taken as input images. In our experiment the SIFT based features have been implemented and trained using filter and convolution operation. In addition to improved accuracy, pixel-level annotation (annotating bounding boxes is approximately an order of magnitude which is quicker to perform). The model is retrained to perform the detection of four types of metal objects, with the entire process of annotation and training for the new model per solid block. A key benefit of feature based Mask-RCNN approach is high precision (~94%) in classification and minimized feature points with SIFT key points.

Keywords: Bigdata, CNN, Deep Learning, Feature points, Mask-RCNN, SIFT

1 Introduction

The deep Learning algorithms are interesting domains of research into the automated extraction of complex data representations (features) at high levels of abstraction. RCNN, FAST-RCNN, Mask-RCNN, Autoencoder are different kinds of deep learning algorithms as mentioned in [16]. Using such algorithms development of a layered, hierarchical architecture of learning and classifying image object, where higher-level (more abstract) features are defined in terms of lower-level features using Convolution Neural Network design. This technique can be very useful in detecting objects such as building materials, mined ores from distinctive features such as texture and regular shape.

Deep learning algorithms [14] lead to abstract representations because more abstract representations are often constructed based on less abstract ones such as shape and texture of image. An important advantage of more abstract representations is that they can be invariant to the local changes in the input data. Learning such invariant features is an ongoing major goal in pattern recognition (for example learning features that are invariant to the object orientation in an object detection task). Beyond being invariant such representations can also keep intact the factors of variation in data. The

DOI: 10.14738/aivp.75.6946
Publication Date: 28th September, 2019
URL: http://dx.doi.org/10.14738/aivp.75.6946

real data used in AI-related tasks mostly arise from complicated interactions of many sources. For example an image is composed of different sources of variations such a light, object shapes, and object materials. The abstract representations provided by deep learning algorithms can separate the different sources of variations in data.

Deep learning algorithms are actually deep architectures of consecutive layers. Each layer applies a nonlinear transformation on its input and provides a representation in its output. The objective is to learn a complicated and abstract representation of the data in a hierarchical manner by passing the data through multiple transformation layers. The sensory data (for example pixels in an image) is fed to the first layer. Consequently the output of each layer is provided as input to its next layer.

Mask R-CNN [5] is a simple and efficient technique to classify and identify objects. Among various design of RCNN techniques Mask RCNN is the extension of Faster R-CNN which has two outputs for each candidate object, a class label and a bounding-box offset. A third branch that out-puts the object mask has been added to this. Here but the additional mask output is distinct from the class and box outputs, requiring extraction of much finer spatial layout of an object. On next, the key element of Mask R-CNN has been introduced, including pixel-to-pixel alignment, which is the main missing piece of Fast/Faster R-CNN.

1.1 Collaboration of Deep learning and Bigdata

CNN based algorithms [21] consists of several level of pooling layer, multiple Region of Interest(ROI) Layer and a number (10-15) of filter layers. It is a very efficient tool to extract meaningful abstract representations (features) of the raw data such as image object through the use of a hierarchical multi-level learning approach. In a higher-level presentation such as image more abstract and complex representations of features are learned based on the less abstract layers, features and representations in the lower levels (pixel values) of the learning hierarchy from lower to higher layer. Hence the pooling layers are used to solve the above issue. While Deep Learning can be applied to learn from labeled data if it is available in sufficiently large amounts, it is primarily efiicient for learning from large amounts of unlabeled or unsupervised data, making it effiecient and robust for extracting meaningful representations and patterns using Big Data storage. Once the hierarchical data abstractions are learned from unsupervised data with RCNN, more conventional discriminative models can be trained with the aid of relatively fewer supervised or labeled data points, where the labeled data is typically obtained through human or expert input. Deep Learning algorithms are shown to perform better at extracting non-local and global relationships and patterns in the data, compared to relatively shallow learning architectures [4] such as segmenting regions. Other useful characteristics of the learnt abstract representations by Deep Learning include: (1) relatively simple linear models can work effectively with the knowledge obtained from the more complex and more abstract data representations, (2) increased automation of data representation extraction from unsupervised data enables its broad application to different data types, such as image, textural, audio, etc., and (3) relational and semantic knowledge can be obtained at the higher levels of abstraction and representation of the raw data. While there are other useful aspects of RCNN based classification of image object, the specific characteristics mentioned above are particularly important for object identification. For each feature data for n number of attributes there will be m^n numbers of feature vector set. Instead of using PCA based redundancy the full feature suit has been selected in order to avoid missing major key points.

Deep Learning algorithms and architectures are normally chosen to address issues on object classification and recognition because of its deep level design (pooling) and efficient filter mechanism. In our proposed design the feature extraction, analysis, and object detection from image is achieved by applying MASK-RCNN(Region Proposal Convolutional Neural Networks) convolution network. For this purpose a huge number feature sets are stored in data storage whose architecture has been discussed in next section.Here in our work 10-15 feature sets are used and stored in Bigdata based storage. Deep Learning inherently exploits the availability of massive amounts of data, i.e. volume in Big Data, where algorithms in terms of mltiple layers with shallow learning hierarchies fail to explore and understand the higher complexities of data patterns.

In this section, dataset design for feature points have been explained and on next RCNN algorithms have been implemented using our current dataset architecture to optimize a few parameters including semantic indexing, discriminative tasks for matching, and data tagging. Our focus is that by presenting these works in Deep Learning algorithms using highly robust design architecture. Deep Learning techniques in Big Data showcases the application domains in the works presented involve large scale data. These algorithms are applicable to different kinds of input data; however, in this section we focus on its application on a particular vision based system which works in day to day life environment such as mining industry based machinery, robotic system.

A key task associated with constructing dataset with bigdata is decisive feature preserving and securing important feature points. In case of using redundant algorithms such as PCA there is probability of missing a few major key points which may affect distinctive features. Efficient storage and retrieval of information is a growing problem in Big Data, particularly since very large-scale quantities of dataset generated from sift operations on image. The operation such as retrieval of matching key point is twofold as shown below.

$$K_i = f(x_1) + f(x_2) + f(x_3) + \ldots\ldots\ldots + f(x_n) \quad (1)$$

$f(x_i)$ = each individual feature point function generated by algorithm.

$f(x_i)$ = HashFunction $(f(x_i)*\alpha)$

α = multiplicative factor after finding mean of same feature point's pixel values in all key points

Earlier mechanism for information storage and retrieval are challenged by the massive volumes of data and different data representations, both associated with huge number of operations. In these systems, massive amounts of data are available that needs semantic indexing rather than being stored as data bit strings. Semantic indexing presents the data in a more efficient manner and makes it useful as a source for knowledge discovery and comprehension, for example instead of using direct pixel values normalized key value (m1 <= key value <= m2) based on generate function (Hash Function)[10].

```
function HashFunction (features: array of string, N:   integer):
    x: = new vector [N]
    for f in features:
       h: = hash (f)
       x[h mod N] += 1
    return x
```

Instead of using raw input for data indexing, RCNN can be used to generate high-level abstract data representations which will be used for semantic indexing. These representations can reveal complex

associations and factors (especially when the raw input was Big Data), leading to knowledge and understanding about set of feature at particular pixel location. Data representations play an important role in the indexing of data, for example by allowing data points/feature points relatively similar representations to be stored closer to one another in memory, aiding in efficient information retrieval. This technique of grouping similar features is described details in equation 1 as mentioned earlier. It should be noted, however, that the high-level abstract data such as image pixels representations need to demonstrate relational and semantic association in order to actually confer a good classification and recognition for understanding and comprehension of the image object.

While Deep Learning aids in providing a semantic and relational understanding of the data, a vector contextual representations (extracted key points based on sift algorithm) of data instances would provide faster searching and information retrieval. Scale Invariant Feature Transform (SIFT) based points tend to be distinctive at major image area such as high peak points. More specifically, vector data representations contain semantic and relational information instead of just raw bit data, they can directly be used for semantic indexing when each data point (for example- a feature set vector) is presented by a vector representation, allowing for a vector-based comparison which is more efficient than comparing instances based directly on raw data.

Feature vector set =

<f1, f2, f3,….,fi,…., fn>

fi = [x1, x2, x3,…..,xm]

Where xi= D(xi) and

$$D(xi) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x \quad (2)$$

The data instances those have similar vector representations are likely to have same equilibrium semantic meaning. Thus, using vector representations as it is done in equation 1 and equation 2, of complex high-level data abstractions for indexing the data makes semantic indexing easy to understand. Semantic indexing has been done using hash function as described in equation 1. In the remainder of this section, the focus has been given on dataset design based on knowledge gained from feature extraction algorithm. However, the general idea of indexing based on data representations obtained from Deep Learning can be extended to other forms of data instances would provide faster searching and information retrieval.

1.2 Dataset Design

The details of feature points have been evaluated for edge point, corner point, highest distance based length and width.Those feature sets are stored in index (key-map) based data store.The traditional relational data bases are not appropriate for this architecture as it needs high volumes of data with faster acess and retrieval process.Henceforth NoSQL based databases with organized map based Bigdata architecture has been adapted.There are two kinds of datasets those are used in this work.

1. RAW dataset: its sizes ranges from 20 gb to 100 gb. The size is huge because it contains a lot of keypoints from feature data readings during training and testing of 5 feature points.
2. PROCESSED dataset: size is of few kb only.

The processed dataset is derived from the raw dataset and contains the data in the format suitable for statistical, mathematical, cognitive or AI analysis. It contains the data in the form features. There are approx 100 features and few hundred rows in the training dataset. The processed dataset completely resembles the raw dataset.

2 Methodology

2.1 Mask RCNN Architecture

Mask RCNN is a deep neural network aimed to solve instance segmentation problem in machine learning or computer vision. In other words, it can separate different objects in a image or a video. You give it a image, it gives you the object bounding boxes, classes and masks.There are two stages of Mask RCNN. First, it generates proposals about the regions where there might be an object based on the input image. Second, it predicts the class of the object, refines the bounding box and generates a mask in pixel level of the object based on the first stage proposal. Both stages are connected to the backbone structure.

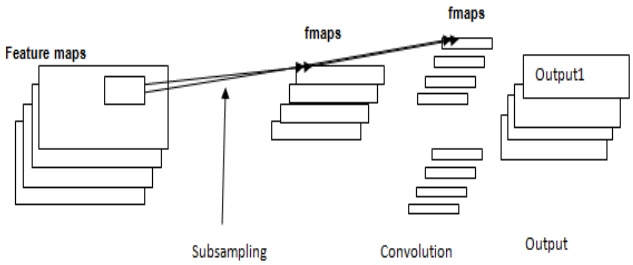

Figure 1: Diagram of Mask RCNN

Achievement of Mask R-CNN is it outperforms all existing, single-model entries on every task. Here in our experiment also in precision and accuracy Mask-RCNN based along with bigdata storage has played a vital role for classifying these solid objects.

3 Experiments and results

The convolution of images is executed using sift technique after applying noise removal filter. Here the median filter has been applied and then using OPENCV SIFT algorithm the feature points are extracted and processed in [224 X 224] matrix. The initial model has been taken custom solid objects and 2-D images. After initial training COCO dataset is used for MASK-RCNN in MATLAB. The comparison of performances parameters has been done in which accuracy and precision holds the significant value showing the proposed techniques.

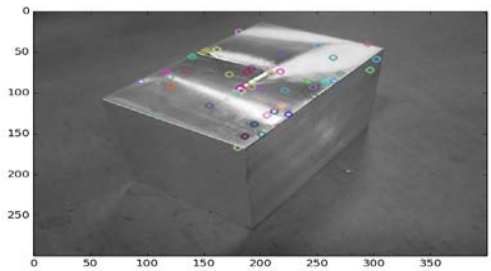

Figure 2: Feature points using sift algorithm

Figure 3: Solid block input images

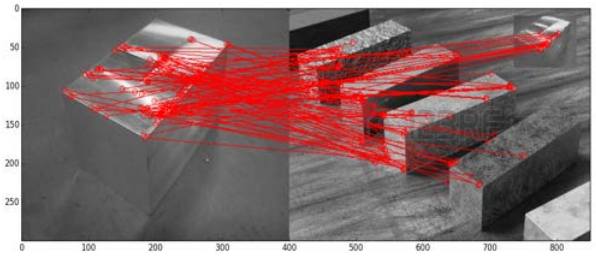

Figure 4: Diagram showing matching points of an identified object using Mask RCNN algorithm

Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN, running at 5 frame per second. Moreover, Mask R-CNN is easy to generalize to other tasks, e.g., allowing us to estimate feature detection in the same framework. Results in all object detection techniques using MRCNN along with the COCO datasets show various parameters for performance measures using instance segmentation, bounding-box object detection, and person keypoint detection.

	Training Data	AP[Val]	AP	AP50	Concrete block	Steel block	Marble block	Wooden Block
MRCNN	Fine	31.5	26.2	49.9	11.7	32.9	18.7	8.4
MRCNN	Fine Coco	36.4	32.0	58.1	14.6	30.21	16.65	9.3

Table 1: Comparison of dataset from COCO for mask RCNN algorithm for various solid blocks

3.1 Performance measures

In any classification model the parameters to be considered to evaluate correctness of the model are as mentioned below.

1) Accuracy: In the fields of science, engineering, industry, and statistics, the accuracy of a measurement system is the degree of proximity of measurements of a parameter to that quantity's actual (true) value. It shows how often the classification model is correct.

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{Total inNumber}}$$

2) Precision: In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the find. If the classification model predicts true precision means how often it happens.

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n.

3) Recall: Recall in information retrieval is the fraction of the documents that

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

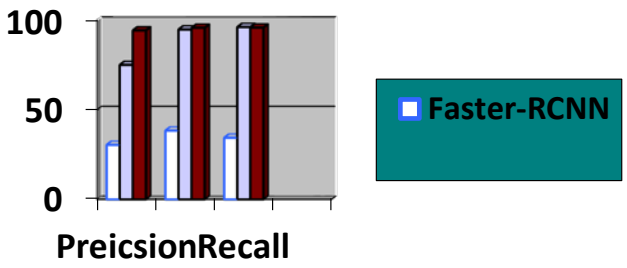

Figure 5: Graph diagram for comparison of results of 3 experiments of our proposed algorithm

4 Conclusion

MASK-RCNN design approach [5] along with big data has made significant contribution for object classification and detection with good precision and accuracy. The proposed detection framework on COCO database consists of two stages. The first stage of the pipeline applies a region proposal method, such as selective search and SIFT algorithm to extract regions of interest from an image and second stage is to feed them to a deep neural network for classification. This approach may be used in various intelligent and automation system for object identification with good accuracy.

REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, Neural Networks and signal processing, 2009 IEEE.

[2] Marcos Eduardo Valle, Complex-Valued Recurrent Correlation Neural Networks ,IEEE Transactions on Neural Networks and Learning SystemsVolume: 25, Issue: 9,Pages: 1600 - 1612,Year: 2014

[3] Liang Zhang, Peiyi Shen , Guangming Zhu , Wei Wei , and Houbing Song, A Fast Robot Identification and Mapping Algorithm Based on Kinect Sensor, Sensors 2015, 15, 19937-19967; doi:10.3390/s150819937

[4] Chen, C.; Liu, M.-Y.; Tuzel, C.O.; Xiao, J., R-CNN for Small Object Detection, TR2016-144, November 2016, Mitsubishi Electric Research Laboratories

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun,"Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 39, No. 6, June 2017

[6] Cheng Wang, Ying Wang, Yinhe Han, Lili Song, Zhenyu Quan, Jiajun Li and Xiaowei Li,"CNN-based object detection solutions for embedded heterogeneous multicore SoCs", 2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)

[7] Kaiming He Georgia Gkioxari Piotr Dollar Ross Girshick,"Mask R-CNN", Facebook AI Research (FAIR), April 2017.

[8] Subhransu Maji, Alexander C. Berg, and Jitendra Malik, “Efficient Classification for Additive Kernel SVMs”, Transactions On Pattern Analysis And Machine Intelligence, Vol. 39, No. 6, June 2017

[9] Cheng Wang, Ying Wang, Yinhe Han, Lili Song, Zhenyu Quan, Jiajun Li and Xiaowei Li,"CNN-based object detection solutions for embedded heterogeneous multicore SoCs", 2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)

[10] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald and Edin Muharemagic, Deep learning applications and challenges in big data analytics ,Journal of Big Data 2015

[11] Xiaojiang Peng, Cordelia Schmid, Multi-region two-stream R-CNN for action detection, European Conference on Computer Vision, Oct 2016, Amsterdam, Netherland

[12] Sapan Naik,Bankim Patel, Machine Vision based Fruit Classification and Grading -A Review, International Journal of Computer Applications (0975 –8887) Volume 170 –No.9, July 2017

[13] R Dalai, KK Senapati, Comparison of Various RCNN techniques for Classification of Object from Image , International Research Journal of Engineering and Technology (IRJET), Volume: 04, Issue: 07, July -2017

[14] T. Hoang Ngan Le, Yutong Zheng, Chenchen Zhu, Khoa Luu, Marios Savvides, Multiple Scale Faster-RCNN Approach to Driver's Cell-Phone Usage and Hands on Steering Wheel Detection ,IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016

[15] Sezer Karaoglu, Yang Liu, Theo Gevers, Detect2Rank: Combining Object Detectors Using Learning to Rank, IEEE Transactions on Image Processing, Year: 2016, Volume: 25, Issue: 1,Pages: 233 – 248

[16] Wanli Ouyang, Xingyu Zeng, Xiaogang Wang, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Hongyang Li, Kun Wang, Junjie Yan, Chen-Change Loy, Xiaoou Tang, "DeepID-Net: Object Detection with Deformable Part Based Convolutional Neural Networks",IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 39, Issue: 7,Pages: 1320 – 1334, Year: 2017

[17] Eel-Wan Lee, Soo-Ik Chae, Fast design of reduced-complexity nearest-neighbor classifiers using triangular inequality, IEEE Transactions on Pattern Analysis and Machine Intelligence Volume: 20, Issue: 5, Pages: 562 - 566, Year: 1998

[18] Ross Girshick, Fast R-CNN Object detection with Caffe , Microsoft Research.

[19] Web resource : https://en.wikipedia.org/wiki/Feature_hashing