Research Article

Performance Analysis of Object Detection Framework: Evolution from SIFT to Mask R - CNN

Ms. Yogitha. R [1], Dr. G. Mathivanan[2]
[1]Department of Computer Science and Engineering, [2]Department of Information Technology,
Sathyabama Institute of Science and Technology, Chennai – 600 119.yogitha.ravi1915@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: In a near of wide spread technological change that has givena positive impact to the society andhelpedinbuildingauser-friendly environment, object detection framework, an importantpart of Computer Vision (CV) plays a vital role. Starting fromasimpleautomaticattendancesystemforstudentsusingfacedetection, recognizing the presence of tumors in medical images,helping with automatic surveillance of cctv cameras to identifypeoplewhobreakstrafficrulescausingroadaccidentstobeingthecentral mechanism behind self-driving cars, object detection haswide range of applications and assist building an easy to cope withsmart environments. This in turn urges the need to evaluate theperformanceofthetechniquesbehindtheseframeworks.Thecentralideabehindthemodern-dayobjectdetectionandclassification is Convolutional Neural Network (CNN) which triesto mimic the occipital lobe, the visual cortex of the human brain.CNNhaswiderangeofvariationsandhascomethroughalongwaystarting from basic CV techniques like Scale Invariant FeatureTransform(SIFT),HistogramsofOrientedGradientsn(HOG)tillRegionbasedCNN's(R-CNN).Theperformanceofeachandeverymethodthathas led throughthe evolutionofobjectdetectionmethods, its advantages and the disadvantages which has pavedway for the innovation of next technique has been discussed andrepresentedindetail.

Keywords: Object Detection, Computer Vision, ConvolutionalNeuralNetwork,HistogramofOrientedGradients,R-CNN.

1. Introduction

Given an input image, Object detection technique involveslocalizing and identifying the artifacts present in the image andclassifying the artifacts into various categories. The whole ideabehind this process of object detection is to impose the humanbrain'saccuracyandspeedindetectingandrecognizingobjectsinto the machine using several machine learning techniques. Itall started in 1959 when Hubel and Wiesel [40] conducted theirresearchoncat's visual recognitionsystemby studying itsprimaryvisualcortexwhichhelpedinidentifyingandrecognizing the objects using the light reflections on the them.They studied the pattern in which the neurons in the visualcortexinthebrainreactedwithlightreflection atvariousanglesoftheobject.Theneuronswhichreactedwithsimpleexhibitoryand inhibitory signals to detect the lines in the object werenamed as simple neurons. In 1961, they further extended theirresearch into two parts, one dealing with neurons which helpsto process more complex level visual information's and theother dealing with binocular interaction by observing certainadditionalpatternsofinformation.Theirresearchonunderstandingtheprocessingofvisualinformationin animal's paved way to the computer vision technique SIFTdescriptor.

Inthebelowparagraphs,SectionIIdescribestheinnovationsinObjectDetectiontechniquesbeforeConvolutionalNeuralNetworkcameintoexistence. SectionIII describes the Object Detection framework that worksbasedonvariationsofdifferentConvolutionalNeuralNetworks. Both the section gives details on performanceanalysis on those techniques based on performance metricnamed mean Average Precision (mAP) which is the directmeasureofaccuracyoftheobjectdetectionframework.

2. Object Detection Techniques before CNN

SIFT

ScaleInvariantFeature Transform,[24][47] analgorithmtechniquethatinvolvesgeneratingfeaturevectorsbyconvoluting Gaussian filters with given sample input images.With the help of the generated feature vectors using SIFTfrom sample images, it is possible to detect the same objectsinthe images that has differentbackground, scaling androtated in divergent angles. It is also invariant in detectingand matching the features in various levels of brightness andcontrastsofthegiveninputimageandcanmatchfeaturesevenwhen the image suffers from occlusion. SIFT can also beusedtostitchtogetherthepanoramicimages.

2.1. HOG

Histogram of Oriented Gradients, feature extraction technique from images in computer vision. Given an image as input, HOG divides it into (8*8) pixel wise grid's, calculate the difference in pixel intensities and compute the gradient magnitude and direction for each grid. The gradient magnitude combined with gradient direction forms the feature vector. For the whole image, after calculating the collection of gradient magnitude and directions, feature vectors are calculated in the form of histogram consisting of n number of bars representing magnitudes equally divided from 00 – 1800based on the object taken into consideration. The histogram can be represented as n vectors or a matrix of size n*1 or as a pictorial representation with n lines witharrows pointing towards the corresponding magnitude and direction. In 1994, [25] HOG feature extraction was used torecognize hand gesture activities which was further extendedand applied to identify and recognize wide range of variety ofobjects ranging from cars, buses, bicycles, animals like dogs,catscowsandevenhuman[31][32]beings.

2.2. Object Detection with HOG and SVM

Support Vector Machine, [49] a classification algorithm, classifies the given set of data into two linearly separable groups with widest possible margins. Given a set of input data, SVM constructs a line equation with corresponding number of co efficient taken from the input data and classifies the data into two groups, each data point in the group represented by positive or negative sign. The signs represent the data belonging to different groups. The distance from each data point to the line represents the magnitude. Higher the magnitude of the resultant data point, higher is the confidence that it belongs to that particulargroup. SVM can be trained with set of inputsamples to generate equation of a line to classify the data, after training it can be tested with new set of data to check its performance in terms of accuracy.

The HOG feature obtained are given as input to the SVMclassifier. The feature vector and coefficients' obtained fromSVM aretaken,dotproductis computedandatlastbiastermisadded to get the final result. In 2005, [11]Dalal and Triggsdesigned an object detector using HOG and SVM classifier todetectthehumansfromthegiveninputimage.Thedrawbackinhereisthatthedetectorwasnot abletoclassifythepeoplewhowerenotinuprightposition.Toovercomethisdrawback,DeformablePartsmodeldetector[48][23]wasdesignedwhichhaddetectorsforindividualbodyparts.Foreg,consideringahumanbody, there were five detectors, one for detecting the face, twofor the left and right side of the body and two more for top andbottomportionsofthelegwhichinturn gaveverygood results.

3. Performance – CNN and its descendants

Hubel and Wiesel's idea also paved way to first set of neuralnetwork model for visual pattern recognition which was namedas Neocognitron. [41] It was based on unsupervised learningtechniques and the network was divided into two layers, firstlayercomposed of simple cells S-cell and the second layercomposedofcomplexcellsC-cells.Itisbasedonselforganization and was able to identify patterns even with littleshiftininthosepatternsifitwasrepeatedlygivenasinputtoit.It was improved further with [42]multi layer cascaded network,againbasedonunsupervisedlearningto learnandidentifyshifted input patterns with a new improvised algorithm whichgavebetterresults.

Theseneuralnetworkarchitectureswerefurtherextended,butthistimebasedonsupervisedlearningalgorithmcalled backcpropogation.[44][45] Givenasetofinput imageswithlabels,thenetworkfirstlearns

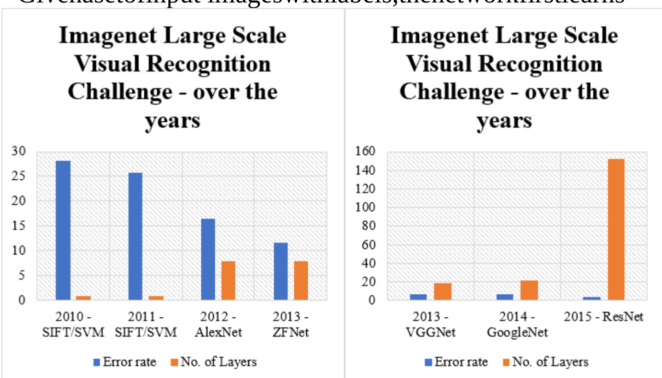

Figure 1. Winners of ILSVRC over the years, their error rate and the number of layers used. The graph gives an inference that with the increase in the number of layers the error rate has dropped down exponentially giving highest accuracy in classification.

theimageandgives aoutput which is the actual output. Now the difference (calculated interms of error) in betweenthe target and actual output iscalculated and backpropogated to the first set of layers todecrease the error and improve the accuracy. This furtherlead to Convolutional Neural Networks [43] which was solely usedforworkingwithimagesandrecognizingvisualpatterns.

Table 1.ILSVCR winners over the years, error percentage and authors name who designed it. [14]

Year	Architecture Name	AuthorName	Winner/Runner	Error in terms of mAP
2012	AlexNet[13]	AlexKrizhevsky, GeoffryHinton,LiyaSuskever	Winner	15.3 percent
2013	ZFNet[38]	MatthewZeiler,RobFergus	Winner	14.8 percent
2014	VGGNet[37]	Karen Simonyan,AndrewZisserman	Runner	8.0 percent
2014	GoogleNet [53]	Christian Szegedy,Wei Liu, YangqingJia, Pierre Sermanet,ScottReed,DragomirAnguelov, DumitruErhan, VincentVanhouckeand AndrewRabinovichich	Winner	6.67 percent
2015	ResNet[52]	Kaiming He, XiangyuZhang,Shaoqing,Ren andJianSun	Winner	3.6 percent
2016	Trimps -Soushen	Trimps researchinstitute,China	Winner	2.99 percent
2016	ResNeXt[2]	Saining Xie1, RossGirshick,PiotrDollar,ZhuowenTuand KaimingHe	Runner	3.03 percent
2017	SENet[57]	Jie Hu, Li Shen,Samuel Albanie, GangSunandEnhuaWu	Winner	2.251 percent

Imagenet[51] Large Scale Visual Recognition Challenge isvery famous object detection challenge, where each yearstartingfrom2010researchersmakesuseofthelargeimagenet database and classify the objects using differentcomputer vision techniques. Before imagenet database cameintoexistence,researchersweremakinguseof PASCALVOC[58]andCOCO[21]datasetwithannotatedimages.In 2010 and 2011 it started with classical CV techniques likeSIFT,HOG,SVMtodetectandclassifytheobjectswhich gaveclassification accuracy upto 70 percent. Gradually, by the year2012 AlextNet[13] which is a CNN based architecture won thechallenge with accuracy upto 86 percent which kick started theinterestinthisfieldofmachinelearning.Inthesubsequentyearsalmost allthewinningmodelswerebased onCNNandtheerrorratebecameincrediblylowereachyear.

3.1 CNN

Convolutional Neural Network, a class of neural networksinspired from biological working of visual cortex of humanbrain. Given an input image, CNN works by taking the inputimage as a matrix of pixel values, convolutes it with standardfilters of specific size to get n feature maps, then applies maxpooling to reduce the feature maps size into half, cascade thefeature maps with more filters and the final set of feature mapsare given to the fully connected layers [16] and classifier toclassify the objects in it. Rectified Linear unit can be used astransfer function since it performs well on linearly separable data [28][46].

3.1.1. Working

CNN filter/kernel is just a matrix of specific size usually 3*3. Feature Map is obtained by sliding and convoluting the filter over the input image of any size by maintaining the stride value as some constant and also by padding the margins of the input image so that after convolution, the output is obtained is same size as the input image. This feature map is then given to the pooling layer which max pools the feature maps into half its size. This convolution and pooling is done in n number of layers using n number of filters at each stage to get the final feature map which is then given to the Fully Connected layer in the form of feature vector for further classification. FC layer takes the feature vectors and convolves with different filters again to get another feature vector which is in turn again convolved with number of filters based on

Table 2. Different methodologies used for generating regions[12]. MS-.Multi-scale Saliency CC-Color Contrast ED- Edge Density SP- Super pixels Straddling

Paper Reference no	Methodologyused-explained	mAP
[39]	Objectness algorithm that combines MS + CC + ED + SP	25.4
[6]	Constrainedparametricmini–cutsusingbottomup process	30.7
[5]	Generaten–regionsaroundtheobject and rank them according tospecifity	31.6
[30]	Combinespixelsaccordingtovalues and hierarchically combinetogethertoformgroundtruth regionofobjects.	32.3
[17]	Usingobjectnessgeneratennumberofwindows in animageand categorize and choose the bestoneaccordingtoorderof magnitude.	30.4
[54]	Generatespartialspanningtreefrom similar pixels and tree withmaximumweightsisidentifiedas themainobjectintheimage	30.9
[4]	By resizing the window size to 8*8andbyusingbinarizednormalgradient,generateobjectr egionproposals.	22.4
[22]	Segmenttheimageusingimagepyramid,combinethevario usaligned hierarchical pairs of imageandgiveobjectproposalsasoutput.	32.7
[10]	Fromthesuper pixelstakenfromthe image, segment the objects bygrouping all the similar super pixelstogether.	31.3
[7]	Generatingboundingboxesfromtheedgespresentintheima ges	32.2

the dataset taken into consideration (in Pascal20, 20 filters are used as the dataset contains 20 different artifacts in the images). Final output of the FC layer is applied with SoftMax function to generate the confidence scores of each object in the dataset. Whichever object has the highest confidence score that will be the classification output of the corresponding image. This basic working of CNN has been explained diagrammatically in Figure 2 for given input image. ILSVCR, ImageNet challenge kindles the interest of researchers over CNN and led to a lot of innovative high performance CNN architectures. It started with AlexNet going through ResNet followed by many wide variations of ResNet-v2 [15] and so on, CNN based architecture gave a breakthrough in the field of object detection. The evolution of object detection frameworks each year has been explained in the Figure 1 and Table 1 where three different architectures are taken into consideration. The Table 3 explains the size of the input image and how it has been reduced after each feature map generation at each layer of the network. The main noticeable difference between the three isthe filter size taken in each layer in the network in convolution and max pooling layers. The shaded portion of the Table 3 denotes the width of the feature map obtained after pooling at each layer.

3.1.2. Bounding Box Regression

To localize the exact location of an object in the image, aboundingboxisdrawnaroundit.Thiscanbedoneinparallelin Fully Connected Layer where the coordinates of the box(x0,x1,y0,y1) is calculated by back propagating the errorsfound using L2 loss function. But it will be very difficult todetect and localize objects using sliding window of specificsame size when the image has multiple objects of varioussize. In 2014, OverFeat[26] networks overcame this problemby scaling the image insix differentscalesusing imagepyramid technique so that at each scale objects of differentsizes will fit fully inside the sliding window making it easiertodetectandlocalizetheobjects.

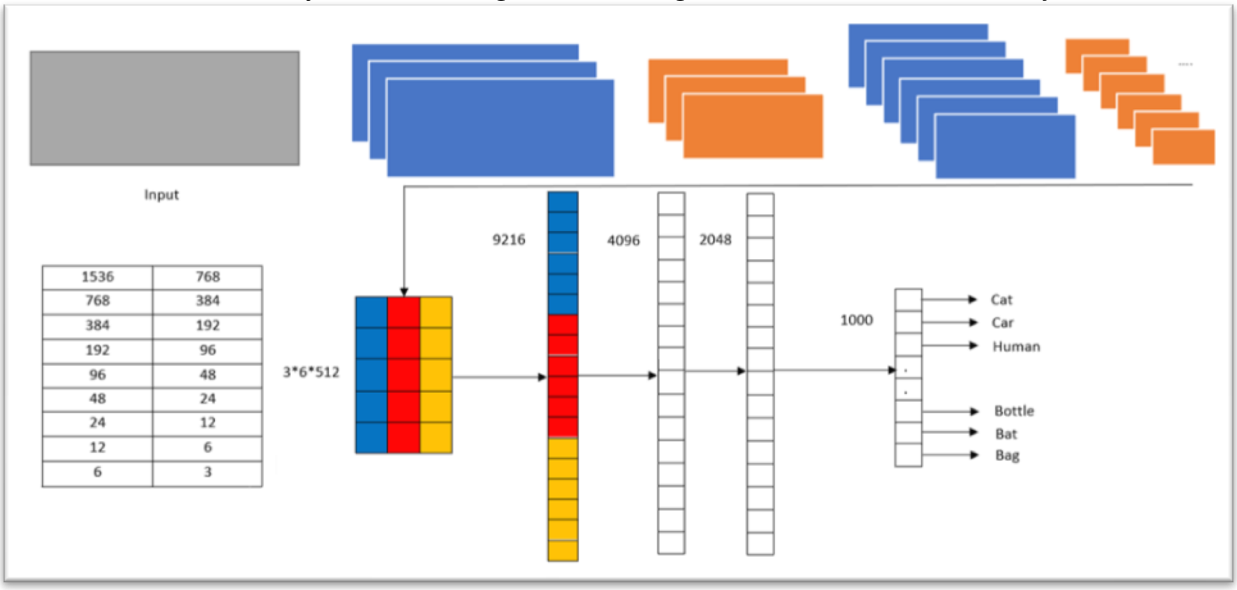

Figure 2. Input image of size 1536*768 is taken and convoluted with three filters to get the same sized output which is then max pooled to get image of size 768*384 which is again convoluted with 6*3 filters and max pooled to get image of size 384*192. This is repeated continuously and max pooled as shown in the table in the image till the image size becomes 6*3 which gives feature map of width 512 which is then taken as feature vector and given to Full Connected Layer to Classify it into 1000 classes of imagenet database

3.1.3. Region Proposal Methods:

Sliding the whole image along with it's background, wherethere are no chances of an object to be found and giving it asinputtothenetworkisgoingtoconsumealottimeunnecessarily. To overcome this disadvantage, the image canbe divided into regions and the region with presence of theobject can alone be given as the input to the network. Thereare a lot of techniques like multi scale saliency which worksbased on Fourier transform, colour contrast which segmentsobjects based on similar colour intensity, edge density whichboundsobjectsbasedontheedges,multi-thresholdingstraddling expansion[59], super pixel stradding which groupspixel with similar values together and some more techniquesalongwithitsmethodology,algorithmusedandmAPpercentageareexplainedindetailintheTable 2.

3.2 R-CNN

Given an input image, using selective search technique, atfirst nearly 2000 regions are generated from the image andthese regions are given as the input to the CNN architecture.Based onthepreferredCNNarchitecturethegeneratedregionboxes are cropped and warped to a specific size and is givenas the input to the first convolution and pooling layer. Insteadof softmax function, linearSVM is used as classifierandthere'snoneedforboundingboxregressionasboundingboxesarealreadygeneratedatthestartingstag eitself.

R CNN [27] [29] [33] network is nine times slower than the overfeat network because of the fact that it gives too many region proposals as the input to the network, but it is 10% more accurate when compared with others. It is also more accurate than the overfeat network because it eliminates all the background in the image through region proposals and doesn't result in any false positives whereas this is not the case with overfeat networks.

Bag of Visual words [55][56] using k-means clustering is amethodwhichclusterssamefeaturestogetherandhistogram gradient is drawn for each distinct features with the help ofcodebook generated from the features. Since the features areclustered according to the pixel intensity the location of theobject cannot be distinguished in here. No matter where theimage is present, it'll result in same histogram. To overcomethis, spatial pyramid technique [3] is used where feature mapsare generated in levels, at each level the image is divided intoparts and feature map is in turn generated for each part, thefeaturevectorobtainedatlastwillclearly distinguishtheobject and the locationof the corresponding object in theimage.

Thisspatialpyramid[34][35]poolingcanbeappliedtoCNNtogiveresultswithbetteraccuracy.InOverfeatnetworks, thelastpooling layer was replaced with spatial pyramid pooling. Theinput images need not be cropped or warped which decreasestheaccuracy.Itcanbefedasinputwithoutchangingtheaspectratio as last level pooling layer has been replaced by spatialpyramid pooling. This has increased the accuracy by 1.5 to 2percentonanaverage.

3.3 Fast R-CNN

Usingtheconceptofspatialpyramidpooling,SPP2stagenetworkwasconstructedwheretheinputimagewasdirectlygivenasinputtotheconvolutionallayersinsteadofgiving2000regionsasinRCNN.Regionproposalsarealsogenerated and they are translated into feature maps using ROI(RegionOfInterest)projection.AftergeneratingtheROI,thatpartaloneispooledusingvariouslevelsofSpatial PyramidPooling.Thenatlaststage,BoundingboxesaregeneratedusingL2loss.SPPNettakes0.3secstoprocesstheinitialinputimagewhereasRCNNtakes9secstoprocessthesamewhichisahugeadvantageousdifference.Intermsof accuracy,RCNN gives 58.5% whereas SPPNet gives 59.2% respectively.Fast R CNN is just an extension of SPPNet. It just makes fewdifferencesinSPPNetarchitecture.Given,aninputimage,FastRCNNgivesittofirstlevelofconvolutionallayers directlyandalsogeneratedregionproposalsoftheimageseparately.Then,itgeneratesfeaturemapsusingROIprojectionandgivesittoROIpoolinglayer,whichisalsolikeSPPpoolinglayer with the only difference that it has only one level of 7*7pooling altogether. And then feature vectors are given to FC.Here, finetuninghappensusingloglossandsoftmaxisusedforclassification instead of SVM used bySPPNet. For BoundingBoxgeneration,smoothL1loss functionis used.

Table 3. Comparison between CNN Architectures from the size of filters and feature maps at each layer and feature vector obtained from the lastconv+maxpoollayertotheconfidencescoresobtainedfromthelastlayerofthenetwork

Architecture	**Input**	**Conv+MaxPoolLayers-FeatureMapGeneration**										**FullyConnectedLayer – FeatureVector**		**Output**
		Layer1		Layer2		Layer3		Layer4		Layer5				
Name	Size	Filtersize	FeatureMap	FeatureMap	FeatureMap	Filtersize	FeatureMap	Filtersize	FeatureMap	Filtersize	FeatureMap	Layer6	Layer7	FinalConfidence Scores afterapplyingSoftMax
AlexNet	224 * 224	11*11	55 *55 96	5*5	27 *27 256	5*5	13 *13 384	3*3	13 *13 384	3*3	13 *13 256	4096	4096	1000
ZFNet	224 * 224	7*7	55 *55 96	5*5	27 *27 256	3*3	13 *13 384	3*3	13 *13 384	3*3	13 *13 256	4096	4096	1000
VGGNet	224 * 224	3*3	55 *55 96	3*3	27 *27 256	3*3	13 *13 384	3*3	13 *13 384	3*3	13 *13 256	4096	4096	1000

Altogether,FastRCNN isonestepprocesswhereitadds thelossobtainedbyclassifierandboundingboxregressor,back propagatesthroughthenetworkup tothirdconvolutionallayerandfinetunethefeatures.Whereas,othernetworksfollow2-stepor3-stepprocessbecauseeacherrorobtainedbyclassifier, SVM and Bounding Box regressor are backpropagatedseparatelyandfeaturesarefinetuned.Duetothe fact that Fast R CNN follows one step process it is 149timesfasterthantheR-CNNand22timesfasterthantheSPPNetarchitecture.Itgives anaccuracyofabout66.9%.

3.4 Faster R-CNN

The region proposals can be obtained using various region proposal networks or by using dense sampling techniques which has been used by the overfeat networks. Now, is it possible to use dense sampling methods to come up with region proposals instead of classical computer vision techniques like selective search or edge box. The minimum criteria to replace the existing region proposal networks and the combinations that have been tried based on those criteria's are shown in the Table 4. The third technique from the Table 5 Fast R CNN + Neural Network is the central idea behind Faster R CNN [9] since it satisfies all the corresponding criterias. The network part of the Fast R CNN is retained as such and the usual Selective Search technique that gives 2000 region proposals is replaced by a region proposal network which has fully connected layer with two parts, one for classifying between foreground and background parts and giving confidence scores for each using Softmax classifierand the other one for bounding box regression which has 9differentregressorsforthreebasicshapesofwindow(squarebox, height wise rectangular box and width wise rectangularbox) with 9 different aspect ratios that will fit all the objectsintheimage.

Faster R CNN gives mAP of 69.9% with only 300 region proposals whereas the previous Fast R CNN with Selective Search which gives 2000 region proposals gives mAP rate of 66.9%. The results clearly show that Faster R CNN gives better accuracy with less computation time compared with the previous existing techniques.

3.5 Mask R-CNN

Mask R – CNN [1] is an extension ofFasterR–CNNwhichfurtheraddsinstance levelsegmentationtothe network.While in Faster R – CNN, output is a bounding box andclassification of the objects in the image, Mask R – CNNgives three classes of output, the two being already said, thethird one is mask generated around the different instances ofthe object. To generate mask, [18][19][20] fully connected layeris added separately and ROI is given as the input to it. Maskwillbegeneratedbycomputingpixeltopixelcalculation.

SinceoneimagemighthaveseveralROIforasingleobject,beforegivingitasinputtothemaskpartofthenetwork,ROI is re calculated by comparing it with ground truth box andfinding IoU (Intersection over Union), if the value is above0.5, it is considered as ROI or else the corresponding box isdiscarded from ROI. Also, here instead of ROI pool, a newtechnique ROI align is used. ROI pool uses SPP and maxpools the features but it might have minor differences whileprojectingthefeatureswhichmightnotaffecttheclassification but these differences makes a major impact inpixeltopixelcomputationwhilegeneratingmask.ToovercomethisROIalignisusedwherethefeaturemapsare aligned not based on pixel grid division, but by using bilinearinterpolation and dividing the pixel in the feature maps intoexactfloatingpointnumbersandaligningwhichwillhelptopinpointtheexactfeaturestherebygeneratingthepixelwisemaskseasily. Mask R – CNN gives better accurate results and alsohelpsatinstancelevelsegmentation.

Table 4. Criteria’s to be satisfied by proposed technique to replaceSelectiveSearchTechnique

(i)	Shouldbeableto propose<2000regionproposals.
(ii)	Shouldbe fasterthanSelectiveSearch
(iii)	Shouldbeaccurateor betterthantheSelectiveSearch
(iv)	Shouldbeabletopropose • OverlappingROI’s • Withdifferentaspectratio’s • Withdifferentscales

Table 5. Analysis of Combination of different dense sampling techniques with Fast R CNN which gives best result.

NetworkCombination	Advantage/Disadvantage	Satisfies theCriteria?
FastRCNN+ SlidingWindow + ImagePyramid	ImagePyramidtechniqueisfourtimesslower	Doesn’tsatisfy(ii)
Fast R CNN+FeaturePyramid	Generates 9 different regionproposals for each FeatureMap.For a standard FM ofsize 40*60 gives40*60*9=20000proposalsto ROIpoolinglayer.	Doesn’tsatisfy(i)
FastRCNN+NeuralNetwork	Gives 300 to 500 regionproposals using SlidingWindow/Dense Samplingtechniques	Satisfies (i),(ii),(iii),(iv)

4. Conclusion

Different ObjectDetectionframework hasbeenanalyzedalongwithitsadvantageanddisadvantage.Howthedisadvantage of each technique has been overcome with eachnew technique all the way along has been discussed. The state-of-artperformanceofeachbreakthroughframeworkduringeveryyearofILSVRCalongwithitstop5errorintermsofm APhas been tabulated and the performance of each network hasbeen analyzed. The techniques and networks considered aboveareadropintheoceanwhencomparedtoeveryexistingtechnique that has paved way to the innovation in the field ofComputer Vision. Though the error rate starting from 28-30percent inclassical CV techniques hasbeenincredibly andexponentially reduced through the years, mainly contributed by theConvolutionalNeuralNetworkfamily,itstillhasalongway to go to match the accuracy and speed of the human visualcortex.

References

1. Kaiming He, Georgia Gkioxari, Piotr Doll´ar and Ross Girshick, Mask R-CNN *Computer Vision and Pattern*
2. *Recognition*, 2018.

3. Saining Xie1, Ross Girshick, Piotr Dollar, Zhuowen Tu and Kaiming He Aggregated Residual Transformations
4. for Deep Neural Networks *Computer Vision and Pattern Recognition*, 2017.

5. Svetlana Lazebnik1, Cordelia Schimid and Jean Ponce Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories *Computer Vision and Pattern Recognition* 2016.
6. Ming-Ming Cheng1, Yun Liu1, Wen-Yan Lin, Ziming Zhang, PaulL. Rosin and Philip H. S. Torr BING: Binarized normed gradients for objectness estimation at 300fps*Computer Visual Media*, 2019.
7. Ian Endres and Derek Hoiem Category Independent Object Proposals*Computer Vision European Conference on Computer Vision* 2010.
8. Joao Carreira and Cristian Sminchisescu Constrained Parametric Min-Cuts for Automatic Object Segmentation *Computer Vision and Pattern Recognition*, 2010.
9. C. Lawrence Zitnick and Piotr Doll´ar Edge Boxes: Locating Object Proposals from Edges *Computer Vision European Conference on Computer Vision* 2014.
10. Ross Girshick Fast R-CNN *Computer Vision and Pattern Recognition*, 2015.
11. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun Faster R- CNN: Towards Real-Time Object Detection with Region Proposal Networks *Pattern Analysis And Machine Intelligence*, 2017.
12. PekkaRantalankila, Juho Kannala and EsaRahtu Generating object segmentation proposals using global and local search Computer *Vision and Pattern Recognition* 2014.
13. Navneet Dalal and Bill Triggs Histograms of Oriented Gradients for Human *Detection Computer Vision and Pattern Recognition*, 2005.
14. Jan Hosang, Rodrigo Benenson and Bernt Schiele How good are detection proposals, really? *Computer Vision and Pattern Recognition*, 2014.
15. Alex Krizhevsky, Ilya Sutskever and Geoffrey E.Hinton ImageNet Classification with Deep Convolutional Neural Networks *Communications of the ACM*, 2017.
16. Rajat Vikram Singh ImageNet Winning CNN Architectures – A Review *semanticscholar*, 2017.
17. Christian Szegedy, SeregyIoeffy and Vincent Vanhoucke Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning *Computer Vision and Pattern Recognition*, 2016.
18. Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren and Jian Sun Instance- sensitive Fully Convolutional Networks *Computer Vision and Pattern Recognition*, 2016.
19. EsaRahtu, Juho Kannala and Mathew Blaschko Learning a Category Independent Object Detection Cascade *International Conference on Computer Vision* 2011.
20. Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert and Piotr Dollar Learning to Refine Object Segments *Computer Vision and Pattern Recognition*, 2016.
21. Tianshui Chen, Liang Lin, Xian Wu, Nong Xiao, and Xiaonan Luo Learning to Segment Object Candidates via Recursive Neural Networks *Computer Vision and Pattern Recognition*, 2016.
22. PedroO.Pinheiro, RonanCollobert and PiotrDoll´ar Learning to Segment Object Candidates *Computer Vision and Pattern Recognition*, 2015.
23. Tsung-Yi Lin, Michael Maire, Serge Belongie, LubomirBourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick and Piotr Doll´ ar Microsoft COCO: Common Objects in Context In*Computer Vision and Pattern Recognition*, 2015.

24. Jordi Pont-Tuset, Pablo Arbel´aez, Jonathan T. Barron and Ferran Marques Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation*Computer Vision and Pattern Recognition*, 2015.
25. Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan Object Detection with Discriminatively Trained Part Based Models*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.
26. David G. Lowe Object Recognition from Local Scale-Invariant Features *International Conference on Computer Vision*, 2002.
27. William T. Freeman, Michal Roth Orientation Histograms for Hand Gesture Recognition *IEEE International Workshop on Automatic Face and Gesture Recognition*, 1995.
28. Pierre Sermanet David Eigen Xiang Zhang Michael Mathieu Rob Fergus Yann LeCunOverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks *Computer Vision and Pattern Recognition*, 2014.
29. Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik Rich feature hierarchies for accurate object detection and semantic segmentation*Computer Vision and Pattern Recognition*, 2014.
30. Vinod Nair and Geoffrey E. Hinton Rectified Linear Units Improve Restricted Boltzmann Machines In*International Conference on Machine Learning*, 2010.
31. Jifeng Dai, Yi Li, Kaiming He and Jian Sun R- FCN: Object Detection via Region-based Fully Convolutional Networks *International Conference on Neural Information Proceeding Systems*, 2016
32. Koen E. A., van de Sande, Jasper R. R., Uijlings† Theo Gevers and Arnold W. M. Smeulders Segmentation as Selective Search for Object Recognition *International Conference on Computer Vision*, 2011.
33. Takuya Kobayashi, Akinori Hidaka, and Takio Kurita Selection of Histograms of Oriented Gradients Features for Pedestrian Detection *International Conference on Neural Information Processing*, 2007.
34. Qiang Zhu, Shai Avidan, Mei-Chen Yeh and Kwang-Ting Cheng Fast Human Detection Using a Cascade of Histogram of Oriented Gradients *Computer Vision and Pattern Recognition*, 2006.
35. Bharath Hariharan, Pablo Arbel´aez1, Ross Girshick and Jitendra Malik Simultaneous Detection and Segmentation*Computer Vision and Pattern Recognition*, 2014.
36. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition *Computer Vision and Pattern Recognition*, 2015.
37. Max Jaderberg, Karen Simonyan, Andrew Zisserman and KorayKavukcuoglu Spatial Transformer Networks In*Computer Vision and Pattern Recognition*, 2016.
38. Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama and Kevin Murphy Speed/accuracy trade-offs form modern convolutional object detectors*Computer Vision and Pattern Recognition*, 2017.
39. Karen Simonyan and Andrew Zisserman Very Deep Convolutional Networks For Large-Scale Image Recognition *Computer Vision and Pattern Recognition*, 2015.
40. Matthew D. Zeiler and Rob Fergus Visualizing and Understanding Convolutional Networks *Computer Vision and Pattern Recognition*, 2013.
41. Bogdan Alexe, Thomas Deselaers and Vittorio Ferrari What is an object? *Computer Vision and Pattern Recognition* 2010.
42. D. H. Hubel and T. N. Wiesel, Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex*The Journal of Physiology*, 1962.
43. Kunihiko Fukushima Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position*Biological Cybernetics*, 1980.
44. Kunihiko Fukushima and Se1 Miyake Neocognitron: A New Algorithm For Pattern Recognition Tolerant Of Deformations And Shifts In Position*Pattern Recognition*, 1981.
45. Yann Le Cun, Leon Bottou, YoshuaBengio and Patrick Haner Gradient Based Learning Applied to Document Recognition*Proceedings of the IEEE*, 1998.
46. Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel Handwritten Digit Recognition with a Back- Propagation Network*Advances in Neural Information Processing Systems*, 1989.
47. Yann Le Cunn Generalization and Network Design Strategies *Connectionism in Perspective*, 1989.
48. Yogitha. R, Mathivanan. G, Performance Analysis of Transfer Funtions in an Artificial Neural Networks In In *ICCSP*, 2018.
49. Lowe, David G. Distinctive Image Features from Scale-Invariant Keypoints*International Journal of Computer Vision*, 2004.
50. Pedro F. Felzenszwalb and Daniel P. Huttenlocher Pictorial Structures for Object Recognition *International Journal of Computer Vision*, 2005.
51. Corrina Cortes and Vladimir Vapnik Support Vector Networks *Machine Language*, 1995.

52. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei- Fei ImageNet: A large-scale hierarchical image database *Computer Vision and Pattern Recognition*, 2009.
53. Kaiming He, Xiangyu Zhang, Shaoqing, Ren and Jian Sun Deep Residual Learning for Image Recognition*Computer Vision and Pattern Recognition*, 2015.
54. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich Going Deeper with Convolutions *Computer vision and Pattern Recognition*, 2014.
55. Santiago Manen, Matthieu Guillaumin and Luc Van Gool Prime Object Proposals with Randomized Prim's Algorithm*Computer Vision*, 2013.
56. David Aldavert, MarcalRusinol, Ricardo Toledo and JosepLlados A Study of Bag-of-Visual-Words Representations for Handwritten Keyword Spotting *Computer Vision*, 2013.
57. Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng and S. Avidan Fast Human Detection Using a Cascade of Histograms of Oriented Gradients *Computer Vision and Pattern Recognition*, 2006.
58. Jie Hu, Li Shen, SamuealAlbanie, Gang Sun, Enhua Wu Squeeze- and-Excitation Networks *Computer Vision and Pattern Recognition*, 2019.
59. Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn and Andrew Zisserman The PASCAL Visual Object Classes (VOC) Challenge *International Journal of Computer Vision*, 2010.
60. Xiaozhi Chen, Huimin Ma, Xiang Wang and Zhichen Zhao Improving Object Proposals with Multi-Thresholding Straddling Expansion *Computer Vision and Pattern Recognition*, 2015.
61. Nagarajan, G., Minu, R. I., Muthukumar, B., Vedanarayanan, V., & Sundarsingh, S. D. (2016). Hybrid genetic algorithm for medical image feature extraction and selection. *Procedia Computer Science*, *85*, 455-462.
62. Minu, R. I., G. Nagarajan, A. Suresh, and A. Jayanthila Devi. "Cognitive computational semantic for high resolution image interpretation using artificial neural network." (2016).
63. Nagarajan, G., Minu, R. I., & Devi, A. J. (2020). Optimal nonparametric bayesian model-based multimodal BoVW creation using multilayer pLSA. *Circuits, Systems, and Signal Processing*, *39*(2), 1123-1132.
64. Dhanalakshmi, A., and G. Nagarajan. "Convolutional Neural Network-based deblocking filter for SHVC in H. 265." Signal, Image and Video Processing 14 (2020): 1635-1645.