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

Ms. Yogitha. R 1, Dr. G. Mathivanan 2
1Department of Computer Science and Engineering. 2Department 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 given a positive impact to the society and helped in building a user-friendly environment, object detection framework, an important part of Computer Vision (CV) plays a vital role. Starting from a simple automatic attendance system for students using face detection, recognizing the presence of tumors in medical images, helping with automatic surveillance of CCTV cameras to identify people breaking traffic rules causing road accidents to being the central mechanism behind self-driving cars, object detection has wide range of applications and assist building an easy to cope with smart environments. This in turn urges the need to evaluate the performance of the techniques behind these frameworks. The central idea behind them modern-day object detection and classification is Convolutional Neural Network (CNN) which tries to mimic the occipital lobe, the visual cortex of the human brain. CNN has a widerange of variations and has come through a long way starting from basic CV techniques like Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) till Region-based CNN’s (R-CNN). The performance of each and every method that has led through the evolution of object detection methods, its advantages and the disadvantages which has paved way for the innovation of next technique has been discussed and represented in detail.

Keywords: Object Detection, Computer Vision, Convolutional Neural Network, Histogram of Oriented Gradients, R-CNN.

1. Introduction

Given an input image, object detection technique involves localizing and identifying the artifacts present in the image and classifying the artifacts into various categories. The whole idea behind this process of object detection is to impose the human brain’s accuracy and speed in detecting and recognizing objects into the machine using several machine learning techniques. It all started in 1959 when Hubel and Wiesel [40] conducted their research on cat’s visual recognition system by studying its primary visual cortex which helped in identifying and recognizing the objects using the light reflections on them. They studied the pattern in which the neurons in the visual cortex in the brain react to light reflection at various angles of the object. The neurons which react with simple excitatory and inhibitory signals to detect the lines in the object were named as simple neurons. In 1961, they further extended their research into two parts, one dealing with neurons which help to process more complex level visual information and the other dealing with binocular interaction by observing certain additional patterns of information. Their research on understanding the processing of visual information in animal’s paved way to the computer vision technique SIFT descriptor. In the below paragraphs, Section II describes the innovations in object detection techniques before Convolutional Neural Network came into existence. Section III describes the Object Detection framework that works based on variations of different Convolutional Neural Networks. Both the section gives details on performance analysis on those techniques based on performance metrics named mean Average Precision (mAP) which is the direct measure of accuracy of the object detection framework.

2. Object Detection Techniques before CNN

SIFT

Scale Invariant Feature Transform [24][47] algorithm technique that involves generating feature vectors by convoluting Gaussian filters with given sample input images. With the help of the generated feature vectors using SIFT from sample images, it is possible to detect the same object in the image that has different background, scaling and rotated in
Research Article

divergent angles. It is also invariant in detecting and matching the features in various levels of brightness and contrast of the given input image and can match features even when the image suffers from occlusion. SIFT can also be used to stitch together the panoramic images.

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 – 1800 based on the object taken into consideration. The histogram can be represented as n vectors or a matrix of size nx1 or as a pictorial representation with n lines with arrows pointing towards the corresponding magnitude and direction. In 1994, [25] HOG feature extraction was used to recognize hand gesture activities which was further extended and applied to identify and recognize wide range of variety of objects ranging from cars, buses, bicycles, animals like dogs, cats, cows and even humans [31][32].

2.2. 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 particular group. SVM can be trained with set of input samples 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 SVM classifier. The feature vector and coefficients’ obtained from SVM are taken, dot products computed and at last bias term added to get the final result. In 2005, [31] Dalal and Triggs designed an object detector using HOG and SVM classifier to detect the humans from the given input image. The drawback here is that the detector was not able to classify the people when not in an upright position. To overcome this drawback, Deformable Parts model detector [48][23] was designed which had detectors for individual body parts. For example, considering a human body, there were five detectors, one for detecting the face, two for the left and right side of the body and two more for top and bottom portions of the legs which in turn gave very good results.

3. Performance – CNN and its descendants

Hubel and Wiesel’s idea also paved way to first set of neural network model for visual pattern recognition which was named Neocognitron. [41] It was based on unsupervised learning techniques and the network was divided into two layers, first layer composed of simple cells S-cell and the second layer composed of complex cells C-cells. It is based on self organization and was able to identify patterns even with little shift in the features. It was repeated by given data in put into it. It was improved further with multi layer cascaded network, again based on unsupervised learning to learn and identify shifted input patterns with a new improvised algorithm which gave better results. These neural network architectures were further extended, but this time based on supervised learning algorithm called backpropagation [44][45]. Given a set of input images with labels, the network first learns...

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.
error and improve the accuracy. This further lead to Convolutional Neural Networks \cite{43} which was solely used for working with images and recognizing visual patterns.

Table 1. ILSVRC winners over the years, error percentage and authors name who designed it. \cite{14}

<table>
<thead>
<tr>
<th>Year</th>
<th>Architecture Name</th>
<th>Author Name</th>
<th>Winner/Runner</th>
<th>Error in terms of mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>AlexNet \cite{13}</td>
<td>Alex Krizhevsky, Geoffrey Hinton, Liya Susskever</td>
<td>Winner</td>
<td>15.3 percent</td>
</tr>
<tr>
<td>2013</td>
<td>ZFNet \cite{38}</td>
<td>Matthew Zeiler, Rob Fergus</td>
<td>Winner</td>
<td>14.8 percent</td>
</tr>
<tr>
<td>2014</td>
<td>VGGNet \cite{37}</td>
<td>Karen Simonyan, Andrew Zisserman</td>
<td>Runner</td>
<td>8.0 percent</td>
</tr>
<tr>
<td>2014</td>
<td>GoogleNet \cite{53}</td>
<td>Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovichich</td>
<td>Winner</td>
<td>6.67 percent</td>
</tr>
<tr>
<td>2015</td>
<td>ResNet \cite{52}</td>
<td>Kaiming He, Xiangyu Zhang, Shaoting Zhang, Ren and Jian Sun</td>
<td>Winner</td>
<td>3.6 percent</td>
</tr>
<tr>
<td>2016</td>
<td>Trimpis - Soushen</td>
<td>Trimps research institute, China</td>
<td>Winner</td>
<td>2.99 percent</td>
</tr>
<tr>
<td>2016</td>
<td>ResNeXt \cite{2}</td>
<td>Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He</td>
<td>Runner</td>
<td>3.03 percent</td>
</tr>
<tr>
<td>2017</td>
<td>SENet \cite{57}</td>
<td>Jie Hu, Li Shen, Samuel Albanie, Gang Sun and Enhua Wu</td>
<td>Winner</td>
<td>2.251 percent</td>
</tr>
</tbody>
</table>

Imagenet\cite{51} Large Scale Visual Recognition Challenge is very famous object detection challenge, where each year starting from 2010 researchers make use of the large imagenet database and classify the objects using different computer vision techniques. Before imagenet database came into existence, researchers were making use of PASCAL VOC\cite{58} and COCO\cite{21} dataset with annotated images. In 2010 and 2011 it started with classical CV techniques like SIFT, HOG, SVM to detect and classify the objects which gave classification accuracy up to 70 percent. Gradually, by the year 2012 AlexNet\cite{13} which is a CNN based architecture won the challenge with accuracy up to 86 percent which kick started the interest in this field of machine learning. In the subsequent years almost all the winning models were based on CNN and the error rate became incredibly low each year.

3.1 CNN

Convolutional Neural Network, a class of neural networks inspired from biological working of visual cortex of humanbrain. Given an input image, CNN works by taking the input image as a matrix of pixel values, convolutes it with standard filters of specific size to get n feature maps, then applies maxpooling to reduce the feature maps size into half, cascade the feature maps with more filters and the final set of feature maps are given to the fully connected layers \cite{16} and classifier to classify the objects in it. Rectified Linear unit can be used as a transfer function since it performs well on linearly separable data \cite{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

<table>
<thead>
<tr>
<th>Paper Reference no</th>
<th>Methodology used-explained</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>[39]</td>
<td>Objectness algorithm that combines MS + CC + ED + SP</td>
<td>25.4</td>
</tr>
<tr>
<td>[5]</td>
<td>Generate regions around the object and rank them according to tospecificity</td>
<td>31.6</td>
</tr>
<tr>
<td>[30]</td>
<td>Combines pixels according to values and hierarchically combines to get the ground truth region of objects.</td>
<td>32.3</td>
</tr>
<tr>
<td>[17]</td>
<td>Using objectness generates number of windows in an image and categorize and choose the best one according to tospectrum.</td>
<td>30.4</td>
</tr>
<tr>
<td>[54]</td>
<td>Generate partial spanning tree from similar pixels and tree with maximum weights is identified as the main object in the image.</td>
<td>30.9</td>
</tr>
<tr>
<td>[4]</td>
<td>By resizing the window size to 8*8 and by using binarized normal gradient, generate object region proposals.</td>
<td>22.4</td>
</tr>
<tr>
<td>[22]</td>
<td>Segment the image using image pyramid, combine the vario us aligned hierarchical pairs of image and give object proposal as output.</td>
<td>32.7</td>
</tr>
<tr>
<td>[10]</td>
<td>From the super pixels taken from the image, segment the objects by grouping all the similar super pixels together.</td>
<td>31.3</td>
</tr>
<tr>
<td>[7]</td>
<td>Generate bounding boxes from the edges present in the image.</td>
<td>32.2</td>
</tr>
</tbody>
</table>

...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. ILSVRC, 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 is the 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, a bounding box is drawn around it. This can be done in parallel using Fully Connected Layer where the coordinates of the box(x0,x1,y0,y1) is calculated by back propagating the error found using L2 loss function. But it will be very difficult to detect and localize objects using sliding window of specific same size when the...
image has multiple objects of various size. In 2014, OverFeat\textsuperscript{26} networks overcame this problem by scaling the image in six different scales using image pyramid technique so that at each scale objects of different sizes will fit fully inside the sliding window making it easier to detect and localize the objects.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image}
\caption{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.}
\end{figure}

3.1.3. Region Proposal Methods:
Sliding the whole image along with its background, where there are no chances of an object to be found and giving it as input to the network is going to consume a lot of time unnecessarily. To overcome this disadvantage, the image can be divided into regions and the region with the presence of the object can be given as the input to the network. There are a lot of techniques like multi scale saliency which works based on Fourier transform, colour contrast which segments objects based on similar colour intensity, edge density which bounds objects based on the edges, multi-thresholding straddling expansion\textsuperscript{59}, super pixel straddling which groups pixels with similar values together and some more techniques along with their methodology, algorithm used and mAP percentage are explained in detail in the Table 2.

3.2 R-CNN

Given an input image, using selective search technique, at first nearly 2000 regions are generated from the image and these regions are given as the input to the CNN architecture. Based on the preferred CNN architecture, generated region boxes are cropped and warped to a specific size and is given as the input to the first convolution and pooling layer. Instead of softmax function, linear SVM is used as classifier and there is no need for bounding box regression as bounding boxes are already generated at the starting stage itself.

R-CNN\textsuperscript{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\textsuperscript{55}[56] using k-means clustering is a method which clusters same features together and histogram gradient is drawn for each distinct features with the help of codebook generated from the features. Since the features are clustered according to the pixel intensity the location of the object cannot be distinguished in here. No matter where the image is present, it’ll result in same histogram. To overcome this, spatial pyramid technique\textsuperscript{3} is used where feature maps are generated in levels, at each level the image is divided into parts and feature map is in turn generated for each part, the feature vector obtained at last will clearly distinguish the object and the location of the corresponding object in the image.

This spatial pyramid\textsuperscript{34}[35] pooling can be applied to CNN to give results with better accuracy. In OverFeat networks, the last pooling layer was replaced with spatial pyramid pooling. The input images need not be cropped or
warped which decreases the accuracy. It can be fed as input without changing the aspect ratio as the last level pooling layer has been replaced by spatial pyramid pooling. This has increased the accuracy by 1.5 to 2 percentage on average.

3.3 Fast R-CNN
Using the concept of spatial pyramid pooling, SPP2 stage network was constructed where the input image was directly given into the convolutional layers instead of giving 2000 regions in RCNN. Region proposals are also generated and they are translated into feature maps using ROI (Region Of Interest) projection. After generating the ROI, that part alone is pooled using various levels of spatial pyramid pooling. Then at last stage, bounding boxes are generated using L2 loss. SPPNet takes 0.3 second to process the initial input image where as RCNN takes 9 seconds to process the same which gives a huge advantage difference. In terms of accuracy, RCNN gives 58.5% whereas SPPNet gives 59.2% respectively. Fast R-CNN is just an extension of SPPNet. It just makes few differences in the SPPNet architecture. Given an input image, Fast R-CNN gives it to the first level of convolutional layers directly and also generates region proposals separately. Then, it generates feature maps using ROI projection and gives it to the ROI pooling layer, which is also like SPP pooling layer with the only difference that it has only one level of 7*7 pooling altogether. And then feature vectors are given to FC. Here, fine-tuning happens using log loss and softmax is used for classification instead of SVM used by SPPNet. For bounding box generation, smooth L1 loss function is used.

Table 3. Comparison between CNN Architectures from the size of filters and feature maps at each layer and feature vector obtained from the last conv+ max pool layer to the confidences scores obtained from the last layer of the network

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Input Size</th>
<th>Conv+MaxPool Layers - FeatureMaps Generation</th>
<th>Fully Connected Layer - FeatureVector</th>
<th>Output Layer</th>
<th>Final Confidence Scores after applying SoftMax</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>224 x 224</td>
<td>Layer1: Filtersize: 11, 55x55, 27x5, 13x3, 13x3, 13x3</td>
<td>11, 55x55, 27x5, 13x3, 13x3, 13x3</td>
<td>4096</td>
<td>100, 0</td>
</tr>
<tr>
<td>ZFNet</td>
<td>224 x 224</td>
<td>Layer1: Filtersize: 7, 55x55, 27x5, 13x3, 13x3, 13x3</td>
<td>7, 55x55, 27x5, 13x3, 13x3, 13x3</td>
<td>4096</td>
<td>100, 0</td>
</tr>
<tr>
<td>VGGNet</td>
<td>224 x 224</td>
<td>Layer1: Filtersize: 3, 55x55, 27x5, 13x3, 13x3, 13x3</td>
<td>3, 55x55, 27x5, 13x3, 13x3, 13x3</td>
<td>4096</td>
<td>100, 0</td>
</tr>
</tbody>
</table>

 Altogether, FastRCNN is one step process whereas it adds the loss obtained by classifier and bounding box regressor back to the convolutional layer and finetunes the features. Whereas other networks follow 2-step or 3-step processes because the error obtained by classifier, SVM and bounding box regressor are backpropagated separately and features are finetuned. Due to the fact that Fast R-CNN follows one step process, it is 149 times faster than the R-CNN and 22 times faster than the SPPNet architecture. It gives an accuracy of about 66.9%.

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 classifier and the other one for bounding box regression which has 9 different regressors for three basic shapes of window (square box, height wise rectangular box and width wise rectangular box) with 9 different aspect ratios that will fit all the objects in the image.

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 of Faster R–CNN which further adds instance level segmentation to the network. While in Faster R – CNN, output is a bounding box and classification of the objects in the image, Mask R – CNN gives three classes of output, the two being already said, the third one is mask generated around the different instances of the object. To generate mask, [18][19][20] fully connected layers are added separately and ROI is given as the input to it. Mask will be generated by computing pixel to pixel calculation. Since one image might have several ROI for a single object, before giving it as input to the mask part of the network, ROI is re-calculated by comparing it with ground truth box and finding IoU (Intersection over Union), if the value is above 0.5, it is considered as ROI or else the corresponding box is discarded from ROI. Also, here instead of ROI pool, a new technique ROI align is used. ROI pool uses SPP and maxpools the features but it might have minor differences while projecting the features which might not affect the classification but these differences makes a major impact in pixel to pixel computation while generating mask. To overcome this, ROI align is used where the feature maps are aligned not based on pixel grid division, but by using bilinear interpolation and dividing the pixel in the feature maps into exact floating point numbers and aligning which will help pinpoint the exact features thereby generating the pixel wise masks easily. Mask R – CNN gives better accurate results and also helps in instance level segmentation.

Table 4. Criteria’s to be satisfied by proposed technique to replace Selective Search Technique

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)should be able to propose  &lt;  2000 region proposals.</td>
<td>✔️</td>
</tr>
<tr>
<td>(ii)should be faster than Selective Search</td>
<td>✔️</td>
</tr>
<tr>
<td>(iii)should be accurate or better than the Selective Search</td>
<td>✔️</td>
</tr>
<tr>
<td>(iv)should be able to propose</td>
<td>✔️</td>
</tr>
<tr>
<td>- Overlapping ROI’s</td>
<td>✔️</td>
</tr>
<tr>
<td>- With different aspect ratio’s</td>
<td>✔️</td>
</tr>
<tr>
<td>- With different scales</td>
<td>✔️</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Network Combination</th>
<th>Advantage/Disadvantage</th>
<th>Satisfies the Criteria?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast R CNN + Sliding Window + Image Pyramid</td>
<td>Image Pyramid technique is four times slower</td>
<td>Doesn’t satisfy (ii)</td>
</tr>
<tr>
<td>Fast R CNN + Feature Pyramid</td>
<td>Generates 9 different region proposals for each Feature Map. For a standard FM of size 40<em>60 gives 40</em>60*9 = 20000 proposals to ROI pooling layer.</td>
<td>Doesn’t satisfy (i)</td>
</tr>
<tr>
<td>Fast R CNN + Neural Network</td>
<td>Gives 300 to 500 region proposals using Sliding Window/Dense Sampling techniques</td>
<td>Satisfies (i), (ii), (iii), (iv)</td>
</tr>
</tbody>
</table>

4. Conclusion
Different Object Detection framework has been analyzed along with its advantage and disadvantage. How the disadvantage of each technique has been overcome with each new technique all along has been discussed. The state-of-the-art performance of each breakthrough framework during every year of ILSVRC along with top 5 errors in terms of mAP has been tabulated and the performance of each network has been analyzed. The techniques and networks considered above achieved the breakthrough when compared to the existing technique that has paved way to the innovation in the field of Computer Vision. Though the error rate starting from 28-30 percent in classical CV techniques has been incredibly and exponentially reduced through the years, mainly contributed by the Convolutional Neural Network family, it still has a long way to go to match the accuracy and speed of the human visual cortex.

References

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.
60. Xiaozhi Chen, Huimin Ma, Xiang Wang and Zhichen Zhao Improving Object Proposals with Multi-Thresholding Straddling Expansion *Computer Vision and Pattern Recognition*, 2015.