A Review on Applications of Object Detection Algorithms in Airports

Nandhini S 1, Easwaramoorthy D 2 , Abinandan 3

1 2 Department of Mathematics, School of Advanced Sciences, VIT, Vellore, India

3School of Mechanical Engineering, VIT, Vellore, India

Email: [nandhini.s@vit.ac.in](mailto:nandhini.s@vit.ac.in) 1

ABSTRACT

*In this paper, the basic algorithms of Object Detection using Computer Vision is been discussed.*

*Keywords: Object Detection, Airports, Aviation, Aircraft Maintenance, Computer Vision, Deep learning, Convolutional Neural Networks*

1. INTRODUCTION

This article will be useful as a guide to implement Computer Vision techniques across various applications.. We first begin by comparing the diﬀerent algorithms of Object Detection. We start with an overview of the algorithm of SVM, HOG and DPM algorithms. After analyzing the traditional Computer Vision algorithms, we look at Convolutional Neural Networks. We also analyze how object detection kept evolving and getting better from Spatial Pyramid Pooling to Faster RCNN. Then we present the study of YOLO algorithm and finally the Multi Scale Deformable Algorithm which stands to be the most eﬃcient of today.

1. REVIEW METHODOLOGY

In this paper, we review the algorithms and applications of Object Detection using Computer Vision. In Section III, we have presented a comparative study of different algorithms.

1. ALGORITHMS OF OBJECT DETECTION
   1. SVM (Support Vector Machine ) [1]

Object detection process involves lots of factors that has to be taken into account for the detection. SVM is one of the effective algorithm, which helps in classification of objects. It looks at the extremes of the dataset and draws a decision boundary. This decision boundary is known as a hyperplane. It segregates the dataset into two groups. The problem arises in drawing the decision boundary, we can draw it in many ways using diﬀerent angles. The optimal decision boundary is important to classify a lion and a tiger. All these boundaries are support vectors. D+ represents the vectors towards the positive direction from the hyperplane and D- represents the vectors towards the negative direction.

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane.

Here, f(x) is the function of the hyperplane. W represents the slope of the line and b represents the y intercept. x is the value along the abscissa.

There might be cases where it might be almost impossible to separate the two classes. In these cases, Linear Support Vector Machine algorithm (LSVM) is used in which, we convert the dimensional plane to dimensional plane. We can also convert 2-D to 3-D and draw a hyperplane. This is non-Linear Support Vector Machine. The only disadvantage is the high computational power required.

ABSTRACTS:

Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," in IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 415-425, March 2002, doi: 10.1109/72.991427.

**Abstract:**

Support vector machines (SVMs) were originally designed for binary classification. How to effectively extend it for multiclass classification is still an ongoing research issue. Several methods have been proposed where typically we construct a multiclass classifier by combining several binary classifiers. Some authors also proposed methods that consider all classes at once. As it is computationally more expensive to solve multiclass problems, comparisons of these methods using large-scale problems have not been seriously conducted. Especially for methods solving multiclass SVM in one step, a much larger optimization problem is required so up to now experiments are limited to small data sets. In this paper we give decomposition implementations for two such "all-together" methods. We then compare their performance with three methods based on binary classifications: "one-against-all," "one-against-one," and directed acyclic graph SVM (DAGSVM). Our experiments indicate that the "one-against-one" and DAG methods are more suitable for practical use than the other methods. Results also show that for large problems methods by considering all data at once in general need fewer support vectors.

Classification of Hyperspectral Remote Sensing Images With Support Vector Machines Farid Melgani, Member, IEEE, and Lorenzo Bruzzone, Senior Member, IEEE

Abstract:

This paper addresses the problem of the classification of hyperspectral remote sensing images by support vector machines (SVMs). First, we propose a theoretical discussion and experimental analysis aimed at understanding and assessing the potentialities of SVM classifiers in hyperdimensional feature spaces. Then, we assess the effectiveness of SVMs with respect to conventional feature-reduction-based approaches and their performances in hypersubspaces of various dimensionalities. To sustain such an analysis, the performances of SVMs are compared with those of two other nonparametric classifiers (i.e., radial basis function neural networks and the K-nearest neighbor classifier). Finally, we study the potentially critical issue of applying binary SVMs to multiclass problems in hyperspectral data. In particular, four different multiclass strategies are analyzed and compared: the one-against-all, the one-against-one, and two hierarchical tree-based strategies. Different performance indicators have been used to support our experimental studies in a detailed and accurate way, i.e., the classification accuracy, the computational time, the stability to parameter setting, and the complexity of the multiclass architecture. The results obtained on a real Airborne Visible/Infrared Imaging Spectroradiometer hyperspectral dataset allow to conclude that, whatever the multiclass strategy adopted, SVMs are a valid and effective alternative to conventional pattern recognition approaches (feature-reduction procedures combined with a classification method) for the classification of hyperspectral remote sensing data.

Abstract

Local space-time features capture local events in video and can be adapted to the size, the frequency and the velocity of moving patterns. In this paper we demonstrate how such features can be used for recognizing complex motion patterns. We construct video representations in terms of local space-time features and integrate such representations with SVM classification schemes for recognition. For the purpose of evaluation we introduce a new video database containing 2391 sequences of six human actions performed by 25 people in four different scenarios. The presented results of action recognition justify the proposed method and demonstrate its advantage compared to other relative approaches for action recognition.

E. Osuna, R. Freund and F. Girosit, "Training support vector machines: an application to face detection," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, USA, 1997, pp. 130-136, doi: 10.1109/CVPR.1997.609310.

Abstract:

**Abstract:**

We investigate the application of Support Vector Machines (SVMs) in computer vision. SVM is a learning technique developed by V. Vapnik and his team (AT&T Bell Labs., 1985) that can be seen as a new method for training polynomial, neural network, or Radial Basis Functions classifiers. The decision surfaces are found by solving a linearly constrained quadratic programming problem. This optimization problem is challenging because the quadratic form is completely dense and the memory requirements grow with the square of the number of data points. We present a decomposition algorithm that guarantees global optimality, and can be used to train SVM's over very large data sets. The main idea behind the decomposition is the iterative solution of sub-problems and the evaluation of optimality conditions which are used both to generate improved iterative values, and also establish the stopping criteria for the algorithm. We present experimental results of our implementation of SVM, and demonstrate the feasibility of our approach on a face detection problem that involves a data set of 50,000 data points.

* 1. HOG Algorithm (Histogram of Oriented Gradients) [2] [3]

The algorithm works by having an input of an image with an aspect ratio of .We convert the image into by dimension. In some cases, Gamma correction can be used to improve the performance gain. Then the horizontal and vertical gradients are calculated and the function used for horizontal and vertical gradient is given by equation and .

------------------(2)

-------------------(3)

The magnitude of the gradient is taken for the entire image. After that, gradients are found in each cell after splitting the image into cells. This is done for a compact representation. Furthermore, an image patch would contain 192 pixels and it would be robust to noise. Then a histogram of gradients is created in these cells. The histogram contains 9 bins corresponding to angles. The gradient magnitude is grouped corresponding to the gradient directions. Also, the angles greater than degree contributes proportionally to the zero degree bin and degree bin.

The next step is to perform Block normalization. This is to make sure the algorithm isn’t aﬀected by lighting. Finally the HOG feature vector is calculated. The histogram obtained demonstrates that the pixel from background give much lower accumulating result than the pixel from the object.

We typically use HOG for feature extraction and SVM for classification. For the training phase, we take images with people and 100 images without people. Then we can take the HOG features from all of these. We can then train the computer to segregate it into human and non-human by providing it to the SVM classifier. Once we have done the image training, we extract Hog features. Each block is represented by a vector. When we concatenate it all into one giant vector, we obtain a dimensional vector. We multiply these and add a bias term. If the result is positive, it’s that of a human and if it’s negative, it’s that of a non-human.

The algorithm can give confidence score for each location in the image. Near the person, the confidence score is really high. Selection of a confidence score threshold accordingly. We have to choose a good confidence score to obtain a proper trade off between specificity and sensitivity. To avoid the use of duplicate boxes on the object, non max suppression is used. Non max suppression is the technique used to remove the boxes which gets overlapped.

To detect the coordinates of multiple people, the algorithm uses a technique of sliding window. It goes through the entire image through a box of small dimensions and slides through all the possibilities. If the case arises that only part of the person is identified, we use the concept of image pyramid. Image Pyramid is the process in which the image gets downscaled till the sliding window identified the entire person.

The main drawback for this image algorithm is that it doesn’t catch the object in certain poses or deformations. Humans are deformable and has many poses unlike non living objects. Thus the algorithm may fail to detect humans in certain poses. DPM is the algorithm which takes care of this aspect.

N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol:. 1, doi: 10.1109/CVPR.2005.177.

Abstract:

We study the question of feature sets for robust visual object recognition; adopting linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, we show experimentally that grids of histograms of oriented gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

Q. V. Le, W. Y. Zou, S. Y. Yeung and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," CVPR 2011, Providence, RI, 2011, pp. 3361-3368, doi: 10.1109/CVPR.2011.5995496.

Abstract:

Previous work on action recognition has focused on adapting hand-designed local features, such as SIFT or HOG, from static images to the video domain. In this paper, we propose using unsupervised feature learning as a way to learn features directly from video data. More specifically, we present an extension of the Independent Subspace Analysis algorithm to learn invariant spatio-temporal features from unlabeled video data. We discovered that, despite its simplicity, this method performs surprisingly well when combined with deep learning techniques such as stacking and convolution to learn hierarchical representations. By replacing hand-designed features with our learned features, we achieve classification results superior to all previous published results on the Hollywood2, UCF, KTH and YouTube action recognition datasets. On the challenging Hollywood2 and YouTube action datasets we obtain 53.3% and 75.8% respectively, which are approximately 5% better than the current best published results. Further benefits of this method, such as the ease of training and the efficiency of training and prediction, will also be discussed. You can download our code and learned spatio-temporal features here: http://ai.stanford.edu/~wzou/.

O. Oreifej and Z. Liu, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences," 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 716-723, doi: 10.1109/CVPR.2013.98.

**Abstract:**

We present a new descriptor for activity recognition from videos acquired by a depth sensor. Previous descriptors mostly compute shape and motion features independently, thus, they often fail to capture the complex joint shape-motion cues at pixel-level. In contrast, we describe the depth sequence using a histogram capturing the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. To build the histogram, we create 4D projectors, which quantize the 4D space and represent the possible directions for the 4D normal. We initialize the projectors using the vertices of a regular polychoron. Consequently, we refine the projectors using a discriminative density measure, such that additional projectors are induced in the directions where the 4D normals are more dense and discriminative. Through extensive experiments, we demonstrate that our descriptor better captures the joint shape-motion cues in the depth sequence, and thus outperforms the state-of-the-art on all relevant benchmarks.

Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng and S. Avidan, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 2006, pp. 1491-1498, doi: 10.1109/CVPR.2006.119.

**Abstract:**

We integrate the cascade-of-rejectors approach with the Histograms of Oriented Gradients (HoG) features to achieve a fast and accurate human detection system. The features used in our system are HoGs of variable-size blocks that capture salient features of humans automatically. Using AdaBoost for feature selection, we identify the appropriate set of blocks, from a large set of possible blocks. In our system, we use the integral image representation and a rejection cascade which significantly speed up the computation. For a 320 × 280 image, the system can process 5 to 30 frames per second depending on the density in which we scan the image, while maintaining an accuracy level similar to existing methods.

R. Chaudhry, A. Ravichandran, G. Hager and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 1932-1939, doi: 10.1109/CVPR.2009.5206821.

**Abstract:**

System theoretic approaches to action recognition model the dynamics of a scene with linear dynamical systems (LDSs) and perform classification using metrics on the space of LDSs, e.g. Binet-Cauchy kernels. However, such approaches are only applicable to time series data living in a Euclidean space, e.g. joint trajectories extracted from motion capture data or feature point trajectories extracted from video. Much of the success of recent object recognition techniques relies on the use of more complex feature descriptors, such as SIFT descriptors or HOG descriptors, which are essentially histograms. Since histograms live in a non-Euclidean space, we can no longer model their temporal evolution with LDSs, nor can we classify them using a metric for LDSs. In this paper, we propose to represent each frame of a video using a histogram of oriented optical flow (HOOF) and to recognize human actions by classifying HOOF time-series. For this purpose, we propose a generalization of the Binet-Cauchy kernels to nonlinear dynamical systems (NLDS) whose output lives in a non-Euclidean space, e.g. the space of histograms. This can be achieved by using kernels defined on the original non-Euclidean space, leading to a well-defined metric for NLDSs. We use these kernels for the classification of actions in video sequences using (HOOF) as the output of the NLDS. We evaluate our approach to recognition of human actions in several scenarios and achieve encouraging results.

* 1. DPM [4] (Deformable Parts Model)

This model builds upon the previous model. DPM is a model that came about to solve the problem of object detection during different poses in images. To take the pose into account, each body part is detected. But there could be a possibility that there may be multiple legs, arms or other body parts in the crowd. For solving this, the idea of penalty scores were introduced. If the body part is closer, the penalty is lesser. The score for the whole body is taken, then the score for each body part is added and finally the penalties are subtracted from the resultant value. HOG and SVM is typically used together for each body part and then they are summed up.

It is useful even for other objects like cars. There could be detectors for the side view, top view and front view. This is especially useful when the object is blocked or is at a different angle. So each part of the car is also having a separate detector. It is useful when the car’s door is open or some part is deformed.

There is a course filter for the entire object detected and there will be multiple higher resolution part filters for each part. The model has root filter 0 and part models .The score of the hypothesis depends on the root filter and part filter.

Equation 5 represents the sum of the root filter and part filters in the first term. The second term represents the penalties and b represents the bias. The root occurs at and the part occurs at . represents the filters. represents the features of subwindows at location. represents the deformation parameters. is the displacement of part I relative to its anchor position. Then b is the bias. The first term is referred to as the data term and the second term is referred to as the spatial prior.

-------- (4)

B is the set which contains all the unknowns which are the filters, deformation and the bias. is the known term above. It’s the response of the algorithm for each part filter and the root filter.

------------- (5)

---------------(6)

We multiply the two terms and to compute the score.

---------------------(7)

is the known term above and is the term which will be computed.

If we notice the deformation cost. It’s a four dimensional vector. is also a four dimensional vector.

Initially when we take,the distance would be given by .Depending on the values dx and dy, the locus of displacement would be a circle or ellipse.

The overall score of a root location is computed by the best possible placement of the parts. After training this model, we want to make sure that the human detection is confirmed by the parts.

--------------(10)

Equation 9 represents the score at the root location .

There might be cases in which a body part of another person might be considered by mistake. For avoiding this error, a dynamic programming and generalized distance transform is implemented. It is the process of finding the closest part, to avoid errors.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627-1645, Sept. 2010, doi: 10.1109/TPAMI.2009.167.

Abstract—We describe an object detection system based on mixtures of multiscale deformable part models. Our system is able to represent highly variable object classes and achieves state-of-the-art results in the PASCAL object detection challenges. While deformable part models have become quite popular, their value had not been demonstrated on difficult benchmarks such as the PASCAL data sets. Our system relies on new methods for discriminative training with partially labeled data. We combine a marginsensitive approach for data-mining hard negative examples with a formalism we call latent SVM. A latent SVM is a reformulation of MI-SVM in terms of latent variables. A latent SVM is semiconvex, and the training problem becomes convex once latent information is specified for the positive examples. This leads to an iterative training algorithm that alternates between fixing latent values for positive examples and optimizing the latent SVM objective function.

P. Felzenszwalb, D. McAllester and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587597.

**Abstract:**

This paper describes a discriminatively trained, multiscale, deformable part model for object detection. Our system achieves a two-fold improvement in average precision over the best performance in the 2006 PASCAL person detection challenge. It also outperforms the best results in the 2007 challenge in ten out of twenty categories. The system relies heavily on deformable parts. While deformable part models have become quite popular, their value had not been demonstrated on difficult benchmarks such as the PASCAL challenge. Our system also relies heavily on new methods for discriminative training. We combine a margin-sensitive approach for data mining hard negative examples with a formalism we call latent SVM. A latent SVM, like a hidden CRF, leads to a non-convex training problem. However, a latent SVM is semi-convex and the training problem becomes convex once latent information is specified for the positive examples. We believe that our training methods will eventually make possible the effective use of more latent information such as hierarchical (grammar) models and models involving latent three dimensional pose.

D. Terzopoulos and D. Metaxas, "Dynamic 3D models with local and global deformations: deformable superquadrics," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 7, pp. 703-714, July 1991, doi: 10.1109/34.85659.

**Abstract:**

The authors present a physically based approach to fitting complex three-dimensional shapes using a novel class of dynamic models that can deform both locally and globally. They formulate the deformable superquadrics which incorporate the global shape parameters of a conventional superellipsoid with the local degrees of freedom of a spline. The model's six global deformational degrees of freedom capture gross shape features from visual data and provide salient part descriptors for efficient indexing into a database of stored models. The local deformation parameters reconstruct the details of complex shapes that the global abstraction misses. The equations of motion which govern the behavior of deformable superquadrics make them responsive to externally applied forces. The authors fit models to visual data by transforming the data into forces and simulating the equations of motion through time to adjust the translational, rotational, and deformational degrees of freedom of the models. Model fitting experiments involving 2D monocular image data and 3D range data are presented.< >

P. F. Felzenszwalb, R. B. Girshick and D. McAllester, "Cascade object detection with deformable part models," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 2241-2248, doi: 10.1109/CVPR.2010.5539906.

**Abstract:**

We describe a general method for building cascade classifiers from part-based deformable models such as pictorial structures. We focus primarily on the case of star-structured models and show how a simple algorithm based on partial hypothesis pruning can speed up object detection by more than one order of magnitude without sacrificing detection accuracy. In our algorithm, partial hypotheses are pruned with a sequence of thresholds. In analogy to probably approximately correct (PAC) learning, we introduce the notion of probably approximately admissible (PAA) thresholds. Such thresholds provide theoretical guarantees on the performance of the cascade method and can be computed from a small sample of positive examples. Finally, we outline a cascade detection algorithm for a general class of models defined by a grammar formalism. This class includes not only tree-structured pictorial structures but also richer models that can represent each part recursively as a mixture of other parts.

M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 1307-1314, doi: 10.1109/ICCV.2011.6126383.

**Abstract:**

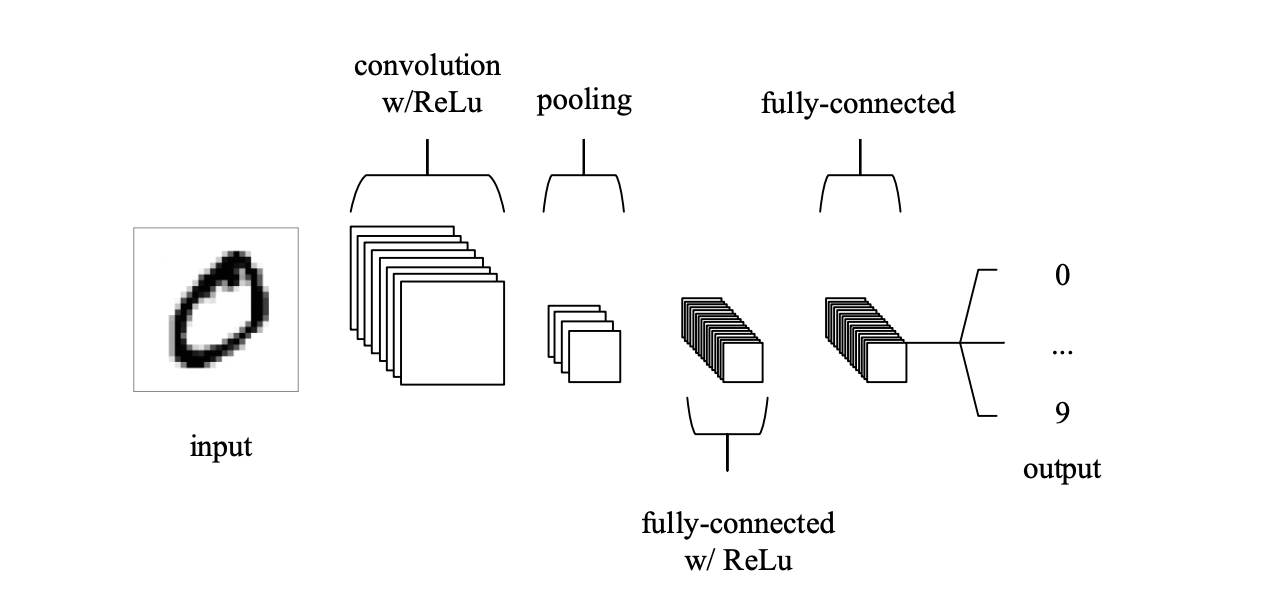
Weakly supervised discovery of common visual structure in highly variable, cluttered images is a key problem in recognition. We address this problem using deformable part-based models (DPM's) with latent SVM training [6]. These models have been introduced for fully supervised training of object detectors, but we demonstrate that they are also capable of more open-ended learning of latent structure for such tasks as scene recognition and weakly supervised object localization. For scene recognition, DPM's can capture recurring visual elements and salient objects; in combination with standard global image features, they obtain state-of-the-art results on the MIT 67-category indoor scene dataset. For weakly supervised object localization, optimization over latent DPM parameters can discover the spatial extent of objects in cluttered training images without ground-truth bounding boxes. The resulting method outperforms a recent state-of-the-art weakly supervised object localization approach on the PASCAL-07 dataset.

de la Escalera, J. M. Armingol, J. M. Pastor and F. J. Rodriguez, "Visual sign information extraction and identification by deformable models for intelligent vehicles," in IEEE Transactions on Intelligent Transportation Systems, vol. 5, no. 2, pp. 57-68, June 2004, doi: 10.1109/TITS.2004.828173.

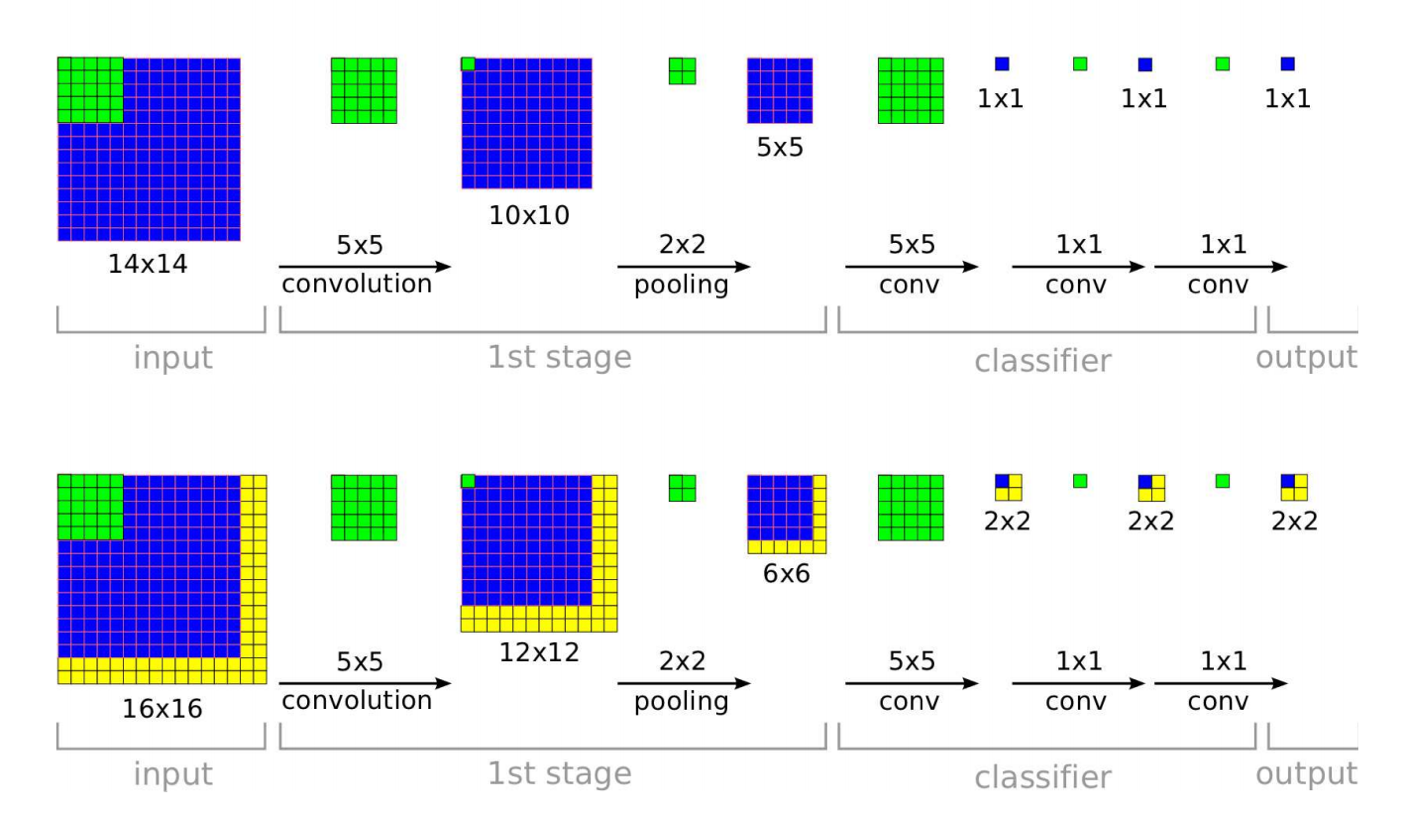
**Abstract:**

This paper deals with the extraction of part of the visual information presented in streets, roads, and motorways. This information, provided by either traffic or road signs and route-guidance signs, is extremely important for safe and successful driving. An automatic system that is capable of extracting and identifying these signs automatically would help human drivers enormously; navigation would be easier and would allow him or her to concentrate on driving the vehicle. The system would indicate to the driver the presence of a sign in advance, so that some incorrect human decisions could be avoided. A deformable model scheme allows us to include the knowledge used while designing the signs in the algorithm and is used for their detection and identification. Two techniques to find the minimum in the energy function are shown: simulated annealing and genetic algorithms. Some problems are addressed, such as uncontrolled lighting conditions; occlusions; and variations in shape, size, and color.

3.5 Convolutional Neural Networks and Overfeat [5][6]

In the previous computer vision techniques, the HOG algorithms acts as feature extractors while SVM acts as a classifier. In a convolutional Neural Network, the convolutional pooling layers would act as feature extractors. The fully connected and softmax layers act as classifiers. The classifier could also be modified to perform localisation which is drawing a bounding box around the detected object. To draw the bounding box we need to get the coordinates of one point and get the height and width. We would have 4 parameters. The last layer of the convolutional layer consists of one fully connected layer followed by softmax. The layer would have the scores for each class. Softmax basically converts the scores to probabilities. In CNN, Bounding Box Regression Training is used to trained to get the bounding box coordinates. We input an image along with the 4 coordinates required for building a bounding box. In the beginning, the layers of CNN would have the weights assigned as 0.1 for every layer. We pass the initial vector of features through the layers to obtain the four coordinates. We calculate the L2 loss which is the difference between the squares of the expected coordinates and the generated coordinates. The value we get is back propogated through the layer. This leads to the value of the weights changing. The neural network is again processed to get 4 coordinates. Then the L2 loss is calculated again. This is done till we get the sum of the L2 losse of each coordinate to be zero. After this, the CNN would be trained to draw a bounding box for the image.

We use the sliding Window technique to detect multiple and crop whenever the object is detected. This would be included as a preprocessing step. Then we convert that image into 224\*224 to draw the bounding box. Finally we merge the cropped images to form the complete images with bounding boxes.

In cases where objects are of different sizes and overlap, we use the sliding window combined with the image pyramid technique. The image pyramid represents the process of resizing the image continuously so that objects of different sizes gets detected at different scales. A confidence score is given to each detection. This would ensure that partial detections don’t get cropped or fed to the CNN. Usually convolutional Neural Networks require a fixed image as the feature vectors would get multiplied at the final layer – the fully connected layer. Due to a lot of advancement, the fully connected layer is used as a convolutional layer. Instead of an array, it is considered as a matrix. So the sliding window technique can be used to detect the objects and it can be fed to the CNN without the requirement of cropping. Furthermore, repeated pixels won’t be run again and this would save computatonal power. We use image pyramid to get spatial output. Spatial output is the matrix of the confidence scores at each sliding window box. This process is the idea of the Overfeat algorithm. The effective stride is the number of pixels in which the algorithm moves if 1 pixel is shifted in the spatial output. The effective stride should be as low as possible.

In the above image, the feature map is a 5X5 output. Using a 5X5 filter, we convolve the feature map to get a 1X1 output. To perform the convolution further we use 1X1 filters as we can’t use a larger filter on a 1X1 output. We use the filter repeatedly to get the output for one class in the dataset. In case of the input images being of different size (as we use the image pyramid), we get a 6X6 feature map. Then when we use a 5X5 filter followed by 1X1 filters to get a 2X2 output. This would lead to the creation of a spatial output for one class. The feature map would be of size 256. We use a 4096 filters each of size 5X5 to get an output of 4096. We then use 4096X4096 no of filters to get the next output as 4096. We aim to get the final output in the size of the number of classes. The size of the classes including the background is represented by C(No of classes + 1). So we then use 4096XC filters to get C number of classes. We always design the filters keeping in mind the final output size we need. The feature maps would usually be run through one set of layers to perform object classification and another set of layer to perform the bounding box regression.

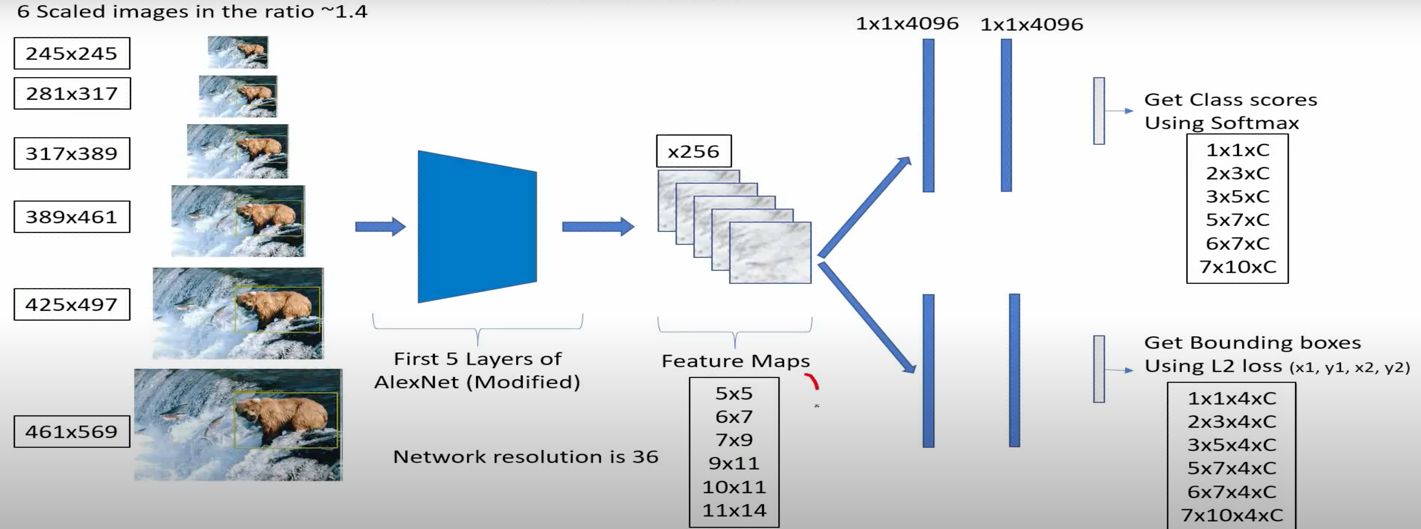


Image credit:https://host.robots.ox.ac.uk/pascal/VOC/voc2012/examples.index.html

A. S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, 2014, pp. 512-519, doi: 10.1109/CVPRW.2014.131.

**Abstract:**

Recent results indicate that the generic descriptors extracted from the convolutional neural networks are very powerful. This paper adds to the mounting evidence that this is indeed the case. We report on a series of experiments conducted for different recognition tasks using the publicly available code and model of the OverFeat network which was trained to perform object classification on ILSVRC13. We use features extracted from the OverFeat network as a generic image representation to tackle the diverse range of recognition tasks of object image classification, scene recognition, fine grained recognition, attribute detection and image retrieval applied to a diverse set of datasets. We selected these tasks and datasets as they gradually move further away from the original task and data the OverFeat network was trained to solve. Astonishingly, we report consistent superior results compared to the highly tuned state-of-the-art systems in all the visual classification tasks on various datasets. For instance retrieval it consistently outperforms low memory footprint methods except for sculptures dataset. The results are achieved using a linear SVM classifier (or L2 distance in case of retrieval) applied to a feature representation of size 4096 extracted from a layer in the net. The representations are further modified using simple augmentation techniques e.g. jittering. The results strongly suggest that features obtained from deep learning with convolutional nets should be the primary candidate in most visual recognition tasks.

P. Chen and S. Ho, "Is overfeat useful for image-based surface defect classification tasks?," *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, pp. 749-753, doi: 10.1109/ICIP.2016.7532457.

**Abstract:**

One of the challenges for real-world image-based surface defect classification task is the lack of labeled training samples to extract useful features to confidently classify defects. In this paper, we present results on our investigation on whether features derived from OverFeat, a variant of Convolution Neural Network, can be used directly for image-based surface defect classification task. We show that the classification performance of two real-world defect images datasets can be significantly different. For the harder classification task, OverFeat features are useful for some types of surface defects, but performs poorly when the defects demonstrate characteristics beyond texture patterns. We propose a simple heuristic approach called Approximate Surface Roughness (ASR) that provides auxiliary information on the relationship between spatial regions in the defect image to be used together with the OverFeat features. Empirical results show improvement in classification performance for those defect types that do not classify well using only OverFeat features.

B. van Ginneken, A. A. A. Setio, C. Jacobs and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, New York, NY, 2015, pp. 286-289, doi: 10.1109/ISBI.2015.7163869.

**Abstract:**

Convolutional neural networks (CNNs) have emerged as the most powerful technique for a range of different tasks in computer vision. Recent work suggested that CNN features are generic and can be used for classification tasks outside the exact domain for which the networks were trained. In this work we use the features from one such network, OverFeat, trained for object detection in natural images, for nodule detection in computed tomography scans. We use 865 scans from the publicly available LIDC data set, read by four thoracic radiologists. Nodule candidates are generated by a state-of-the-art nodule detection system. We extract 2D sagittal, coronal and axial patches for each nodule candidate and extract 4096 features from the penultimate layer of OverFeat and classify these with linear support vector machines. We show for various configurations that the off-the-shelf CNN features perform surprisingly well, but not as good as the dedicated detection system. When both approaches are combined, significantly better results are obtained than either approach alone. We conclude that CNN features have great potential to be used for detection tasks in volumetric medical data

L. Trottier, P. Giguere and B. Chaib-draa, "Parametric Exponential Linear Unit for Deep Convolutional Neural Networks," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, 2017, pp. 207-214, doi: 10.1109/ICMLA.2017.00038.

**Abstract:**

Object recognition is an important task for improving the ability of visual systems to perform complex scene understanding. Recently, the Exponential Linear Unit (ELU) has been proposed as a key component for managing bias shift in Convolutional Neural Networks (CNNs), but defines a parameter that must be set by hand. In this paper, we propose learning a parameterization of ELU in order to learn the proper activation shape at each layer in the CNNs. Our results on the MNIST, CIFAR-10/100 and ImageNet datasets using the NiN, Overfeat, All-CNN and ResNet networks indicate that our proposed Parametric ELU (PELU) has better performances than the non-parametric ELU. We have observed as much as a 7.28% relative error improvement on ImageNet with the NiN network, with only 0.0003% parameter increase. Our visual examination of the non-linear behaviors adopted by Vgg using PELU shows that the network took advantage of the added flexibility by learning different activations at different layers.

Y. Saleh and E. Edirisinghe, "3D face reconstruction and recognition using the overfeat network," 2017 8th International Conference on Information and Communication Systems (ICICS), Irbid, 2017, pp. 116-119, doi: 10.1109/IACS.2017.7921956.

**Abstract:**

Although face recognition is considered a popular area of research and study, it still has few unresolved challenges, and with the appearance of devices such as the Microsoft Kinect, new possibilities for researchers were uncovered. With the goal of enhancing face recognition techniques, this paper presents a novel way to reconstruct face images in different angles, through the use of the data of one front image captured by the Kinect, using faster techniques than ever before, also, this paper utilizes a deep learning network called Overfeat, where it functioned as a feature extractor that was used on normal images and on the new 3D created images, which introduced a new application for the network. To check the capabilities of the new created images, they were used as a testing set in three main experiments. Finally, results of the experiments are presented to prove the ability of the created images to function as new data sets for face recognition; also, proving the capability of the Overfeat network, working with computer generated face images.

C. Desai, J. Eledath, H. Sawhney and M. Bansal, "De-correlating CNN Features for Generative Classification," 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, 2015, pp. 428-435, doi: 10.1109/WACV.2015.63.

**Abstract:**

The problem of training a classifier from a handful of positive examples, without having to supply class specific negatives is of great practical importance. The proposed approach to solving this problem builds on the idea of training LDA classifiers using only class specific foreground images and a large collection of unlabelled images, as described in [11]. While we adopt the LDA training methodology of [11], we depart from HOG features and work with those extracted from a Convolutional Neural Network (CNN) pre-trained on Image Net (Over feat). We combine Over feat features with the LDA training methodology to derive generative classifiers. When evaluated on a K-way classification problem, these classifiers are almost as good as those trained discriminatively using the same features. Unlike the HOG based approach of [11], our classifiers do not need any post-processing step of calibration, a step that requires positives and negatives. Finally, we show that in an instance retrieval setup, we can employ these generative classifiers to derive a novel query-expansion framework that achieves a significant performance boost by utilizing only the top ranked positive examples from an initial nearest-neighbor list.

M. Korytkowski, P. Staszewski, P. Woldan and R. Scherer, "Fast Computing Framework for Convolutional Neural Networks," 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), Atlanta, GA, 2016, pp. 118-123, doi: 10.1109/BDCloud-SocialCom-SustainCom.2016.28.

**Abstract:**

In the case of building large convolutional neural networks, signal propagation speed is one of priority factors. Training large neural structures requires enormous time for achieving satisfying accuracy. In addition, the networks need to be learn by very large sets of good quality training images, which is another time-consuming factor. The paper presents a fast computing framework with some methods to optimize the signal propagation speed. We compare our implementation with the original OverFeat implementation.

* 1. SPPNet [7][8] (Spatial Pyramid Pooling)

Spatial Pyramid pooling is an extension of the Bag of words model. If there is an image which is black and white, black would have the pixel intensity value of and the white would have pixel intensity value of . For plotting a histogram, we plot the number of pixels according to the pixel intensity which ranges from . The more colors that are present, the more detailed the image. Even if the images are expanded or cropped, the histograms would be the same.

There is one more way of drawing histograms. Usually we actually calculated each pixel intensity value from 0 to 255. Instead we could group pixels that are similar to each other into one single bin. For example, we group pixels from 0-50 in a bin. Then we group pixels from 50 -100 into another bin. This is done till 255. This is ho we plot histograms by using intensity bins. This is a more sophisticated way of classifying all the images into one bin. This leads to less memory being used. Only 5 bytes would be used instead of 255 bytes as used in the usual method. Labels are usually allotted to each intensity bin. The list of labels are called as the codebook.

We can also take the histogram of each pixel using HOG feature descriptors. We first divide the image into a grid. The features are clustered according to their similarity. Then, a histogram is created. Once we extract the features, we make the codebook. We use the technique on k-means clustering to group similar features into a single layer.

The main steps of k-means++ clustering are:

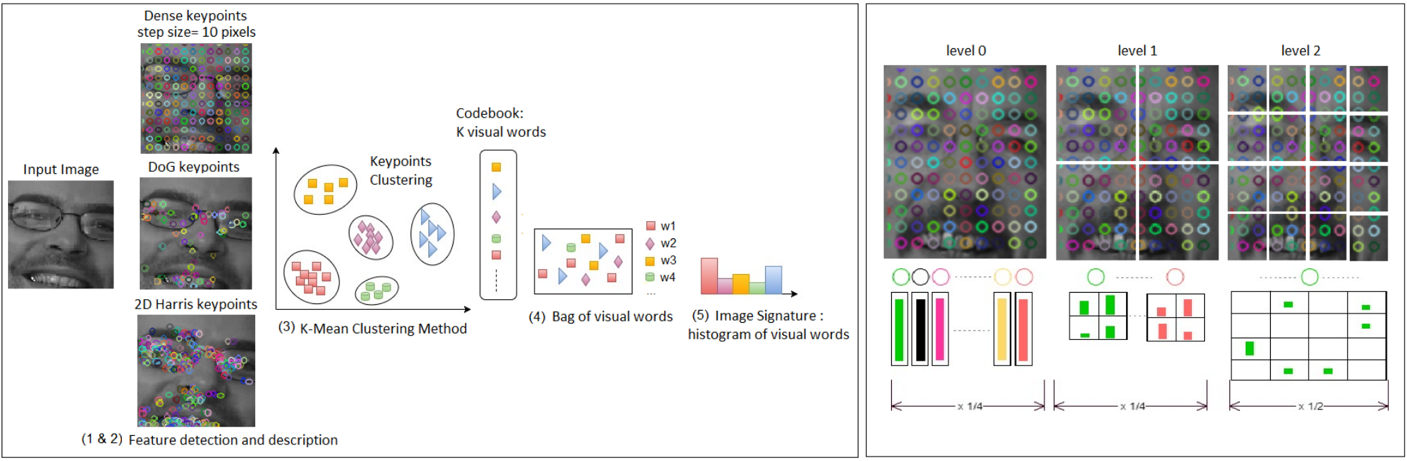
1. Choose an initial center uniformly at random from the data points.

2. For each data point x, compute D(x), the distance between x and the nearest center that has already been chosen.

3. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to D(x) 2 .

4. Repeat steps 2 and 3 until a total of k centers has been selected.

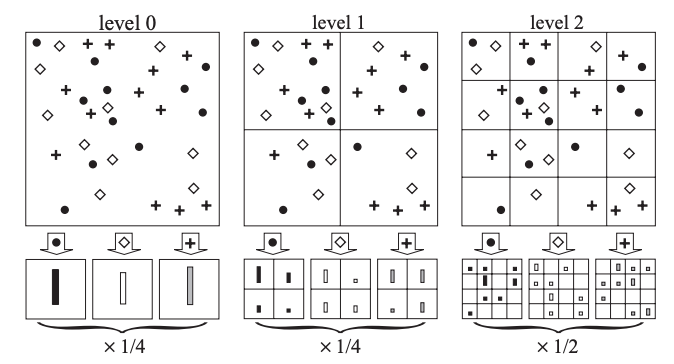
5. Proceed as with standard k-means algorithm.



This technique is called bag of visual words. It can be used for faces as shown above.

Using this technique, we can determine what kind of image it is.

When there are two images - one zoomed in and one zoomed out, the histogram looks the same. Spatial Pyramid matching is been performed as follows. The image is divided into blocks and represent histograms separately. It is further divided into blocks and a histogram is created. So the first level would be of length vectors, the second level would be of length vectors and the third level would be of level vectors. So the total length would add up to vectors. This would allow us to notice the diﬀerence in the images as a lot of blocks would have diﬀerent histograms.



The bag of words method categorises images even if it is cropped and transformed. If the image is divided into 4 distinct parts and the 4 parts are interchanged. The bag of visual words method creates the same histogram. To differentiate between these images, we mainly use spatial pyramid matching. We do the processing in three levels. Level 0 is just the ordinary bag of visual words representation. In level one, we divide the image into a 2X2 grid. We calculate the histogram separately for these 4 boxes. In the next level, we do the same with a 4X4 grid. The total histogram would be a concatenated version of the three levels. In level 0, we had a vector size of 3. In level one, we have a vector size of 12. In level two, the vector size is 48. The three levels are concatenated to get a vector of size 63. This will allow us to differentiate different images. This method doesn’t matter on the size or aspect ratio of the feature map.

Spatial pyramid matching just takes the maximum values in each block. This leads to dimension. For images of diﬀerent sizes, the aspect ratio is maintained to be the same instead of varying. The smaller size is maintained at a fixed dimension while the other size could vary.

SPP partitions the image into divisions from finer to coarser levels, and aggregates local features in them. SPP is able to generate a fixed length output regardless of the input size, while the sliding window pooling used in the previous algorithms could not. SPP uses multi-level spatial bins, while the sliding window pooling uses only a single window size. SPP can pool features extracted at variable scales thanks to the flexibility of input scales

In SPPNet, first the image is passed through a network like Selective Search. [9]

It leads to the generation of 2000 Region proposals. To pass this on to a pertained model on ImageNet takes a lot of time. To solve this problem, we pass the image through the network proposal separately. Next we translate it directly on to the feature map. Then we apply SPP pooling for each region separately. To translate the ROI proposal onto feature map, we pass the image through 3 pooling layers. We do it till we get an image of dimensions . The ratio between the initial and final image dimension is called the subsampling ratio. This allows us to correctly identify the object and draw a bounding box around it alone.

K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.

**Abstract:**

Existing deep convolutional neural networks (CNNs) require a fixed-size (e.g., 224 × 224) input image. This requirement is “artificial” and may reduce the recognition accuracy for the images or sub-images of an arbitrary size/scale. In this work, we equip the networks with another pooling strategy, “spatial pyramid pooling”, to eliminate the above requirement. The new network structure, called SPP-net, can generate a fixed-length representation regardless of image size/scale. Pyramid pooling is also robust to object deformations. With these advantages, SPP-net should in general improve all CNN-based image classification methods. On the ImageNet 2012 dataset, we demonstrate that SPP-net boosts the accuracy of a variety of CNN architectures despite their different designs. On the Pascal VOC 2007 and Caltech101 datasets, SPP-net achieves state-of-the-art classification results using a single full-image representation and no fine-tuning. The power of SPP-net is also significant in object detection. Using SPP-net, we compute the feature maps from the entire image only once, and then pool features in arbitrary regions (sub-images) to generate fixed-length representations for training the detectors. This method avoids repeatedly computing the convolutional features. In processing test images, our method is 24-102 × faster than the R-CNN method, while achieving better or comparable accuracy on Pascal VOC 2007. In ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, our methods rank #2 in object detection and #3 in image classification among all 38 teams. This manuscript also introduces the improvement made for this competition.

Q. Liu, R. Hang, H. Song and Z. Li, "Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 1, pp. 117-126, Jan. 2018, doi: 10.1109/TGRS.2017.2743243.

**Abstract:**

In this paper, we propose a multiscale deep feature learning method for high-resolution satellite image scene classification. Specifically, we first warp the original satellite image into multiple different scales. The images in each scale are employed to train a deep convolutional neural network (DCNN). However, simultaneously training multiple DCNNs is time-consuming. To address this issue, we explore DCNN with spatial pyramid pooling (SPP-net). Since different SPP-nets have the same number of parameters, which share the identical initial values, and only fine-tuning the parameters in fully connected layers ensures the effectiveness of each network, thereby greatly accelerating the training process. Then, the multiscale satellite images are fed into their corresponding SPP-nets, respectively, to extract multiscale deep features. Finally, a multiple kernel learning method is developed to automatically learn the optimal combination of such features. Experiments on two difficult data sets show that the proposed method achieves favorable performance compared with other state-of-the-art methods.

C. Huang, Z. He, G. Cao and W. Cao, "Task-Driven Progressive Part Localization for Fine-Grained Object Recognition," in IEEE Transactions on Multimedia, vol. 18, no. 12, pp. 2372-2383, Dec. 2016, doi: 10.1109/TMM.2016.2602060. The problem of fine-grained object recognition is very challenging due to the subtle visual differences between different object categories. In this paper, we propose a task-driven progressive part localization (TPPL) approach for fine-grained object recognition. Most existing methods follow a two-step approach that first detects salient object parts to suppress the interference from background scenes and then classifies objects based on features extracted from these regions. The part detector and object classifier are often independently designed and trained. In this paper, our major finding is that the part detector should be jointly designed and progressively refined with the object classifier so that the detected regions can provide the most distinctive features for final object recognition. Specifically, we develop a part-based SPP-net (Part-SPP) as our baseline part detector. We then establish a TPPL framework, which takes the predicted boxes of Part-SPP as an initial guess, and then examines new regions in the neighborhood using a particle swarm optimization approach, searching for more discriminative image regions to maximize the objective function and the recognition performance. This procedure is performed in an iterative manner to progressively improve the joint part detection and object classification performance. Experimental results on the Caltech-UCSD-200-2011 dataset demonstrate that our method outperforms state-of-the-art fine-grained categorization methods both in part localization and classification, even without requiring a bounding box during testing.

X. Ouyang, K. Gu and P. Zhou, "Spatial Pyramid Pooling Mechanism in 3D Convolutional Network for Sentence-Level Classification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 11, pp. 2167-2179, Nov. 2018, doi: 10.1109/TASLP.2018.2852502.

**Abstract:**

In this paper, we investigate the usage of the convolutional neural network (CNN) to propose a novel end-to-end language processing structure to model textual data for this task. In particular, we propose a 3D CNN structure for the task, which is featured by spatial pyramid pooling (SPP). To our knowledge, it is the first time that 3D convolution and SPP structure are applied together in language processing issues. Compared with methods of 2D CNNs, the proposed method can effectively and efficiently capture the complicated internal relations in sentences. Furthermore, in previous work, the issue of sentence length variety is usually addressed by padding zero to make all sentences vectors to a fixed length, which causes too much redundant and useless noise. Inspired by the SPP structure for object detection in image processing, this issue can be well handled with the SPP, which divides the sentences into several length sections for respective pooling processing. Experiments are conducted for the task of sentence classification as well as relation classification. Experiments on Stanford Treebank, TREC, subj, and Yelp datasets demonstrate that our proposed method can outperform other state-of-the-art models, with respect to classification accuracy. Auxiliary attempts to leverage our method to SemEval-2010 Task 8 dataset further substantiate the model's capability of extracting features efficiently.

Jianchao Yang, Kai Yu, Yihong Gong and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 1794-1801, doi: 10.1109/CVPR.2009.5206757.

**Abstract:**

Recently SVMs using spatial pyramid matching (SPM) kernel have been highly successful in image classification. Despite its popularity, these nonlinear SVMs have a complexity O(n 2 ~ n 3 ) in training and O(n) in testing, where n is the training size, implying that it is nontrivial to scaleup the algorithms to handle more than thousands of training images. In this paper we develop an extension of the SPM method, by generalizing vector quantization to sparse coding followed by multi-scale spatial max pooling, and propose a linear SPM kernel based on SIFT sparse codes. This new approach remarkably reduces the complexity of SVMs to O(n) in training and a constant in testing. In a number of image categorization experiments, we find that, in terms of classification accuracy, the suggested linear SPM based on sparse coding of SIFT descriptors always significantly outperforms the linear SPM kernel on histograms, and is even better than the nonlinear SPM kernels, leading to state-of-the-art performance on several benchmarks by using a single type of descriptors.

Z. Ren, S. Gao, L. Chia and I. W. Tsang, "Region-Based Saliency Detection and Its Application in Object Recognition," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 24, no. 5, pp. 769-779, May 2014, doi: 10.1109/TCSVT.2013.2280096.

**Abstract:**

The objective of this paper is twofold. First, we introduce an effective region-based solution for saliency detection. Then, we apply the achieved saliency map to better encode the image features for solving object recognition task. To find the perceptually and semantically meaningful salient regions, we extract superpixels based on an adaptive mean shift algorithm as the basic elements for saliency detection. The saliency of each superpixel is measured by using its spatial compactness, which is calculated according to the results of Gaussian mixture model (GMM) clustering. To propagate saliency between similar clusters, we adopt a modified PageRank algorithm to refine the saliency map. Our method not only improves saliency detection through large salient region detection and noise tolerance in messy background, but also generates saliency maps with a well-defined object shape. Experimental results demonstrate the effectiveness of our method. Since the objects usually correspond to salient regions, and these regions usually play more important roles for object recognition than background, we apply our achieved saliency map for object recognition by incorporating a saliency map into sparse coding-based spatial pyramid matching (ScSPM) image representation. To learn a more discriminative codebook and better encode the features corresponding to the patches of the objects, we propose a weighted sparse coding for feature coding. Moreover, we also propose a saliency weighted max pooling to further emphasize the importance of those salient regions in feature pooling module. Experimental results on several datasets illustrate that our weighted ScSPM framework greatly outperforms ScSPM framework, and achieves excellent performance for object recognition.

* 1. R CNN and fast RCNN [10][11] (Region Convolutional Neural Network)

The standard CNN can be used to generate proposals and classify an image in diﬀerent regions. The problem is that the object that is required to be found could have a diﬀerent aspect ratio. Furthermore, it might have diﬀerent spatial locations. To solve this, RCNN can be used. Initially, the Selective Search Algorithm[9] is used to extract just region proposals. So we work with just regions instead of an infinite number of diﬀerent regions. These regions are then merged into a square. It is then fed into a CNN which generates -dimensional feature vectors. It is then fed into an SVM algorithm to find if the object is present or not. Also, an oﬀset value is taken into account to avoid cases where an object is cut into half and wrong detections are made. The oﬀset values adjust the boundary boxes of the region proposal. The disadvantage of this method is that, the Selective Search Algorithm is fixed and so no learning takes place. Now when we have image proposals classified on every object class, we bring the entire image back using greedy non-maximum suppression. Non-maximum suppression is just the process where the computer takes the intersection of union of each proposals and select the region with the higher score.

Unlike the RCNN, in the Fast RCNN, object detection is made more eﬃcient by avoiding feeding the region proposals to the CNN. Instead a convolutional feature map is obtained by feeding the image. Then ROI(Region of interest) pooling layer is used to reshape them into fixed size and then we feed them into a fully connected layer. It’s called Fast RCNN because it is not required to feed region proposal to CNN.

R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.

**Abstract:**

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012 -- achieving a mAP of 53.3%. Our approach combines two key insights: (1) one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. Since we combine region proposals with CNNs, we call our method R-CNN: Regions with CNN features. We also present experiments that provide insight into what the network learns, revealing a rich hierarchy of image features. Source code for the complete system is available at http://www.cs.berkeley.edu/~rbg/rcnn.

R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

**Abstract:**

This paper proposes a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection. Fast R-CNN builds on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy. Fast R-CNN trains the very deep VGG16 network 9x faster than R-CNN, is 213x faster at test-time, and achieves a higher mAP on PASCAL VOC 2012. Compared to SPPnet, Fast R-CNN trains VGG16 3x faster, tests 10x faster, and is more accurate. Fast R-CNN is implemented in Python and C++ (using Caffe) and is available under the open-source MIT License at https://github.com/rbgirshick/fast-rcnn.

R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 142-158, 1 Jan. 2016, doi: 10.1109/TPAMI.2015.2437384.

**Abstract:**

Object detection performance, as measured on the canonical PASCAL VOC Challenge datasets, plateaued in the final years of the competition. The best-performing methods were complex ensemble systems that typically combined multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 50 percent relative to the previous best result on VOC 2012-achieving a mAP of 62.4 percent. Our approach combines two ideas: (1) one can apply high-capacity convolutional networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data are scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly. Since we combine region proposals with CNNs, we call the resulting model an R-CNN or Region-based Convolutional Network. Source code for the complete system is available at http://www.cs.berkeley.edu/~rbg/rcnn.

Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 6154-6162, doi: 10.1109/CVPR.2018.00644.

**Abstract:**

In object detection, an intersection over union (IoU) threshold is required to define positives and negatives. An object detector, trained with low IoU threshold, e.g. 0.5, usually produces noisy detections. However, detection performance tends to degrade with increasing the IoU thresholds. Two main factors are responsible for this: 1) overfitting during training, due to exponentially vanishing positive samples, and 2) inference-time mismatch between the IoUs for which the detector is optimal and those of the input hypotheses. A multi-stage object detection architecture, the Cascade R-CNN, is proposed to address these problems. It consists of a sequence of detectors trained with increasing IoU thresholds, to be sequentially more selective against close false positives. The detectors are trained stage by stage, leveraging the observation that the output of a detector is a good distribution for training the next higher quality detector. The resampling of progressively improved hypotheses guarantees that all detectors have a positive set of examples of equivalent size, reducing the overfitting problem. The same cascade procedure is applied at inference, enabling a closer match between the hypotheses and the detector quality of each stage. A simple implementation of the Cascade R-CNN is shown to surpass all single-model object detectors on the challenging COCO dataset. Experiments also show that the Cascade R-CNN is widely applicable across detector architectures, achieving consistent gains independently of the baseline detector strength. The code is available at https://github.com/zhaoweicai/cascade-rcnn.

G. Gkioxari, R. Girshick and J. Malik, "Contextual Action Recognition with R\*CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1080-1088, doi: 10.1109/ICCV.2015.129.

**Abstract:**

There are multiple cues in an image which reveal what action a person is performing. For example, a jogger has a pose that is characteristic for jogging, but the scene (e.g. road, trail) and the presence of other joggers can be an additional source of information. In this work, we exploit the simple observation that actions are accompanied by contextual cues to build a strong action recognition system. We adapt RCNN to use more than one region for classification while still maintaining the ability to localize the action. We call our system R\*CNN. The action-specific models and the feature maps are trained jointly, allowing for action specific representations to emerge. R\*CNN achieves 90.2% mean AP on the PASAL VOC Action dataset, outperforming all other approaches in the field by a significant margin. Last, we show that R\*CNN is not limited to action recognition. In particular, R\*CNN can also be used to tackle fine-grained tasks such as attribute classification. We validate this claim by reporting state-of-the-art performance on the Berkeley Attributes of People dataset.

K. Kang, W. Ouyang, H. Li and X. Wang, "Object Detection from Video Tubelets with Convolutional Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 817-825, doi: 10.1109/CVPR.2016.95.

**Abstract:**

Deep Convolution Neural Networks (CNNs) have shown impressive performance in various vision tasks such as image classification, object detection and semantic segmentation. For object detection, particularly in still images, the performance has been significantly increased last year thanks to powerful deep networks (e.g. GoogleNet) and detection frameworks (e.g. Regions with CNN features (RCNN)). The lately introduced ImageNet [6] task on object detection from video (VID) brings the object detection task into the video domain, in which objects' locations at each frame are required to be annotated with bounding boxes. In this work, we introduce a complete framework for the VID task based on still-image object detection and general object tracking. Their relations and contributions in the VID task are thoroughly studied and evaluated. In addition, a temporal convolution network is proposed to incorporate temporal information to regularize the detection results and shows its effectiveness for the task. Code is available at https://github.com/ myfavouritekk/vdetlib.

* 1. Faster RCNN 12

Faster CNN is set out to find a way to replace the techniques of Selective Search and Edge boxes with a Dense Sampling technique like sliding windows. The objects obtained from these CNN are either squarish or rectangular. Very broadly objects can be said to be either big, small or of medium size. So proposals can be said have scales and aspect ratios. A lot of different techniques such as using Overfeat network were tried to replace the selective search but it didn’t go well. Feature pyramids were also tried but it had it’s own drawbacks. So to avoid these issues, we need to simplify the network so that it would work to give the Bounding Box proposals without using the Image Pyramid or the Feature Pyramids. It becomes especially needed to come up with a new technique in many cases as images overlap.

In this algorithm, the concept of anchor boxes is used. Usually slide from the sliding window is taken as reference. The anchor boxes of fixed size can be used as reference and regress from it. So if there is aspect ratios and scales, then there will be reference boxes. Each of these anchor boxes are used with a different BBox Regressor. Furthermore, for every BBox Regressor a classifier is added. This classifier would tell us if it’s a background or an object. This way the number of proposals are reduced from to a fewer. Also, Softmax is used after this to give a rank for every proposal based on the background. This allows us to pick the best few proposals. After this, the feature maps and use it both on RPN and Fast RCNN. This would just leave just a few regions.

The RPN would then look at a small portion. It could roughly infer the size of the object and accordingly give the proposal. Even the approximate proposal is enough as it would get refined in the faster RCNN convolutional networks.

Training loss for RPN is given by equation 10.

------------(11)

Here refers to the anchor box. The LHS part of the equation is the log loss over two classes. is the output score from the classification branch for anchor

H. Jiang and E. Learned-Miller, "Face Detection with the Faster R-CNN," 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, 2017, pp. 650-657, doi: 10.1109/FG.2017.82.

**Abstract:**

While deep learning based methods for generic object detection have improved rapidly in the last two years, most approaches to face detection are still based on the R-CNN framework [11], leading to limited accuracy and processing speed. In this paper, we investigate applying the Faster RCNN [26], which has recently demonstrated impressive results on various object detection benchmarks, to face detection. By training a Faster R-CNN model on the large scale WIDER face dataset [34], we report state-of-the-art results on the WIDER test set as well as two other widely used face detection benchmarks, FDDB and the recently released IJB-A.

G. Papandreou et al., "Towards Accurate Multi-person Pose Estimation in the Wild," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3711-3719, doi: 10.1109/CVPR.2017.395.

**Abstract:**

We propose a method for multi-person detection and 2-D pose estimation that achieves state-of-art results on the challenging COCO keypoints task. It is a simple, yet powerful, top-down approach consisting of two stages. In the first stage, we predict the location and scale of boxes which are likely to contain people, for this we use the Faster RCNN detector. In the second stage, we estimate the keypoints of the person potentially contained in each proposed bounding box. For each keypoint type we predict dense heatmaps and offsets using a fully convolutional ResNet. To combine these outputs we introduce a novel aggregation procedure to obtain highly localized keypoint predictions. We also use a novel form of keypoint-based Non-Maximum-Suppression (NMS), instead of the cruder box-level NMS, and a novel form of keypoint-based confidence score estimation, instead of box-level scoring. Trained on COCO data alone, our final system achieves average precision of 0.649 on the COCO test-dev set and the 0.643 test-standard sets, outperforming the winner of the 2016 COCO keypoints challenge and other recent state-of-art. Further, by using additional in-house labeled data we obtain an even higher average precision of 0.685 on the test-dev set and 0.673 on the test-standard set, more than 5% absolute improvement compared to the previous best performing method on the same dataset.

Y. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng and R. Sukthankar, "Rethinking the Faster R-CNN Architecture for Temporal Action Localization," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 1130-1139, doi: 10.1109/CVPR.2018.00124.

**Abstract:**

We propose TAL-Net, an improved approach to temporal action localization in video that is inspired by the Faster RCNN object detection framework. TAL-Net addresses three key shortcomings of existing approaches: (1) we improve receptive field alignment using a multi-scale architecture that can accommodate extreme variation in action durations; (2) we better exploit the temporal context of actions for both proposal generation and action classification by appropriately extending receptive fields; and (3) we explicitly consider multi-stream feature fusion and demonstrate that fusing motion late is important. We achieve state-of-the-art performance for both action proposal and localization on THUMOS'14 detection benchmark and competitive performance on ActivityNet challenge.

T. H. N. Le, Y. Zheng, C. Zhu, K. Luu and M. Savvides, "Multiple Scale Faster-RCNN

Approach to Driver’s Cell-Phone Usage and Hands on Steering Wheel Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, 2016, pp. 46-53, doi: 10.1109/CVPRW.2016.13.

**Abstract:**

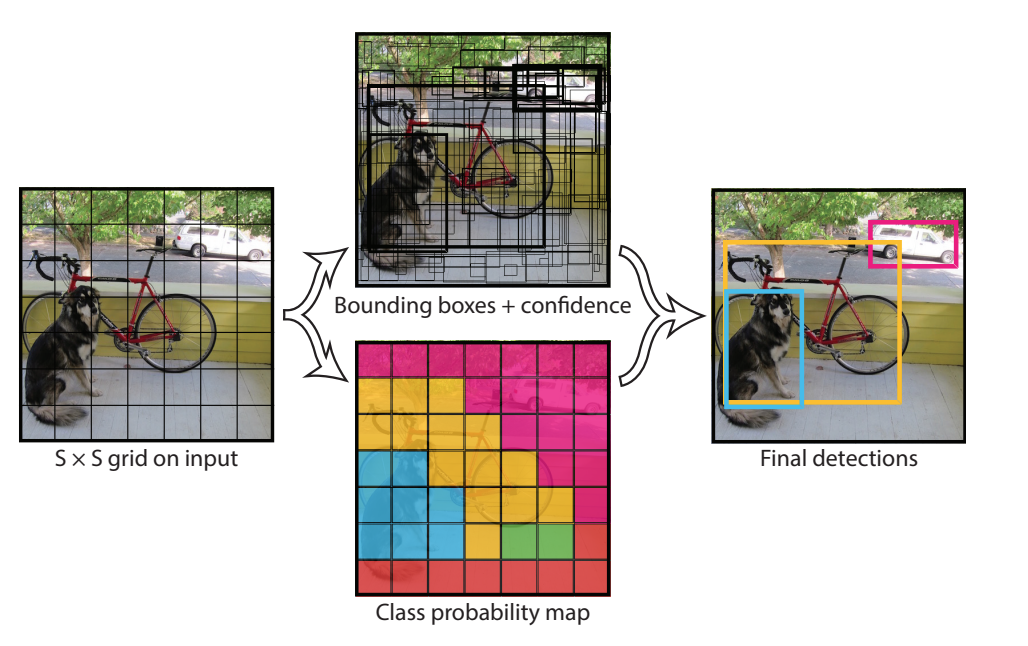
In this paper, we present an advanced deep learning based approach to automatically determine whether a driver is using a cell-phone as well as detect if his/her hands are on the steering wheel (i.e. counting the number of hands on the wheel). To robustly detect small objects such as hands, we propose Multiple Scale Faster-RCNN (MSFRCNN) approach that uses a standard Region Proposal Network (RPN) generation and incorporates feature maps from shallower convolution feature maps, i.e. conv3 and conv4, for ROI pooling. In our driver distraction detection framework, we first make use of the proposed MS-FRCNN to detect individual objects, namely, a hand, a cell-phone, and a steering wheel. Then, the geometric information is extracted to determine if a cell-phone is being used or how many hands are on the wheel. The proposed approach is demonstrated and evaluated on the Vision for Intelligent Vehicles and Applications (VIVA) Challenge database and the challenging Strategic Highway Research Program (SHRP-2) face view videos that was acquired to monitor drivers under naturalistic driving conditions. The experimental results show that our method archives better performance than Faster R-CNN on both hands on wheel detection and cell-phone usage detection while remaining at similar testing cost. Compare to the state-of-the-art cell-phone usage detection, our approach obtains higher accuracy, is less time consuming and is independent to landmarking. The groundtruth database will be publicly available.

* 1. YOLO( You Only look once ) 13

-------(12)

In the yolo algorithm, a matrix is created as above. The anchor boxes are represented by . The coordinates required for drawing the bounding boxes is represented by , , and . c1,c2 and c3 represents the classes. Usually more than one ancho box is present for the algorithm. In that case, the arguments in the target vector y is followed by the second anchor box followed by boundary boxes and classes again. The anchor boxes is used to detect more than one images in the same grid cell. If the anchor box is zero, the following boxes are given dummy values. Initially, a training set is constructed. If two anchor boxes are used, the output would be of size . Then the target vector would be . We multiply by 8 as the target vector y is of size 8. We go through the target grid cells and form the target vector . Also the class which the detected object belongs to is denoted by and the other classes are denoted by . For each of the three classes, non-max suppression is used to generate final predictions.

Among the many convolutional layers, the final layer predicts class probabilities and bounding box coordinates. A linear activation function is used for the final layer and all other layers use the following leaky rectified linear activation. This leads to specialization between the bounding box predictors. Each predictor gets better at predicting certain sizes, aspect ratios, or classes of object, improving overall recall.



* 1. Multi Scale Deformable R-CNN [14]

Most recently a novel Multi-Scaled Deformable Convolution Network Model efficiently lets the CNN improve the generalization ability of extracting image features under different geometric deformation followed by up-sampling to merge the multi-scaled feature information. It was able to achieve a better single-model performance than any other model according to PASCAL VOC dataset.

Multi Scale Deformable R-CNN adopts YOLO’s backbone network and adds the new trick in convolution operation and feature information fusion. The first backbone network is the Darknet53 network . The second element is the detection network section. The upper-level feature maps will be up-sampled and merged with the low-level layer features by the channel.

This detection network uses a deformable convolutional structure instead of an ordinary convolutional operation in order to increase the learning ability of the model with respect to object geometric deformation, as well as increasing the accuracy of object detection. This study also uses multi-scaled feature maps that combine low-level features by up-sampling to extract target object position information. This increases the ability of the model to detect small target objects and dense objects, and also greatly makes up for the defect in missing detections, which is always present in other object detection models.

J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

**Abstract:**

We present YOLO, a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance. Our unified architecture is extremely fast. Our base YOLO model processes images in real-time at 45 frames per second. A smaller version of the network, Fast YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on background. Finally, YOLO learns very general representations of objects. It outperforms other detection methods, including DPM and R-CNN, when generalizing from natural images to other domains like artwork.

J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6517-6525, doi: 10.1109/CVPR.2017.690.

**Abstract:**

We introduce YOLO9000, a state-of-the-art, real-time object detection system that can detect over 9000 object categories. First we propose various improvements to the YOLO detection method, both novel and drawn from prior work. The improved model, YOLOv2, is state-of-the-art on standard detection tasks like PASCAL VOC and COCO. Using a novel, multi-scale training method the same YOLOv2 model can run at varying sizes, offering an easy tradeoff between speed and accuracy. At 67 FPS, YOLOv2 gets 76.8 mAP on VOC 2007. At 40 FPS, YOLOv2 gets 78.6 mAP, outperforming state-of-the-art methods like Faster RCNN with ResNet and SSD while still running significantly faster. Finally we propose a method to jointly train on object detection and classification. Using this method we train YOLO9000 simultaneously on the COCO detection dataset and the ImageNet classification dataset. Our joint training allows YOLO9000 to predict detections for object classes that dont have labelled detection data. We validate our approach on the ImageNet detection task. YOLO9000 gets 19.7 mAP on the ImageNet detection validation set despite only having detection data for 44 of the 200 classes. On the 156 classes not in COCO, YOLO9000 gets 16.0 mAP. YOLO9000 predicts detections for more than 9000 different object categories, all in real-time.

R. Laroca et al., "A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-10, doi: 10.1109/IJCNN.2018.8489629.

**Abstract:**

Automatic License Plate Recognition (ALPR) has been a frequent topic of research due to many practical applications. However, many of the current solutions are still not robust in real-world situations, commonly depending on many constraints. This paper presents a robust and efficient ALPR system based on the state-of-the-art YOLO object detector. The Convolutional Neural Networks (CNNs) are trained and finetuned for each ALPR stage so that they are robust under different conditions (e.g., variations in camera, lighting, and background). Specially for character segmentation and recognition, we design a two-stage approach employing simple data augmentation tricks such as inverted License Plates (LPs) and flipped characters. The resulting ALPR approach achieved impressive results in two datasets. First, in the SSIG dataset, composed of 2,000 frames from 101 vehicle videos, our system achieved a recognition rate of 93.53% and 47 Frames Per Second (FPS), performing better than both Sighthound and OpenALPR commercial systems (89.80% and 93.03%, respectively) and considerably outperforming previous results (81.80%). Second, targeting a more realistic scenario, we introduce a larger public dataset 1 dataset, designed to ALPR. This dataset contains 150 videos and 4,500 frames captured when both camera and vehicles are moving and also contains different types of vehicles (cars, motorcycles, buses and trucks). In our proposed dataset, the trial versions of commercial systems achieved recognition rates below 70%. On the other hand, our system performed better, with recognition rate of 78.33% and 35 FPS.The UFPR-ALPR dataset is publicly available to the research community at https://web.inf.ufpr.br/vri/databases/ufpr-alpr/ subject to privacy restrictions.

L. Xie, T. Ahmad, L. Jin, Y. Liu and S. Zhang, "A New CNN-Based Method for Multi-Directional Car License Plate Detection," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 2, pp. 507-517, Feb. 2018, doi: 10.1109/TITS.2017.2784093.

**Abstract:**

This paper presents a novel convolutional neural network (CNN) -based method for high-accuracy real-time car license plate detection. Many contemporary methods for car license plate detection are reasonably effective under the specific conditions or strong assumptions only. However, they exhibit poor performance when the assessed car license plate images have a degree of rotation, as a result of manual capture by traffic police or deviation of the camera. Therefore, we propose the a CNN-based MD-YOLO framework for multi-directional car license plate detection. Using accurate rotation angle prediction and a fast intersection-over-union evaluation strategy, our proposed method can elegantly manage rotational problems in real-time scenarios. A series of experiments have been carried out to establish that the proposed method outperforms over other existing state-of-the-art methods in terms of better accuracy and lower computational cost.

D. T. Nguyen, T. N. Nguyen, H. Kim and H. Lee, "A High-Throughput and Power-Efficient FPGA Implementation of YOLO CNN for Object Detection," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 8, pp. 1861-1873, Aug. 2019, doi: 10.1109/TVLSI.2019.2905242.

**Abstract:**

Convolutional neural networks (CNNs) require numerous computations and external memory accesses. Frequent accesses to off-chip memory cause slow processing and large power dissipation. For real-time object detection with high throughput and power efficiency, this paper presents a Tera-OPS streaming hardware accelerator implementing a you-only-look-once (YOLO) CNN. The parameters of the YOLO CNN are retrained and quantized with the PASCAL VOC data set using binary weight and flexible low-bit activation. The binary weight enables storing the entire network model in block RAMs of a field-programmable gate array (FPGA) to reduce off-chip accesses aggressively and, thereby, achieve significant performance enhancement. In the proposed design, all convolutional layers are fully pipelined for enhanced hardware utilization. The input image is delivered to the accelerator line-by-line. Similarly, the output from the previous layer is transmitted to the next layer line-by-line. The intermediate data are fully reused across layers, thereby eliminating external memory accesses. The decreased dynamic random access memory (DRAM) accesses reduce DRAM power consumption. Furthermore, as the convolutional layers are fully parameterized, it is easy to scale up the network. In this streaming design, each convolution layer is mapped to a dedicated hardware block. Therefore, it outperforms the “one-size-fits-all” designs in both performance and power efficiency. This CNN implemented using VC707 FPGA achieves a throughput of 1.877 tera operations per second (TOPS) at 200 MHz with batch processing while consuming 18.29 W of on-chip power, which shows the best power efficiency compared with the previous research. As for object detection accuracy, it achieves a mean average precision (mAP) of 64.16% for the PASCAL VOC 2007 data set that is only 2.63% lower than the mAP of the same YOLO network with full precision.

1. Conclusion

We have discussed various algorithms of Object detection starting from SVM and Ending at YOLO and Multi deformable RCNN. We bagan by discussing SVM individually. Then we slowly built up chronologically the evolution of object detection. Histogram of gradients were used along with SVM followed by depormable parts model to take care of the change in poses and posture of the object. Then we discussed about CNNs and Overfeat. Then SPP which is a more advanced form of bag of visual words were discussed. Then we reviewed RCNN and faster RCNN. Finally we reviewed Yolo as well as Multi scale deformable RCNN.

1. References
   * 1. *Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7\_12.*
     2. *T. Surasak, I. Takahiro, C. Cheng, C. Wang and P. Sheng, "Histogram of oriented gradients for human detection in video," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 172-176, doi: 10.1109/ICBIR.2018.8391187.*
     3. *N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.*
     4. *P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627-1645, Sept. 2010, doi: 10.1109/TPAMI.2009.167.*
     5. *O'Shea, Keiron & Nash, Ryan. (2015). An Introduction to Convolutional Neural Networks. ArXiv e-prints.*
     6. *Sermanet, Pierre & Eigen, David & Zhang, Xiang & Mathieu, Michael & Fergus, Rob & Lecun, Yann. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. International Conference on Learning Representations (ICLR) (Banff).*
     7. S. Lazebnik, C. Schmid and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, 2006, pp. 2169-2178, doi: 10.1109/CVPR.2006.68.
     8. *K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.*
     9. *K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers and A. W. M. Smeulders, "Segmentation as selective search for object recognition," 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 1879-1886, doi: 10.1109/ICCV.2011.6126456.*
     10. *R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.*
     11. *R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.*
     12. *S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.*
     13. *J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.*
     14. *Cao, Danyang & Chen, Zhixin & Gao, Lei. (2020). An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks. Human-centric Computing and Information Sciences. 10. 14. 10.1186/s13673-020-00219-91*
     15. *S. Cao, Y. Liu, W. Song, Z. Cui, X. Lv and J. Wan, "Toward Human-in-the-Loop Prohibited Item Detection in X-ray Baggage Images," 2019 Chinese Automation Congress (CAC), Hangzhou, China, 2019, pp. 4360-4364, doi: 10.1109/CAC48633.2019.8996933.*
     16. *S. Lahiri, N. Jyoti, S. Pyati and J. Dewan, "Abnormal Crowd Behavior Detection Using Image Processing," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697356.*
     17. *J. Molineros, R. Behringer and C. Tam, "Vision-based augmented reality for pilot guidance in airport runways and taxiways," Third IEEE and ACM International Symposium on Mixed and Augmented Reality, Arlington, VA, USA, 2004, pp. 302-303, doi: 10.1109/ISMAR.2004.66.*
     18. *Cusano, Claudio & Napoletano, Paolo. (2017). Visual recognition of aircraft mechanical parts for smart maintenance. Computers in Industry. 86. 26-33. 10.1016/j.compind.2017.01*