Object Detection and Handwriting Recognition System (FLEX LENS): Review Paper

Rohan Ghosh (3180700554102)

The program, Flex Lens is based on Deep Learning. We have introduced the concept of Convolutional Neural Network (CNN) to create an AI system that can detect handwriting and visualize an object, as it automatically detects important features without human monitoring and has very high accuracy in image recognition problems. Using Jupyter Notebook we recorded our data sets and used Tensorflow to store image data and object discovery.

Our system is able to identify objects based on previously saved image sets. In addition, it is capable of capturing, storing and setting new photo sets. As such, it is a self-study program that will gather information for selfimprovement.

The aim of this project is to identify handwritten and digital objects for use for a variety of purposes in this rapidly evolving environment, where Metaverse is a common expression. And the acquisition of an object and recognition for immediate processing of intermediate operations. It has a comprehensive program for security cameras, to improve programs for people with physical disabilities, etc.

Overall, the “Flex Lens” is a clever, fast and efficient system for flexibility in our daily lives. So, making the world a better place.

I. INTRODUCTION

In order to gain a complete understanding of the image, we must not only focus on the differences between the two images, but also try to accurately measure the concepts and locations of the objects in each image. This function is called object detection, which usually involves various sub-functions such as facial detection, pedestrian detection and skeletal detection. As one of the basic problems of computer perception, object detection is able to provide valuable information for the semantic understanding of images and videos. At the same time, inheriting neural networks and related learning programs, progress in these areas will improve neural network algorithms, and will also have a significant impact on discovery strategies that can be considered learning programs. However, due to the wide variety of views, shapes, closures and lighting conditions, it is difficult to achieve a complete acquisition of an object with additional local performance. A lot of attention has been drawn to this field in recent years.

The definition of an acquisition problem is to determine where the objects are located in a given image (local placement of the object) and what category each item is (item classification). The pipeline models of common acquisition models can therefore be divided mainly into three categories: informed regional selection, feature extraction and classification.

Selecting an informed region. Since different objects may appear in any area of the image and have different aspect ratio or sizes, it is a natural decision to scan the entire image with a slide window that has multiple dimensions. While this complete strategy can cover all possible areas of material, its shortcomings are also evident. Due to the large number of candidate windows, it is computerized and produces a lot of non-functional windows. However, if only a fixed number of sliding window templates are used, unsatisfactory circuits can be generated.

Feature domain. In order to see different things, we need to extract visual elements that can provide semantic and solid representation. However, due to the variety of appearance, lighting conditions and backgrounds, it is difficult to design a solid feature dictionary to accurately describe all types of objects.

Separation. In addition, classification is required to separate the target object into all other categories and to make presentations segmented, semantic and teach visual recognition.

And the discriminatory study of image models allows for the creation of models that are based on an accurate component in a variety of object classes.

Based on these descriptive feature descriptions and shallow readable structures, state-of-the-art results were obtained from the PASCAL VOC acquisition competition and real-time embedded systems were acquired at low load. However, small gains are made in 2010-2012 by building only integration systems and using a slightly different alternative to successful methods. This fact is due to the following reasons: 1) The production of candidate binding boxes with a sliding window design is empty, inefficient and inaccurate. 2) The semantic gap cannot be closed by combining low-level engineering definitions with shallowly moderately trained models.

Due to the emergency of Deep Neural Networks (DNNs), the most significant benefit is gained through the introduction of CNN (R-CNN) regions. DNNs, or more representative CNNs, operate in a very different way from traditional methods. They have deep structures with the ability to learn complex features rather than small ones. And expressivity and robust training algorithms allow for learning informative presentation without the need to design features in person.

Since the launch of R-CNN, a number of advanced models have been proposed, including Fast R-CNN which integrates segmentation and box retrieval functions, Faster R-CNN taking another sub-network to produce regional and YOLO proposals. which achieves the acquisition of an object by a fixed grid rotation. They all bring different degrees of acquisition performance enhancement over the main R-CNN and make real-time and accurate acquisitions more realistic.

In this paper, systematic reviews are provided on models representing the various abbreviations and characteristics in a few application domains, including common object discovery, face detection and pedestrian detection. Their relationship is illustrated in Figure 1. Based on CNN's basic architecture, the acquisition of a normal object is achieved by a constraint of a limited box, while the acquisition of the object is achieved by the enhancement of spatial brightness and pixel level separation.

Face detection and pedestrian detection are closely related to the acquisition of a common object and are achieved primarily by multi-dimensional transformation and a multi-faceted / motivating forest, respectively. Dotted lines indicate that corresponding domains are associated with each other under certain conditions. It should be noted that the covered domains are different. Pedestrian and face photographs have common structures, while common objects and scenes have a complex variety of geometric structures and structures.
Therefore, different deep models are required for different images.

There has been a well-established pioneer effort that focuses on the right software tools to apply in-depth learning techniques for image editing and object acquisition, but he pays little attention to defining certain algorithms. In contrast, our work not only reviews in-depth learning-based acquisition models and regulations that cover different application domains in detail, but also provides their corresponding comparative evaluation and rational analysis.

II. Overview of Deep Learning

Before we look at all about in-depth learningbased approaches, we provide an overview of in-depth reading history and an introduction to the basic structures and benefits of CNN.

A. History: Birth, Decline and Prosperity:

Deep models can be called neural networks with deep structures. The history of neural networks can be traced back to the 1940s, and the original purpose was to mimic the human brain system in order to solve common learning problems in a legal way. However, due to overcrowding of training, lack of large training data, limited calculation power and insignificant performance compared to other machine learning tools, neural networks became obsolete in the early 2000s.

In-depth reading has been popular since 2006 with a successful speech recognition. The restoration of deep learning can be triggered by the following factors.

The emergence of high-level training data, such as ImageNet, to fully reflect its enormous capacity for learning;

Rapid development of highly compatible compatible computer systems, such as GPU clusters;

Significant improvements in network infrastructure and training strategies. With pre-supervised and horizontally preprogrammed training guided by AutoEncoder (AE) or Boltzmann Limited Machine (RBM), a good startup is provided. With the cessation and addition of data, the problem of overuse in training has been relieved and further explored to improve performance.

What makes in-depth learning a major impact on the entire academic community? It may be due to the role of the Hinton team, whose ongoing efforts have shown that in-depth learning will bring dynamic success to larger challenges rather than just obvious advances in small data sets.

B. Architecture and Advantages of CNN

CNN is the most representative model for indepth learning [26]. CNN's standard format, called VGG16 can be found. Each CNN layer is known as a feature map. The input element feature map is a 3D pixel density matrix for different color channels (e.g., RGB). A feature map of any interior layer is a multichannel image, the ‘pixel’ of which can be viewed as a specific feature. Every neuron is connected to a small portion of nearby neurons from the previous layer (receiving field). Different types of changes can be made to feature maps, such as filtering and merging. The filtering function (convolution) combines the filter matrix (studied weights) with the reception field values of the neurons and takes non-linear function (such as sigmoid, ReLU) to obtain final responses. Combination functionality, such as high integration, intermediate integration, L2 integration and general field comparisons, summarizes the reception field responses into a single value in order to produce descriptions of the strongest factor.

By distinguishing between merging and merging, the first element of the feature is constructed, which can be processed in a controlled manner by adding a few fully integrated layers (FC) layers to suit a variety of viewing functions. According to the functions involved, a final layer with different opening functions is added to detect specific conditional opportunities for each outgoing neuron. And the whole network can be upgraded to objective function (e.g., mean square error or cross-entropy loss) in the form of stochastic gradient descent (SGD). The standard VGG16 has 13 complete convolutional (conv) layers, 3 fully integrated layers, 3 layers of high integration and SoftMax split layer. Con feature featuremaps are generated by merging 3 * 3 filter windows, and the map aspect adjustment is minimized by 2 stride max merge layers. An unofficial test image with the same size and training samples can be processed with a trained network. Redesign or cropping functions may be required if different sizes are provided.

CNN's advantages against traditional methods can be summarized as follows.

Hierarchical feature presentation, which is a multi-level presentation from pixel to highlevel semantic features studied by multiphase formatting, can be read automatically and hidden input data features can be categorized using a multi-level indirect map.

Compared to shallow traditional models, the deep architecture offers a remarkable presentation.

The structure of CNN provides the opportunity to collaborate on a number of related tasks (e.g., Fast R- CNN incorporates the division and arrangement of a binding box into a multi-functional approach).

Benefiting from CNN's great reading capacity, some old computer vision challenges can be repeated as high-level data transforms problems and is solved with a different perspective.

III. OBJECT DETECTION

Normal object detection is intended to detect and separate objects present in any single image, as well as to label them with rectangular boxes to indicate the reliability of the presence. The frameworks of standard acquisition methods can be divided mainly into two types. One follows a common procurement pipeline, produces regional proposals initially and divides each proposal into different categories of items. One views the acquisition of an object as a retrospective or separation problem, which takes a combined framework to achieve the final results (categories and locations) directly.

A. *Region Proposal Based Framework*

The framework based on a regional application, a two-step process, is similar to the degree of attention of the human brain to some degree, which provides a solid scan of the whole situation first and then focuses on the regions in which you are interested. Among the previously related works, the most representative of Overbeat. This model puts CNN in the path of the sliding window, which predicts binding boxes directly from the top-level map areas after the reliable discovery of the root object categories.

B. *Regression/Classification Based Framework*

The frameworks based on the regional proposals are made up of a number of related components, which include the production of the regional proposal, the feature release by CNN, the classification and the reversal of the binding box, which is usually trained separately. Even in the latest end-to-end module Faster R-CNN, some training is still needed to obtain the parameters for sharing between RPN and network acquisition. Therefore, the time spent on handling different components becomes bottled in real-time use.

Step-by-step steps based on landscaping / categories, mapping directly from pixels to links to binding boxes and class opportunities, can reduce time costs.

Combined losses are introduced in bias in both localization and multi-component reliability to predict link links that include anonymous categories. However, a large number of additional parameters are included in the final layer.

C. *Experimental Evaluation*

We compare various acquisition methods on three benchmark datasets, including PASCAL VOC 2007, PASCAL VOC 2012 and Microsoft COCO.

In the absence of specific instructions for the accepted framework provided, the model used is VGG16 pre-trained in the ImageNet 1000 division function. Due to the paper length limit, we only offer a comprehensive overview, including proposal, learning method, job loss, programming language and forum, for outstanding structures.

Properly integrated, CNN models are more robust that can enhance object acquisition function (comparisons between R-CNN and AlexNet, R-CNN and VGG16 and SPP-net and ZF-Net).

With the introduction of the SPP (SPP-net), end-to-end multi-task architecture (FRCN) and RPN (Faster R- CNN), acquisition performance is gradually and transparently improved.

Due to the large number of trained parameters, in order to obtain strong multilevel features, data enhancement is very important for in-depth learning models (Faster R-CNN with '07 ', '07 + 12' and '07 + 12 + coco ') .

In addition to the basic models, there are still many other factors that affect the performance of an object acquisition, such as extracting a multi-dimensional and multilocation feature (e.g. MR-CNN), modified segmentation networks (e.g. NOC), additional information from other related functions (e.g. StuffNet, HyperNet), multidimensional representation (e.g. ION) and the extraction of solid negative samples (e.g.
OHEM).

As YOLO does not have the capacity to produce high-end local IoU products, it gets a very negative effect on VOC 2012. However, with relevant information from Fast R-CNN (YOLO + FRCN) and the help of other strategies, such as anchor boxes. , BN and well-analyzed features, local performance errors are corrected (YOLOv2).

By combining many of the latest strategies and modeling the entire network as a complete transformation, R-FCN achieves the most obvious improvement in performance acquisition than alternatives.

IV. SALIENT OBJECT DETECTION

Acquisition of visual aesthetics, one of the most important and challenging tasks in computer vision, is aimed at highlighting the regions of the objects in the picture. Many applications include visual acuity to enhance their functionality, such as image capture and segmentation, image retrieval and object recovery.

In general, there are two branches of methods in finding the most important thing, namely bottom-up (BU) and top-down (TD).

Local element variability plays a central role in the acquisition of an important BU object, regardless of the semantic content of the scene. To learn the brightness of a local element, various local and international elements are extracted in pixels, e.g. edges, location information. However, advanced semantic knowledge and multiple scales cannot be considered for these low-level features. As a result, maps are available that show lower variables instead of highlights. The discovery of an important TD object tends to work and takes prior knowledge about the categories of the object to direct the production of important maps. If we take the semantic segmentation as an example, a saliency map is generated in phase to distribute pixels into categories of an object in the form of TD. In a nutshell, TD saliency can be viewed as a focus, which targets key BU points that may be part of the object.

TABLE I

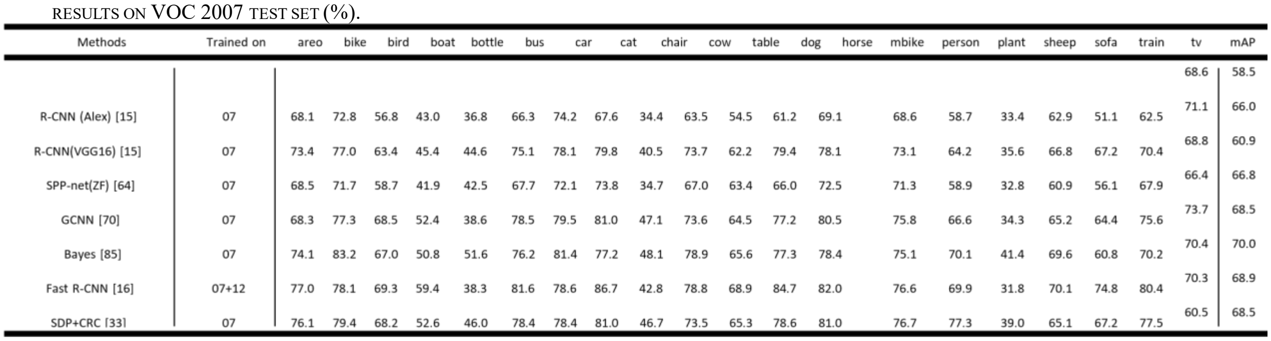
AN OVERVIEW OF PROMINENT GENERIC OBJECT DETECTION ARCHITECTURES

Framework	Proposal	Multi-scale Input	Learning Method	Loss Function	Softmax Layer	End-to-end Train	Platform	Language
R-CNN [15]	Selective Search	-	SGD,BP	Hinge loss (classification),Bounding box regression	+	-	Caffe	Matlab
SPP-net [64]	EdgeBoxes	+	SGD	Hinge loss (classification),Bounding box regression	+	-	Caffe	Matlab
Fast RCNN [16]	Selective Search	+	SGD	Class Log loss+bounding box regression	+	-	Caffe	Python
Faster R-CNN [18]	RPN	+	SGD	Class Log loss+bounding box regression	+	+	Caffe	Python/Matlab
R-FCN [65]	RPN	+	SGD	Class Log loss+bounding box regression	-	+	Caffe	Matlab
Mask R-CNN [67]	RPN	+	SGD	Class Log loss+bounding box regression +Semantic sigmoid loss	+	+	TensorFlow/Keras	Python
FPN [66]	RPN	+	Synchronized SGD	Class Log loss+bounding box regression	+	+	TensorFlow	Python C
YOLO [17]	-	-	SGD	Class sum-squared error loss+bounding box regression +object confidence+background confidence	+	+	Darknet	C++
SSD [71]	-	-	SGD	Class softmax loss+bounding box regression	-	+	Caffe	C
YOLOv2 [72]	-	-	SGD	Class sum-squared error loss+bounding box regression +object confidence+background confidence	+	+	Darknet	

TABLE II

COMPARATIVE

RESULTS ON VOC 2007 TEST SET (%).



Methods	Trained on	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
																					68.6	58.5
R-CNN (Alex) [15]	07	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	71.1	66.0
R-CNN(VGG16) [15]	07	73.4	77.0	63.4	45.4	44.6	75.1	78.1	79.8	40.5	73.7	62.2	79.4	78.1	73.1	64.2	35.6	66.8	67.2	70.4	68.8	60.9
SPP-net(ZF) [64]	07	68.5	71.7	58.7	41.9	42.5	67.7	72.1	73.8	34.7	67.0	63.4	66.0	72.5	71.3	58.9	32.8	60.9	56.1	67.9	66.4	66.8
GCNN [70]	07	68.3	77.3	68.5	52.4	38.6	78.5	79.5	81.0	47.1	73.6	64.5	77.2	80.5	75.8	66.6	34.3	65.2	64.4	75.6	73.7	68.5
Bayes [85]	07	74.1	83.2	67.0	50.8	51.6	76.2	81.4	77.2	48.1	78.9	65.6	77.3	78.4	75.1	70.1	41.4	69.6	60.8	70.2	70.4	70.0
Fast R-CNN [16]	07+12	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.3	68.9
SDP+CRC [33]	07	76.1	79.4	68.2	52.6	46.0	78.4	78.4	81.0	46.7	73.5	65.3	78.6	81.0	76.7	77.3	39.0	65.1	67.2	77.5	60.5	68.5

v. FACE DETECTION

Face detection is important for many facial systems and serves as an important pre-screening process for facial recognition, facial expressions, and facial expressions. Unlike the common object detection, this function is to detect and detect facial circuits that cover the width of the scales (30-300 pts vs. 10-1000 pts). At the same time, the face has the structure of a different object structure (e.g., distribution of different facial features) and features (e.g. skin color). All of these differences lead to special attention in this work. However, large variations in facial expressions, such as closure, cause variability and changes in light, posing major challenges to this function in real systems.

The most popular face detector proposed by Viola and information fusion. Due to the correlations between Jones trains class dividers with Haar-Like features and different tasks within and outside object detection, AdaBoost, which achieves real-time performance. multi-task joint optimization has already been studied However, this scanner can significantly reduce the by many researchers. However, apart from the tasks level on real-world systems due to the greater visual mentioned in Subs. III-A8, it is desirable to think over variability of human faces. In contrast to this cascade the characteristics of different sub-tasks of object structure, Felzenszwalb et al. proposed a partially detection (e.g., super pixel semantic segmentation in salient object detection) and ex- tend multi-task

disabled partial (DPM) model for facial detection. optimization to other applications such as instance However, in these standard face recognition methods, segmentation, multi-object tracking and multi-person a high cost of calculation and a large number of pose estimation. Besides, given a specific application, annotations are required to achieve the optimal result. the information from different modalities, such as text, Apart from that, their performance is severely limited thermal data and images, can be fused together to by hand-crafted features and shallow structures. achieve a more discriminant network.

Scale adaption. Objects usually exist in different

VI. PEDESTRIAN DETECTION

scales, which is more apparent in face detection and pedestrian detection. To increase the robustness to

Recently, pedestrian detection has been extensively scale changes, it is demanded to train scale-invariant, investigated, closely related to pedestrian tracking re- multi-scale or scale- adaptive detectors. For scaleidentification, as well as robotic navigation. Prior to invariant detectors, more powerful backbone the recent advances in DCNN-based approaches, some architectures (e.g., ResNext [123]), negative sample researchers have combined advanced forestry with mining reverse connection and sub- category hand-made features to find pedestrians. At the same modelling are all beneficial. For multi-scale detectors, time, making a clear model of flexibility and closure, both the FPN [66] which produces multi-scale feature component-based models and clear closing grip are maps and Generative Adversarial Network which exhausting. narrows

representation differences between small

As there are many cases of small-sized pedestrians in objects and the large ones with a low-cost architecture normal pedestrian acquisition situations (e.g., provide insights into generating meaningful feature automatic driving and smart surveillance), the use of pyramid. For scale-adaptive detectors, it is useful to an integrated

RoI layer in a standard acquisition combine knowledge graph, attentional mechanism, cascade network and scale distribution estimation [to
pipeline may result in ‘empty’ features due to drum collapse. At the moment, the main source of false detect objects adaptively.

speculation in pedestrian discovery is the confusion of Spatial correlations and contextual modelling. Spatial

solid back conditions, as opposed to distractions from distribution plays an important role in object detection. many stages in the discovery of a common object. As So, region proposal generation and grid regression are a result, different configurations and components are taken to obtain probable object locations. However, required to achieve accurate pedestrian detection. the correlations between multiple proposals and object categories are ignored. Besides, the global structure

VII. **PROMISING FUTURE** information is abandoned by the positionscore maps in R-FCN. To solve these problems, we -sensitive

DIRECTIONS AND TASKS can refer to diverse subset selection and sequential reasoning tasks for possible solutions. It is also In spite of rapid development and achieved promising meaningful to mask salient parts and couple them with progress of object detection, there are still many open the global structure in a joint-learning manner.

issues for future work. The second one is to release the burden on manual The first one is small object detection such as labor and accomplish real-time object detection, with occurring in COCO dataset and in face detection task. the emergence of large-scale image and video data. To improve localization accuracy on small objects The following three aspects can be taken into account. under partial occlusions, it is necessary to modify Cascade network. In a cascade network, a cascade of network architectures from the following aspects. detectors is built in different stages or layers. And Multi-task joint optimization and multi-modal easily distinguishable examples are rejected at shallow

layers so that features and classifiers at latter stages can handle more difficult samples with the aid of the decisions from previous stages. However, current cascades are built in a greedy manner, where previous stages in cascade are fixed when training a new stage. So, the optimizations of different CNNs are isolated, which stresses the necessity of end-to- end optimization for CNN cascade. At the same time, it is also a matter of concern to build contextual associated cascade networks with existing layers. Unsupervised and weakly supervised learning. It’s very time consuming to manually draw large quantities of bounding boxes. To release this burden, semantic prior unsupervised object discovery, multiple instances learning and deep neural network prediction can be integrated to make best use of imagelevel supervision to assign object category tags to corresponding object regions and refine object boundaries. Furthermore, weakly annotations (e.g., center-click annotations) are also helpful for achieving high-quality detectors with modest annotation efforts, especially aided by the mobile platform.

VIII. CONCLUSION

Because of its powerful reading ability and the benefits of dealing with closure, scale conversion and back-to-back switches, the acquisition of in-depth learning-based material has become a research center in recent years. This paper provides a detailed review of deep learning-based acquisition frameworks that address a variety of minor issues, such as closure, clutter and low resolution, with varying degrees of conversion on R-CNN. The review begins with a standard acquisition pipeline that provides basic structures for other related tasks. Then, three other common tasks, namely visual acuity, facial recognition, and pedestrian detection, are also briefly reviewed. Finally, we suggest a few promising directions for the future to gain a complete understanding of the acquisition area. This review is also useful for the development of neural networks and related learning programs, which provide important information and guidelines for future progress.

REFERENCES

[1] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, p. 1627, 2010.

[2] K. K. Sung and T. Poggio, “Example-based learning for view-based human face detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, 2002.

[3] C. Wojek, P. Dollar, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, p. 743, 2012.

[4] H. Kobatake and Y. Yoshinaga, “Detection of spicules on mammogram based on skeleton analysis.” *IEEE Trans. Med. Imag.*, vol. 15, no. 3, pp. 235–245, 1996.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *ACM MM*, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017.

[8] Z. Yang and R. Nevatia, “A multi-scale cascade fully convolutional network face detector,” in *ICPR*, 2016.

[9] C. Chen, A. Seff, A. L. Kornhauser, and J. Xiao, “Deepdriving: Learning affordance for direct perception in autonomous driving,” in *ICCV*, 2015.

[10] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3d object detection network for autonomous driving,” in *CVPR*, 2017.

[1] D. Chen, G. Hua, F. Wen, and J. Sun, “Supervised transformer network for efficient face detection,” in *ECCV*, 2016.

[2] D. Ribeiro, A. Mateus, J. C. Nascimento, and P. Miraldo, “A real-time pedestrian detector using deep learning for human-aware navigation,” *arXiv:1607.04441*, 2016.

[3] F. Yang, W. Choi, and Y. Lin, “Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers,” in *CVPR*, 2016.

[4] P. Druzhkov and V. Kustikova, “A survey of deep learning methods and software tools for image classification and object detection,” *Pattern Recognition and Image Anal.*, vol. 26, no. 1, p. 9, 2016.

[5] W. Pitts and W. S. McCulloch, “How we know universals the perception of auditory and visual forms,” *The Bulletin of Mathematical Biophysics*, vol. 9, no. 3, pp. 127–147, 1947.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representation by back-propagation of errors,” *Nature*, vol. 323, no. 323, pp. 533–536, 1986.

[11] A. Dundar, J. Jin, B. Martini, and E. Culurciello, “Embedded streaming deep neural networks accelerator with applications,” *IEEE Trans. Neural Netw. & Learning Syst.*, vol. 28, no. 7, pp. 1572–1583, 2017.

[12] R. J. Cintra, S. Duffner, C. Garcia, and A. Leite, “Low-complexity approximate convolutional neural networks,” *IEEE Trans. Neural Netw. & Learning Syst.*, vol. PP, no. 99, pp. 1–12, 2018.

[13] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, “Cost-sensitive learning of deep feature representations from imbal- anced data.” *IEEE Trans. Neural Netw. & Learning Syst.*, vol. PP, no. 99, pp. 1–15, 2017.

[14] A. Stuhlsatz, J. Lippel, and T. Zielke, “Feature extraction with deep neural networks by a generalized discriminant analysis.” *IEEE Trans. Neural Netw. & Learning Syst.*, vol. 23, no. 4, pp. 596–608, 2012.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *CVPR*, 2014.

[16] R. Girshick, “Fast r-cnn,” in *ICCV*, 2015.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *CVPR*, 2016.

[18] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real- time object detection with region proposal networks,” in *NIPS*, 2015, pp. 91–99.

[19] D. G. Lowe, “Distinctive image features from scale-invariant key- points,” *Int. J. of Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[20] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*, 2005.

[21] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” in *ICIP*, 2002.

[22] C. Cortes and V. Vapnik, “Support vector machine,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] Y. Freund and R. E. Schapire, “A desicion-theoretic generalization of online learning and an application to boosting,” *J. of Comput. & Sys. Sci.*, vol. 13, no. 5, pp. 663–671, 1997.

[24] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1627–1645, 2010.

[25] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zis- serman, “The pascal visual object classes challenge 2007 (voc 2007) results (2007),” 2008.

[26] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[27] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, “Predicting eye fixations using convolutional neural networks,” in *CVPR*, 2015.

[28] E. Vig, M. Dorr, and D. Cox, “Large-scale optimization of hierarchical features for saliency prediction in natural images,” in *CVPR*, 2014.

[29] H. Jiang and E. Learned-Miller, “Face detection with the faster r-cnn,” in *FG*, 2017.

[30] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, “Joint cascade face detection and alignment,” in *ECCV*, 2014.

[7] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Sci.*, vol. 313, pp. 504–507, 2006.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009.

[10] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, “Binary coding of speech spectrograms using a deep auto- encoder,” in *INTERSPEECH*, 2010.

[11] G. Dahl, A.-r. Mohamed, G. E. Hinton *et al.*, "Phone recognition with the mean-covariance restricted boltzmann machine," in *NIPS*, 2010.

[12] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co- adaptation of feature detectors," *arXiv:1207.0580*, 2012.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[14] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv:1312.6229*, 2013.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[18] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[19] M. Oquab, L. Bottou, I. Laptev, J. Sivic *et al.*, "Weakly supervised object recognition with convolutional neural networks," in *NIPS*, 2014.

[20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*, 2014.

[21] F. M. Wadley, "Probit analysis: a statistical treatment of the sigmoid response curve," *Annals of the Entomological Soc. of America*, vol. 67, no. 4, pp. 549–553, 1947.

[22] K. Kavukcuoglu, R. Fergus, Y. LeCun *et al.*, "Learning invariant features through topographic filter maps," in *CVPR*, 2009.

[23] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. LeCun, "Learning convolutional feature hierarchies for visual recognition," in *NIPS*, 2010.

[24] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolu- tional networks," in *CVPR*, 2010.

[25] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.

[26] Z.-Q. Zhao, B.-J. Xie, Y.-m. Cheung, and X. Wu, "Plant leaf iden- tification via a growing convolution neural network with progressive sample learning," in *ACCV*, 2014.

[27] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *ECCV*, 2014.

[28] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *ACM MM*, 2014.

[29] D. Tome`, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro, "Deep convolutional neural networks for pedestrian detec- tion," *Signal Process.: Image Commun.*, vol. 47, pp. 482–489, 2016

[30] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategoryaware convolutional neural networks for object proposals and detection," in *WACV*, 2017.

[31] Z.-Q. Zhao, H. Bian, D. Hu, W. Cheng, and H. Glotin, "Pedestrian detection based on fast r-cnn and batch normalization," in *ICIC*, 2017.

[32] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[33] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial- temporal clues in a hybrid deep learning framework for video classifi- cation," in *ACM MM*, 2015.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[35] Y. Li, K. He, J. Sun *et al.*, "R-fcn: Object detection via region-based fully convolutional networks," in *NIPS*, 2016, pp. 379–387.

[36] T.-Y. Lin, P. Dolla´r, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[37] K. He, G. Gkioxari, P. Dolla´r, and R. B. Girshick, "Mask r-cnn," in *ICCV*, 2017.

[38] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *CVPR*, 2014.

[39] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. So Kweon, "Attentionnet: Aggregating weak directions for accurate object detection," in *CVPR*, 2015.

[40] M. Najibi, M. Rastegari, and L. S. Davis, "G-cnn: an iterative grid based object detector," in *CVPR*, 2016.

[41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016.

[42] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv:1612.08242*, 2016.

[43] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," *arXiv:1701.06659*, 2017.

[44] Z. Shen, Z. Liu, J. Li, Y. G. Jiang, Y. Chen, and X. Xue, "Dsod: Learning deeply supervised object detectors from scratch," in *ICCV*, 2017.

[45] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto- encoders," in *ICANN*, 2011.

[46] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus, "Learning invariance through imitation," in *CVPR*, 2011.

[47] X. Ren and D. Ramanan, "Histograms of sparse codes for