Application of Deep Learning to Computer Vision: A study

Avinash H. Hedaoo
Dept of Computer Science
Prerna College of Commerce, India

Abstract- *Gradually traditional machine learning algorithms are being replaced with deep learning as it exhibits strong advantages in the feature extraction, thus been widely used in the field of computer vision and among others. The usability of computer vision is everywhere, whereas deep learning revolutionized the concept of artificial intelligence including computer vision. Deep Learning has pushed the limits of what was possible in the domain of Digital Image Processing. However, that is not to say that the traditional computer vision techniques which had been undergoing progressive development in years prior to the rise of DL have become obsolete. This paper first reviews the main ideas of deep learning, and displays several related frequently-used algorithms for computer vision. Afterwards, the current research status of computer vision field is demonstrated in this paper, particularly the main applications of deep learning in the research field. Various types of deep learning algorithms have been described. In this paper, our focus is on CV. Author provides a critical review of recent achievements in terms of techniques and applications. The summarizations, knowledge accumulations, and creations could benefit researchers in the academia and participators in the CV industries.*

Keywords- CNN, Study, Deep learning, VGG, Computer Vision, Machine Learning encoders, Computer vision Literature review

I. INTRODUCTION

In recent years, deep learning [2] has become popular in the field of machine learning and computer vision. In the field of object detection [3], text classification [4], image classification [5], face verification [6], gender classification [7], scene-classification [8], digits and traffic signs recognition [9], etc., many deep learning models achieve high performance by using large architectures with numerous features. Some of the available deep learning models are; AlexNet model [5], VGG S model [10], Berkeley-trained model [11], Places-CNN model [8], Fully Convolutional Semantic Segmentation Model (FCN-Xs) [12], CNN Models for Salient Object Subitizing [13], Places CNDS models on scene recognition [14], Models for age and gender classification [15], GoogLeNet model [16], etc. All these models tried to optimize issues like preventing from over-fitting, connection of nodes between adjacent layers, large learning capacity, etc. The factors need to be taken care of while working with deep learning network, such as availability of large training sets, powerful GPU for training and testing, better model regularization strategies, the amount of training time that we can tolerate etc.[1].

Deep learning is the sub-fields of machine learning, where learning happens with high-level data using hierarchical structures. Using advanced machine learning algorithms it improves chip programming abilities on low cost computing hardware. In recent years lot of research has been done to improve deep learning algorithms. It is found that deep learning algorithms are superior to numerous other state-of-the- art schemes. Despite of several successful attempts, deep learning still remains nascent field. Keeping it in mind, this paper surveys the recent advances in deep learning and the application of these algorithms in the field of computer vision [26].

Since being reignited by [31], compared to traditional methods due to substantially better performance DL has dominated the domain. Several questions about the existence of traditional Computer Vision (CV) techniques have been brought up in the community in recent years [30], which this paper intends to address [29]. The DL developments in past decades are rather rapid. In [28] the authors studied the literature and made the comparison and their respective performance on different CV problems , including image classification, object detection, image retrieval, semantic segmentation, and human pose estimation. After studying several models such as CNN, RBM, Autoencoder, and Sparse Coding, they found that performance of CNN was best for CV. At that time several problems were there in practical application due to the limitation of precisions and model sizes. They included (a) there is not enough literature on performance of architectures; (b) training with limited data; (c) hard to achieve real-time applications; (d) need more powerful models [27].

The rest of the paper is organized as follows: Section II, provides an overview of the existing work in deep learning related to computer vision, Section III presents the potential applications of these techniques. Finally, Section IV gives the concluding remarks [26].

II. EXISTING WORK IN DEEP LEARNING RELATED TO COMPUTER VISION

A. Bits and Pieces together

In this section, we will discuss some of the existing researches which are conducted using deep convolutional neural network. Dan et al. [9] presented Multi-column Deep Neural Network (MCDNN) which was used for handwritten digits and traffic signs recognition. In MNIST dataset Wan et al. [17] used DropConnect method which achieved current best error rate (21%) in digit recognition. In Liu et al. [6] proposed idea ofAU-aware Deep Networks (AUDN) by creating a deep architecture to recognize facial expression. Restricted Boltzmann Machines (RBMs) was used for high level features extraction from each AU-aware Receptive Fields (AURF). Krizhevsky et al. [5] presented a new CNN architecture which achieved top-1 and top-5 error rates of 37.5% and 17.0% respectively on the test data. In the work of Lee et al. [18] constructed a model for scaling with realistic image sizes. This model could translate invariant and supports efficient bottom-up and top-down probabilistic inference as well. Dey et al. [19] constructed deep learning model using “Berkeley-trained” model, [11] for classification of texture based garment design. Using this technique they achieved 73.54% accuracy in Clothing Attribute dataset. Zhoub et al. [8] presented the idea extracting the difference between the density and diversity of image datasets. In Zeiler et al. [20] gave a visualization technique to understand the function of intermediate feature layers. Here, authors addressed large CNN model to improve classification performance. Authors obtained 86.5% and 74.2% accuracy in Caltech-101 and Caltech-256 datasets respectively. In the work of Simon et al. [21] put forward a model which learn part of the model in an unsupervised fashion, which could select generic parts for fine-grained and generic image classification. By using CNN, authors found out neural activation patterns. On the CUB200-2011 and Caltech-256 datasets, this method achieved 81.0% and 84.1% accuracies. Xia et al. [22] gave an idea of DRAE to learn discriminating reconstructions in an auto encoder. The main focus was on automatically removing outliers from noisy data, as outliers were not well reconstructed and would produce more discriminative errors. In the study of Luus et al. [23] achieved 93.48% and 90.26% accuracies by optimizing Deep Convolutional Neural Network (DCNN) hyper parameters using a heuristic approach. Using deep learning, Xua et al. [24] proposed a feature fusion based image retrieval technique, where colors, texture and shape represent the features. Levi et al. [15] proposed a DCNN architecture for gender and age estimation. By using Deep Belief Nets (DBN), Helou et al. [25] proposed a new Convolutional Deep Belief Network (CDBN).However, none of the researches has been conducted to perform comprehensive analysis among these approaches which may help to select an appropriate model for a specific application. Therefore, it is essential to gather a comprehensive knowledge of these models[1].

B. Research Status Of Computer Vision

The deep learning in the field of computer vision is the earliest attempt among many areas [34]. While designing visual algorithms we have to perform in most cases the four processes such as image pre-processing, feature extraction [35], feature selection [36], prediction and recognition [33]. In case of traditional algorithms the first three processes have to be designed manually which is very time consuming and cumbersome work [37].

Computer vision started late in China[38]. People acquired images and then analyse them with the help of machines and at the final stage get results as output. Machines outperformed the human vision in most of the cases [39].

Recently, many researchers working how to perform feature extraction from video and detect the moving target [40]. In the videos and realistic scenes, 3D modelling is performed using the algorithms composed of stereoscopic vision theory. How to make 3D modelling have better synchronization and possess higher degree of reduction, are the main topic of stereoscopic research recently.

Now a days automatic driving technology which is a new field emerging as essential technology for the development of the future automobile industry and there is intense competition between technology companies working in automatic driving technology. In the field of automatic driving technology computer vision methods are also being used widely [41], including vehicle positioning [42], object detecting, object tracking and path planning [43] [32].

C. Techniques

Deep learning has shown great success in the field of Image and Video processing, Computer vision[5,9] and Bioinformatics to name a few. This leads to more research and development of several subfields of deep learning in the above mentioned fields. But prominently deep learning techniques are generally divided into three categories namely

Convolutional Neural Networks(CNN), Restricted Boltzmann Machines(RBM) and Auto encoders. Moreover, Recurrent Neural Networks and Extreme Learning are also a few techniques frequently used in this field. For better clarity, the architectural descriptions along with layer information and detailed functionalities of these techniques are described in a nutshell [26].

Convolutional Neural Network (CNNs) : In case of convolutional neural network there is no need of complicated image pre-processing which we need to do for the algorithm in the past [86], and in CNN we can provide directly original image as an input. Because of this feature the CNN algorithm widely used for image processing and could become more and more popular in many scientific and technological fields at this stage. In CNN there are four key ideas: local connections, shared weights, pooling and the use of many layers including convolutional layers and pooling layers [88].

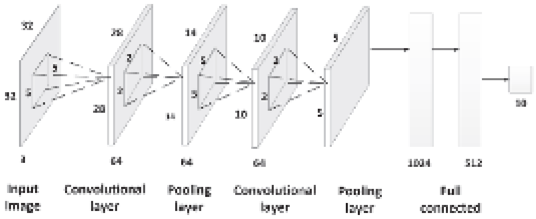

Fig. 1. CNN framework

Each neuron is only connected to a small area near the neuron and the neurons of the network are locally connected. In this way huge amount of computation is avoided that each neuron connects to all the images. Operation of the first layer convolution and activation function produce the feature graph, the dimension of thisfeature graph is significantly reduced compared to the original image of the first layer. Thus obtained feature map is inputted to the operation similar to polymerization which is applied in pooling layer, further reduce the dimension of the output image . Convolution and pooling are constantly crossed. Finally, result is obtained through the fully connected layer. The weights between each layer is adjusted using BP(backpropagation) algorithm [87] in the training process.

***Image Understanding With Deep Convolutional Networks* :**Since the early 2000s, ConvNets achieved great success to the detection, segmentation and recognition of objects and regions in image processing. These were all supervised learning tasks , such as traffic sign recognition, the segmentation of biological images particularly for connect omics, and the detection of faces, text, pedestrians and human bodies in natural images. ConvNetsalso achieved greatly in face recognition task. As images can be labelled at the pixel level, which will be very helpfulin technology, including autonomous mobile robots and self-driving cars. Other applications gaining importance involve natural language processing and speech recognition. ConvNets are now gaining popularity in almost all recognition and detection tasks and outperform humansin some tasks. ConvNets and recurrent net modules jointly demonstrated stunning performance for the generation of image captions recently. ConvNets are easily amenable to efficient hardware implementations in chips or field-programmable gate arrays. A number of companies such as NVIDIA, Mobileye, Intel, Qualcomm and Samsung are developing ConvNet chips to enable real-time vision applications in smartphones, cameras, robots and self-driving cars (YannLeCun et al., 2015). The different CNN architectures include Deep Max-Pooling Convolutional Neural Networks, Very Deep Convolutional Neural Networks, Network In Network, Region-based Convolutional Neural, Fast R-CNN, Faster R-CNN, Mask R-CNN, Multi-Expert R-CNN, Deep Residual Networks, Resnet in Resnet, ResNeXt and Capsule Networks[46].

2) *Recurrent Neural Networks (RNNs) and the LSTM* :RNNs [91] are applicable in the tasks like processing sequential data, such as speech, text, videos, and time-series, where data at any given time/position depends on previously encountered data. At each time-stamp the model collects the input from the current time Xi and the hidden state from the previous step hi-1, and outputs a target value and a new hidden state (Figure 2). RNNs are typically suffers with gradient vanishing or exploding problems in many real-world applications.

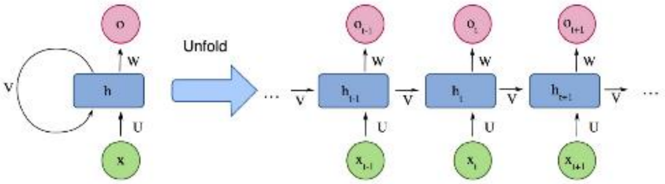

Fig. 2. Architecture of a simple recurrent neural network

These problems in RNN could be avoided using algorithm such as Long Short Term Memory (LSTM) [90]. The LSTM architecture (Figure 3) includes three gates (input gate, output gate, forget gate), which regulate the flow of information into and out from a memory cell, which stores values over arbitrary time intervals[89].

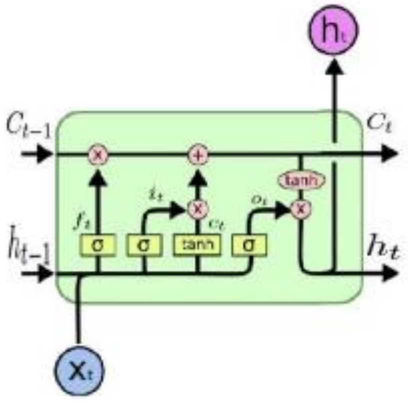

Fig. 3. Architecture of a standard LSTM module. Courtesy of Karpathy

3) Restricted Boltzmann Machines(RBMs): The contrastive divergence proposed by Hinton[47]helped in solving the training efficiency problem in RBM. The RBM earlier would use Stochastic gradient descent method requires a large number of sampling steps, still not makes the training efficiency of RBM high. Restricted Boltzmann Machines (RBM) are special type of Markov random field containing one layer of stochastic hidden units [50, 51]. In Hinton and Salakhutdinov (2011) for document processing presented a Deep Generative Model using Restricted Boltzmann Machines (RBM) [46]. (Hinton and Sejnowski) proposed Boltzmann Machine (BM) in 1986. Boltzmann Machine belongs to the type of feedback neural network is a random neural network. Boltzmann Machine made up of some visible units (visible variables, i.e. data samples) and some hidden units (hidden variables), there is connection between each visible unit to all the hidden units, the visible variables and hidden variables represent in binary form (0 or 1), 0 means the neuron is in suppressed state, and 1 means the neuron is in active state. (Sejnowski et al.) further presented restricted Boltzmann machine (RBM). Visible layer is used to trainthe data, then the hidden layer produces the features of input data. The training of Restricted Boltzmann machine is faster than Autoencoder. In (Le Q V, 2011) proposed a more efficient optimization algorithm based on the stochastic gradient descent method. Contrastive divergence proposed by Hinton solved training efficiency problem in RBM [47].

4) Auto-encoder :In deep neural network AE is an approach applied with efficient data encoding and decoding for unsupervised feature learning. The main objective of auto encoder is to reduce noise in data and learn and representation (encoding) of data, typically for data dimensionality reduction, compression, fusion and many more. Auto encoder technique consists of two parts: the encoder and the decoder. Encoder maps the input samples usually in the lower dimensional features space with a constructive feature representation. This process is repeated until the desired feature dimensional space is reached. Whereas decoder performs reverse processing to regenerate actual features from lower dimensional features [45]. The learning algorithm is based on the implementation of the backpropagation. Auto encoders extend the idea of principal component analysis (PCA) [49]. In a deep AE, lower hidden layers are used for encoding and higher ones for decoding, and error back-propagation is used for training [50].[51][46] Following are the types of Auto-encoders:

- *De-noising Auto-encoder* :Itis a modification on the autoencoder. To address the identity functions, these encoders corrupt the input and afterwards, reconstruct them. It is also called the stochastic version of the auto-encoders[53] . In early Auto-Encoders (AE), encoding layer had smaller dimensions than the input layer. In Stacked Denoising Auto-Encoders (SDAE), encoding layer is wider than the input layer[50][46].

- *Sparse Auto-encoder* :A sparse autoencoder is one of a range of types of autoencoder artificial neural networks that work on the principle of unsupervised machine learning. These auto-encoders have the learning methods that automatically extract the features from the unlabelled data. Here the word sparse indicates that hidden units are allowed to fire only for the certain type of inputs and not too frequently [53].

- *Variational Auto-Encoder (VAE)* : It is made up of an encoder, decoder and a loss function. VAEs are used for the designing of the complex models of the data that too with large datasets. It is also known as high resolution network [53]. VAEs are built upon standard neural networks and can be trained with stochastic gradient descent (Doersch, 2016) [46].

- *Contractive Auto-encoder (CAE)* : These are robust networks as de-noising auto-encoders but the difference is that the contractive auto-encoders generate robustness in the networks through encoder function whereas de-noising auto-encoders work with the reconstruction process [53].

- *Transforming Autoencoders* : Deep Auto-Encoders (DAE) can be transformation-variant, i.e., the extracted features from multilayers of non-linear processing could be changed due to learner. Transforming Auto-Encoders (TAE) work with both input vector and target output vector to apply transformation-invariant property and lead the codes towards a desired way [50] [46] [44].

5) Extreme Learning **:**Extreme Learning is a feed-forward neural network used for regression and classification tasks proposed by Guang-Bin Huang. It consists of a solitary layer of masked nodes in which the weights which are assigned as inputs to the masked nodes are random and are never corrected. In one step, the weights between the masked nodes and outputs are learned, which leads to learning of a linear model. It doesn't require gradient-based backpropagation to work. It uses Moore-Penrose generalized inverse to set its weights. These are better than networks trained by using back-propagation because of their faster learning ability and a good generalization capability [26].

III. APPLICATIONS IN COMPUTER VISION

In this section, we survey works that have leveraged deep learning methods to address key tasks in computer vision, such as object detection, face recognition, action and activity recognition, and human pose estimation etc.

A. Classification

The task of identifying what an image represents is called image classification. The applications include identifying gender given an image of a person's face, identifying the type of pet, tagging photos, and so on.

B. Detection and Localization

Object localization refers to identifying the location of one or more objects in an image and drawing abounding box around their extent. Object detection combines these two tasks and localizes and classifies one or more objects in an image. This has many real-world applications, especially in the automotive industry where self-driving cars detect objects through their camera sensors.

C. (Semantic) Segmentation

Image segmentation involves converting an image into a collection of regions of pixels that are represented by a mask or a labeled image. By dividing an image into segments, you can process only the important segments of the image instead of processing the entire image. It is useful for processing medical images and satellite imagery.

D. Similarity Learning

Similarity learning is the process of learning how two images are similar.A score can be computed between two images based on the semantic meaning. There are several applications of this, from finding similar products to performing facial identification.

E. Image Captioning

Image Captioning is the process of generating a textual description for given images. It has been a very important and fundamental task in the Deep Learning domain.

F. Generative Models

Generative models generate images. In style transfer application to generate an image uses the content of that image and the style of other images. For an example, an image of a temple is generated using the style of a pencil sketch. Generative models help in for other purposes such as new training examples, super-resolution images, and so on [54].

G. Action and Activity Recognition

Deep learning techniques have application in human activity recognition and works on this have been proposed in the literature in the last few years [57]. In [58] presented study on deep learning using for complex event detection and recognition in video sequences. CNN-based approach used in [59]for activity recognition in beach volleyball, in [60] for event classification from large-scale video datasets and in [61] used for activity recognition based on smartphone sensor data. In [56], the study is presented on applicability of CNN as joint feature extraction and classification model for fine-grained activities. In [62], put forward the idea for recognizing group activities in crowded scenes collected from the web using mixed appearance and motion features. In [63] recognized complex event based on combination of heterogeneous features . Study in [64], uses both the video and sensor data and employing a dual CNNs and Long Short-Term Memory architecture to construct a multimodal multistream deep learning framework which is used in egocentric activity recognition problem. Multimodal fusion with a combined CNN and LSTM approach is also proposed in [65]. Finally, in [66] input video sequences that also include depth information to DBNs for activity recognition.

H. Human Pose Estimation

In this data is provided by motion capturing hardware such as images, image sequences, depth images, or skeleton data human pose estimation system determines the position of human joints using this data[71]. In [69] they proposed a model Deep Pose which is a holistic model that formulates the human pose estimation method as a joint regression problem and does not explicitly define the graphical model or part detectors for the human pose estimation. In [70], they trained the network using the local part patches and background patches to train a CNN, to determine conditional probabilities of the part presence and spatial relationships. In[72] presented the ideato train multiple smaller CNNs to perform independent binary body-part classification, followed with a higher-level weak spatial model to remove strong outliers and to enforce global pose consistency. Finally, in [73], performed heat-map likelihood regression for each body part, followed with an implicit graphic model to further promote joint consistency using CNN.

I. Datasets

The applicability of deep learning approaches has been evaluated on numerous datasets, whose content varied greatly, according the application scenario. Regardless of the investigated case, the main application domain is (natural) images. A brief description of utilized datasets (traditional and new ones) for benchmarking purposes isprovided below.

1) *Grayscale Images* : MNIST [74], NIST and perturbed NIST.
2) *RGB Natural Images* : Caltech RGB image datasets [75], CIFAR datasets [76], COIL datasets [77] .
3) *Hyperspectral Images* : SCIEN hyperspectral image data [78] and AVIRIS sensor based datasets [79].
4) *Facial Characteristics Images* :Adience benchmark dataset [80].
5) *Medical Images* : Chest X-ray dataset [81] Lymph Node Detection and Segmentation datasets [82].
6) *Video Streams* : The WR datasets [83, 84] and YouTube-8M [85][BB].

J. Semantic Segmentation

CNN models are used for semantic segmentation tasks, as it is potent of handling the pixel-level predictions. Output masks having a 2-dimensional spatial spread are required by semantic segmentation. The process of semantic segmentation is as Detection based Segmentation, FCN-CRFs Based Segmentation and Weakly supervised annotations [26].

IV. CONCLUSION

In this paper we have given a survey of Deep learning and its recent development. The analysis of prevailing deep learning architectures is done by developing a categorical layout. Deep learning algorithms are divided into three categories: Convolutional Neural Network, Restricted Boltzmann Machines, Autoencoder. Apart from that, RNN and extreme learning are also quite popular. In this paper we mainly dealt with the recent advancement of CNN dependent strategies, since it is mostly used for images. CNNs have the unique capability of feature learning, that is, of automatically learning features based on the given dataset. CNNs are also invariant to transformations, which is a great asset for certain computer vision applications. On the other hand, they heavily rely on the existence of labelled data, in contrast to DBNs/DBMs and SdAs, which can work in an unsupervised fashion. Of the models investigated, both CNNs and DBNs/DBMs are computationally demanding when it comes to training, whereas SdAs can be trained in real time under certain circumstances. Benchmark data sets are not available in every field. Transfer learning can be very useful in this case. But, the requirement of high computing makes it difficult to implement it on low computing devices especially handheld IoT devices. Apart from this, the requirement of high memory also becomes an obstacle for handheld devices. In DL, the choice of hyper parameters (number of layers, learning rate, kernel size, stride size, pooling, etc) are very vital. A good choice of hyper parameters requires good skills and experience. As a closing note, in spite of the promising—in some cases impressive— results that have been documented in the literature, significant challenges do remain, especially as far as the theoretical groundwork that would clearly explain the ways to define the optimal selection of model type and structure for a given task or to profoundly comprehend the reasons for which a specific architecture or algorithm is effective in a given task or not. These are among the most important issues that will continue to attract the interest of the machine learning research community in the years to come.

REFERENCES

[1] X. Y. Jing, F. Wu, Z. Li, R. Hu and D. Zhang, "Multi-Label Dictionary Learning for Image Annotation," in IEEE Transactions on Image Processing, vol. 25, no. 6, pp.

[2] S.M. Sofiqul Islam, Shanto Rahman, Emon Kumar Dey, Application of Deep Learning to Computer Vision: A Comprehensive Study, 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, IEEE

[3] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised featurelearning for audio classification using convolutional deep belief networks,” in Advances in neural information processing systems, 2009,pp. 1096–1104.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014, pp. 580–587.

[5] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[6] Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[7] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 2013, pp. 1–6.

[8] E. M¨akinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 30, no. 3, pp. 541–547, 2008.

[9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in Advances in Neural Information Processing Systems, 2014, pp. 487–495.

[10] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 3642– 3649.

[11] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," arXiv preprint arXiv:1405.3531, 2014.

[12] P. Heit, "The berkeley model," Health education, vol. 8, no. 1, pp. 2–3, 1977.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," arXiv preprint arXiv:1411.4038, 2014.

[14] J. Zhang, M. Sameki, S. Ma, B. Price, R. Mech, X. Shen, M. Betke, S. Sclaroff, and Z. Lin, "Salient object subitizing," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," arXiv preprint arXiv:1505.02496, 2015

[16] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks."

[17] Z. Ge, C. Mccool, and P. Corke, "Content specific feature learning for fine-grained plant classification," in Working notes of CLEF 2015 conference, 2015.

[18] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 1058–1066.

[19] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, pp. 609–616.

[20] E. K. Dey, M. N. A. Tawhid, and M. Shoyaib, "An automated system for garment texture design class identification," Computers, vol. 4, no. 3, pp. 265–282, 2015.

[21] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Computer Vision–ECCV 2014. Springer, 2014, pp. 818–833.

[22] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," arXiv preprint arXiv:1504.08289, 2015.

[23] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1511– 1519.

[24] F. Luus, B. Salmon, F. van den Bergh, and B. Maharaj, "Multiview deep learning for land-use classification," Geoscience and Remote Sensing Letters, IEEE, vol. 12, no. 12, pp. 2448–2452, 2015.

[25] Q. Xua, S. Jiangb, W. Huangb, F. Yeb, and S. Xub, "Feature fusion based image retrieval using deep learning."

[26] A. Helou and C. Nguyen, "Unsupervised deep learning for scene recognition," 2011.

[27] RajatKumar Sinha, Ruchi Pandey, Rohan Pattnaik, Deep Learning For Computer Vision Tasks: A review, International Conference on Intelligent Computing and Control (I2C2) ,2017

[28] Junyi Chai, Hao Zeng, Anming Li, Eric W.T. Ngai, Deep learning in computer vision: A critical review of emerging techniques and application scenarios, www.elsevier.com/locate/mlwa, Machine Learning with Applications 6 (2021) 100134.

[29] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. Neurocomputing, 187, 27–48. http://dx.doi.org/10.1016/j.neucom.2015.09.116.

[30] Niall O' Mahony, Sean Campbell, Anderson Carvalho, SumanHarapanahalli, Gustavo Velasco Hernandez, LenkaKrpalkova, Daniel Riordan, Joseph Walsh, Deep Learning vs. Traditional Computer Vision.

[31] Nash W, Drummond T, Birbilis N (2018) A Review of Deep Learning in the Study of Materials Degradation. npj Mater Degrad 2:37. https://doi.org/10.1038/s41529-018-0058-x

[32] O'Mahony N, Murphy T, Panduru K, et al (2017) Real-time monitoring of powder blend composition using near

infrared spectroscopy. In: 2017 Eleventh International Conference on Sensing Technology (ICST). IEEE, pp 1–6

[33] Qing Wu, Yungang Liu, Qiang Li, Shaoli Jin and Fengzhong Li, The Application of Deep Learning in Computer Vision, IEEE,2017.

[34] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[35] Y. Jia, E. Shelhamer, J. Donahue, et al., “Caffe: Convolutional architecture for fast feature embedding,” in Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678, 2014.

[36] Y. Zhu, T. Tan and Y. Wang, “Biometric personal identification based on iris patterns,” in Proceedings of the 15th International Conference on Pattern Recognition, pp. 801–804, 2000.

[37] Guyon and A. Elisseeff, “An introduction to variable and feature selection, ”Journal of Machine Learning Research, vol.3, pp. 1157–1182, 2003.

[38] O. W. Salomons, F. J. A. M. van Houten and H. J. J. Kals, “Review of research in feature-based design,” Journal of Manufacturing Systems, vol. 12, no. 2, pp. 113–132, 1993.

[39] T. Brosnan and D. W. Sun, “Improving quality inspection of food products by computer vision: A review,” Journal of Food Engineering, vol. 61, no. 1, pp. 3–16, 2004.

[40] D. M. Gavrila, “The visual analysis of human movement: A survey,” Computer Vision and Image Understanding, vol. 73, no. 1, pp. 82–98, 1999.

[41] A. J. Lipton, H. Fujiyoshi and R. S. Patil, “Moving target classification and tracking from real-time video,” in Proceedings of the 4th IEEE Conference on the Workshop on Applications of Computer Vision, pp. 8–14, 1998.

[42] Geiger, P. Lenz and R. Urtasun, “Are we ready for autonomous driving? thekitti vision benchmark suite,” in Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354– 361, 2012.

[43] M. M. Trivedi, T. Gandhi and J. McCall, “Looking-in and lookingout of a vehicle: Computer-vision-based enhanced vehicle safety,” IEEE Transactions on Intelligent Transportation Systems, vol. 8, no. 1, pp. 108– 120, 2007.

[44] Y. K. Hwang and N. Ahuja, “A potential field approach to path planning,” IEEE Transactions on Robotics and Automation, vol. 8, no. 1, pp. 23–32, 1992.

[45] Avinash H. Hedaoo, Deep Learning and its Application: A Review, International Journal of Research Publication and Reviews, Vol 3, no 10, pp 340-352, October 2022.

[46] MdZahangir Alom1, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, PahedingSidike, MstShamimaNasrin, Brian C Van Essen, Abdul A S. Awwal, and Vijayan K. Asari, The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches, Computer Vision and Pattern Recognition, 2018.

[47] Matiur Rahman Minar, JibonNaher Jul 2018, Recent Advances in Deep Learning , arXiv:1807.08169v1 [cs.LG].

[48] Ruihui Mu, Xiaoqin Zeng A Review of Deep Learning Research, KSII Transactions On Internet And Information Systems Vol. 13, no. 4, Apr. 2019

[49] LaithAlzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Du`an, Omran Al-Shamma, J. Santamaria, Mohammed A. Fadhel, Muthana Al-Amidie and LaithFarhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, Journal of Big Data, 2021

[50] Ajay Shrestha And AusifMahmood, Review of Deep Learning Algorithms and Architectures, Vol 7 2019 IEEE Access

[51] Li Deng and Dong Yu, Deep Learning Methods and Applications, Foundations and Trends R_ in Signal Processing, Vol. 7, 2013

[52] J. Goodfellow et al., `Generative adversarial networks," ArXiv e-prints, Jun. 2014. [Online]. Available: http://adsabs.harvard.edu/abs/ 2014arXiv1406.2661G

[53] Hinton, G. et al.: Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process.Mag., 29 (6) (2012), 82–97.

[54] Shaveta Dargan, Munish Kumar, MaruthiRohitAyyagari, Gulshan Kumar, A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning , CIMNE, Barcelona, Spain 2019

[55] SunilaGollapudi Foreword by V Laxmikanth Learn Computer Vision Using OpenCV With Deep Learning CNNs and RNNs, , Apress, https://doi.org/10.1007/978-1-4842-4261-2

[56] AthanasiosVoulodimos, NikolaosDoulamis, AnastasiosDoulamis, and EftychiosProtopapadakis, Deep Learning for Computer Vision: A Brief Review, Hindawi Computational Intelligence and Neuroscience Volume 2018, Article ID 7068349, 13 pages https://doi.org/ 10.1155/ 2018/ 7068349

[57] S. Cao and R.Nevatia, “Exploring deep learning based solutions in fine grained activity recognition in the wild,” in Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 384–389, Cancun, December 2016.

[58] Makantasis, A. Doulamis, N. Doulamis, and K. Psychas, “Deep learning based human behavior recognition in industrial workflows,” in Proceedings of the 23rd IEEE International Conference on Image Processing, ICIP 2016, pp. 1609–1613, September 2016.

[59] C.Gan,N. Wang, Y. Yang, D.-Y. Yeung, and A.G. Hauptmann, “DevNet: A Deep Event Network for multimedia event detection and evidence recounting,” in Proceedings of the IEEE Conference on Computer Vision

and Pattern Recognition, CVPR 2015, pp. 2568–2577, USA, June 2015.

[60] T. Kautz, B. H. Groh, J. Hannink, U. Jensen, H. Strubberg, and B. M. Eskofier, “Activity recognition in beach volleyball using a DEEpConvolutionalNeuralNETwork: leveraging the potential of DEEp Learning in sports,” Data Mining and Knowledge Discovery, vol. 31, no. 6, pp. 1678–1705, 2017.

[61] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.-F. Li, “Large-scale video classification with convolutional neural networks,” in Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, (CVPR ’14), pp. 1725–1732, Columbus,OH, USA, June 2014.

[62] A. Ronao and S.-B. Cho, “Human activity recognition with smartphone sensors using deep learning neural networks,” Expert Systems with Applications, vol. 59, pp. 235–244, 2016.

[63] Shao, C. C. Loy, K. Kang, and X. Wang, “Crowded Scene Understanding by Deeply Learned Volumetric Slices,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 3, pp. 613–623, 2017.

[64] Tang, B. Yao, L. Fei-Fei, and D. Koller, “Combining the right features for complex event recognition,” in Proceedings of the 2013 14th IEEE International Conference on Computer Vision, ICCV 2013, pp. 2696–2703, Australia, December 2013.

[65] S. Song, V. Chandrasekhar, B.Mandal et al., “MultimodalMulti- Stream Deep Learning for Egocentric Activity Recognition,” in Proceedings of the 29th IEEE Conference on Computer Vision and Pattern RecognitionWorkshops, CVPRW2016, pp. 378–385, USA, July 2016.

[66] R. Kavi, V. Kulathumani, F. Rohit, and V. Kecojevic, “Multiview fusion for activity recognition using deep neural networks,” Journal of Electronic Imaging, vol. 25, no. 4, Article ID 043010, 2016.

[67] H. Yalcin, “Human activity recognition using deep belief networks,” in Proceedings of the 24th Signal Processing and Communication Application Conference, SIU2016, pp. 1649–1652, tur, May 2016.

[68] A. Toshev and C. Szegedy, “DeepPose: Human pose estimationvia deep neural networks,” in Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 1653–1660, USA, June 2014.

[69] X. Chen and A. L. Yuille, “Articulated pose estimation by a graphical model with image dependent pairwise relations,” in Proceedings of the NIPS, 2014.

[70] Liu, N. Lay, Z.Wei et al., “Colitis detection on abdominal CT scans by rich feature hierarchies,” in Proceedings of the Medical Imaging 2016: Computer-Aided Diagnosis, vol. 9785 of Proceedings of SPIE, San Diego, Calif, USA, February 2016.

[71] G. Luo, R.An, K.Wang, S.Dong, andH. Zhang, “ADeepLearning Network for Right Ventricle Segmentation in Short:Axis MRI,” in Proceedings of the 2016 Computing in Cardiology Conference

[72] Kitsikidis, K. Dimitropoulos, S.Douka, andN.Grammalidis, “Dance analysis usingmultiplekinect sensors,” in Proceedings of the 9th International Conference on Computer VisionTheory and Applications, VISAPP 2014, pp. 789–795, prt, January 2014.

[73] Jain, J. Tompson, and M. Andriluka, “Learning human pose estimation features with convolutional networks,” in Proceedings of the ICLR, 2014.

[74] J. Tompson, A. Jain, Y. LeCun et al., “Joint training of a convolutional network and a graphical model for human pose estimation,” in Proceedings of the NIPS, 2014.

[75] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2323, 1998.

[76] Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” IEEE Transactions on Pattern Analysis andMachine Intelligence, vol. 28, no. 4, pp. 594–611, 2006.

[77] Krizhevsky and G. Hinton, Learning multiple layers of features from tiny images, 2009.

[78] S. A. Nene, S. K. Nayar, and H.Murase, Columbia object image library (coil-20), 1996.

[79] T. Skauli and J. Farrell, “A collection of hyperspectral images for imaging systems research,” in Proceedings of the Digital Photography IX, USA, February 2013.

[80] F. Baumgardner, L. L. Biehl, andD.A. Landgrebe, “220 band avirishyperspectral image data set: June 12, 1992 indian pine testsite 3,” Datasets, 2015.

[81] E. Eidinger, R. Enbar, and T. Hassner, “Age and gender estimation of unfiltered faces,” IEEE Transactions on Information Forensics and Security, vol. 9, no. 12, pp. 2170–2179, 2014.

[82] X.Wang,Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R.M. Summers, “ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471, Honolulu, HI, May 2017.

[83] Seff, L. Lu, A. Barbu, H. Roth, H.-C. Shin, and R. M. Summers, “Leveraging mid-level semantic boundary cues for automated lymph node detection,” Lecture Notes in Computer Science (including subseries

LectureNotesinArtificial Intelligence and Lecture Notes in Bioinformatics): Preface, vol. 9350, pp. 53– 61, 2015.

[84] Voulodimos, D. Kosmopoulos, G. Vasileiou et al., “A dataset for workflow recognition in industrial scenes,” in Proceedings of the 2011 18th IEEE International Conference on Image Processing, ICIP 2011, pp. 3249–3252, Belgium, September 2011.

[85] Voulodimos, D. Kosmopoulos, G. Vasileiou et al., “A threefold dataset for activity and workflow recognition in complex industrial environments,” IEEE MultiMedia, vol. 19, no. 3, pp. 42–52, 2012.

[86] S. Abu-El-Haija et al., “YouTube-8M: A large-scale video classification benchmark,” Tech. Rep., 2016, https://arxiv.org/abs/ 1609.08675.

[87] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[88] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning internal representations by error propagation, MIT Press, Cambridge, 1986.

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

[90] ShervinMinaee, Yuri Boykov, FatihPorikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos, Image Segmentation Using Deep Learning: A Survey, arXiv:2001.05566v5 [cs.CV] 15 Nov 2020

[91] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[92] D. E. Rumelhart, G. E. Hinton, R. J. Williams et al., “Learning representations by back-propagating errors,” Cognitive modeling, vol. 5, no. 3, p. 1, 1988.