Self-Learning by Word Localization from Images

Saranya Manoharan
Vellore Institute of Technology University
India

Muthu Kumar B
Amrita Vishwa Vidyapeethem
India

ABSTRACT
Artificial Intelligence is an interdisciplinary research area which aims at making the machines more human. Extensive research is going on to teach them to perform the tasks. Deep learning is a collection of algorithms in Machine Learning. In this paper we implement deep learning for learning and gaining the knowledge of the text from real time images. An algorithm namely word localization is proposed to able to make the machine to understand the words extracted from the images. In comparison with traditional Optical character recognition (OCR) it has many advantages over it which is been analyzed.

General Terms
Machine Learning with self-teaching algorithm

Keywords
Machine Learning, Deep Learning Unsupervised Learning, Supervised Learning, Robot Vision, Robot Grasping, Machine Vision, Word Localization, Knowledge transfer of text

1. INTRODUCTION
Artificial Intelligence has been widely focused as a new upcoming field of study. It makes the machine to think and act like humans. AI focuses on different problems or goals like perception, thinking multi dimensional movement, psychology, speech, and Machine Learning. Machine Learning is a study of computer algorithms that teaches a machine to learn from data. There are several learning methods in this area: Supervised learning, unsupervised learning, Semi-supervised learning, Reinforcement learning are namely the few methods. In this paper we are likely to apply both Supervised and Unsupervised learning – Semi-supervised learning.

Supervised learning instructs the learning system on the labels to map with training data, where the training data is a set of training examples and labels are the class labels obtained from classification methodology. It works in such a way by taking a known set of inputs and known set of outputs and builds a model as in Fig 1, that generates reasonable predictions for the response of new data.

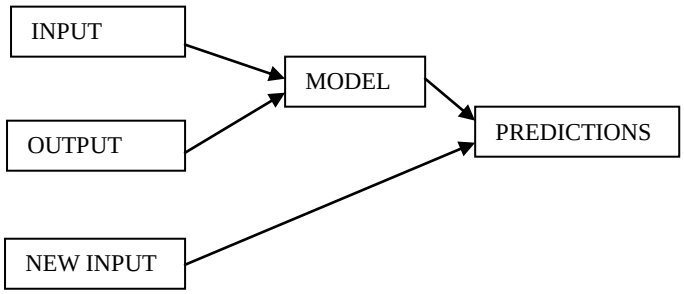

Fig 1: Supervised Learning

Unsupervised learning enables a learning system to represent particular set of inputs in a way that reflects the structure of the overall collection of inputs. The unsupervised learning like Deep belief networks[1], Restricted Boltzmann Algorithm[2] and auto encoders[3],[4] were proposed in deep learning. In comparison with Supervised learning there are no labeled outputs associated to each input. It is used to draw conclusion with input data without mapping to labeled responses.

Semi-supervised learning [5],[6] resides between Supervised and Unsupervised learning. It is generally determined as a class of supervised learning which also uses unlabelled data. It works by considering small amount of labeled data with a large amount of unlabelled data. In our paper initially the system is fed with small number of data to be able to start with basic understanding of words. Later on the vocabulary and the understanding is increased by adding the unlabelled data into the storage. Hence it is considered to be semi-supervised learning algorithm. The words are self taught by itself.

Deep learning is a set of algorithms used to model high-level abstractions in data which is composed of multiple non-linear transformations. It is based upon learning representation or Feature learning which learns a transformation of raw inputs to a representation that is exploited in a supervised learning task- classification. Feature learning algorithms may be either unsupervised or supervised. Also Google is been successful in identifying house address number through the street view which is also implemented in Facebook. It basically uses the deep learning image segmentation.

The coming sections of this paper are explained as following. In section 2, optical character recognition is explained. In section 3, deep learning and the text extraction methods used to categorize the words are explained. In section 4 the results are compared between the OCR and Deep learning algorithms. Section 5 & 6 explains about the conclusion and Future work of the paper.

2. OPTICAL CHARACTER RECOGNITION
The optical character recognition technique is used in a single plane image. The image is converted to grayscale image and from that grayscale image the binary image is formed which is used in edge detection. After the text extraction is done from the image the feature extraction is done.

Feature extraction extracts the text from the image. Then it is given as input to classifier of a Neural Network, which is shown in Fig 2, where the classification of the character is done. The neural network has a single hidden layer and it is trained with a training set. The OCR can extract text only up to a certain limit. When a depth image is given as input to the OCR it cannot extract the text efficiently. The OCR can be explained by the Fig 3.

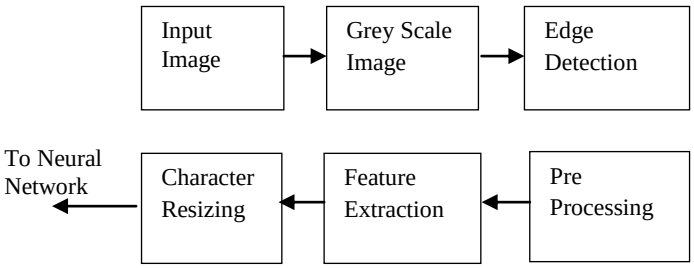

Fig 2: Basic Text Extraction

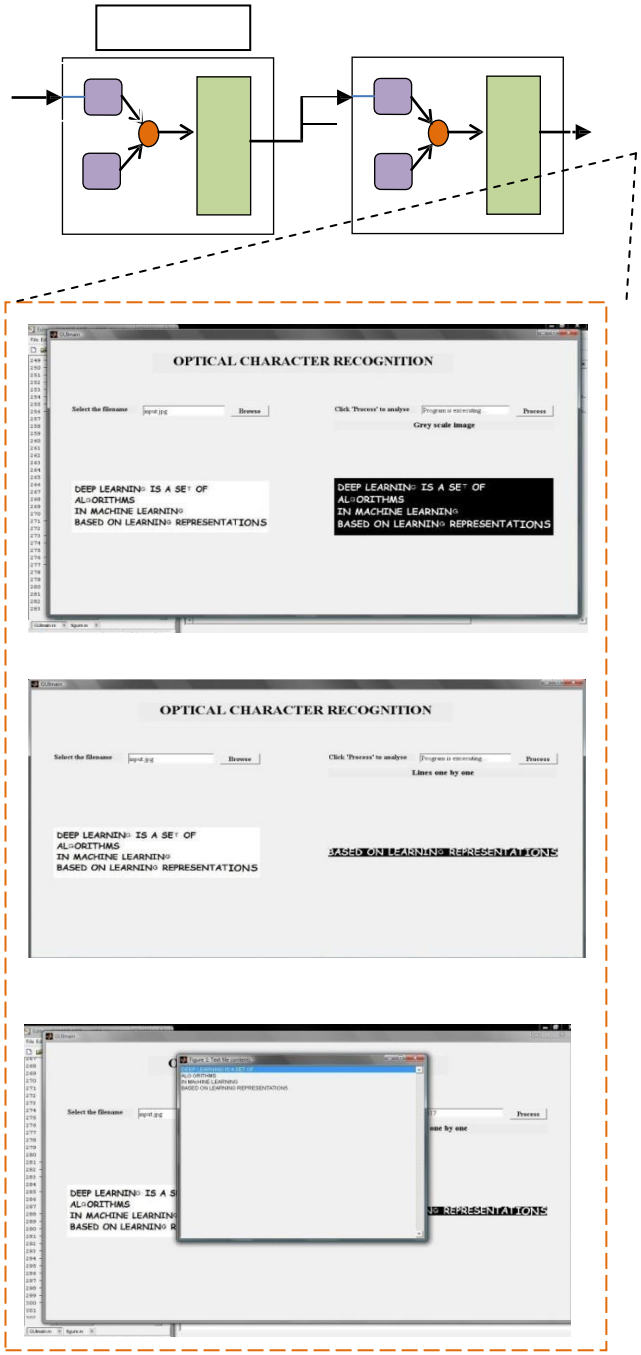

Fig 3: Model of Neural Network

3. DEEP LEARNING

Humans have complex type of neurons and the images they see are processed in more detailed manner. But for a machine it is difficult to analyze the image and understand it. In this paper we propose a method for grasping the data from an image. Deep learning is usually used in identifying objects from image using their edges. It can also be implemented to extract the texts in image. The text recognition is done from the image and for grasping the knowledge of the grasped text we have mapped the data to a trained set.

3.1 Machine Self Learning

The self-taught unsupervised learning is done by sparse filtering method. In deep learning the Deep belief networks and convolution methods are inferred to make the processing more complex in depth images, natural images and high dimensional image. The complexity also grows. So the Sparse filtering [9], [10] method is used in our algorithm to extract the text from depth images.

The sparse filtering uses three levels of hidden layer. When more than one hidden layer is used the processing becomes higher but the feature extraction is lesser. This indicates to deep learning methods whereas shallow learning needs less processing but more features to be extracted.

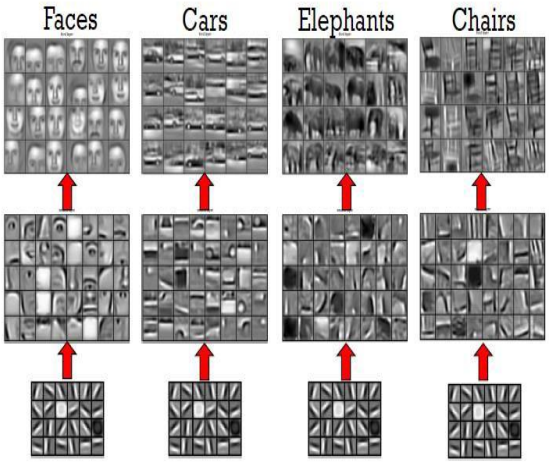

Fig 4: Sparse Filtering Deep Learning

Like in the Fig 4 there are different neurons[7] which can be trained for identifying the objects or alphabets. For text extraction we need only two neurons, one is alphabet identification neuron and another is a number identifying neuron.

When a depth image is given as input to the sparse filtering the first hidden layer as in Fig 7 forms the input similar to Gabor filtering algorithm. Gabor filtering is helpful in detecting edges, blobs and corners. Then its output is given to second hidden layer as input. The second hidden layer pools the data of the first hidden layer. Pooling is a technique in which the lower level pixels are grouped together with same intensities. It goes on in a hierarchical manner.

Pixel → Edges → Characters → Words → Sentences

The text is of different angles, fonts and styles. It is extracted from the image and resized.

The system detects the text from a stored or preloaded data. But to detect the text from a live image and to get the

knowledge of the text which is detected and act accordingly is self teaching of in machine learning.

First it detects the pixels then by edge detection it extracts the character and by pooling it forms the sentences if any sentence is there in the image. The character recognition helps the robot in self-teaching of the words which can be used in future or second time the robot finds the same word in the image. The self-teaching vocabulary in the robot can be built by this technique.

The grasping of data or word from an image is done. The text recognition is done from the image and for grasping the knowledge of the grasped text we have mapped the data to a trained set.

3.1.1 Effects of Variable Font Size

When the text characters differ in font size there is a variation of performance as in [8] which is overcome by scaling the image to a standard size of 10 x 10 and then applying the morphological comparison of the character.

Precision ——————————————

For texts having skewness the skewed text line detection method is applied and the skewed text is converted to horizontal text. The increase in font size to a range till 170 does not have any effect in precision after which it deteorates gradually which is shown in Fig 5.

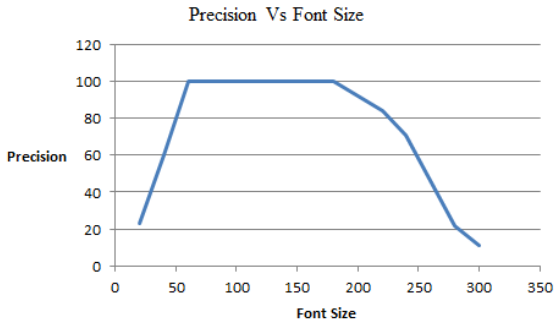

Fig 5: Precision Vs Font Size

3.2 Knowledge Transfer by Text

The text is extracted from the input image and the words or sentences are available in a text format. The vocabulary can be built by the machine by adding the non-duplicate words which are self-taught by itself as in Fig 6.

The GPS based driving cars or robots can be taught to read the traffic signals, distance of destinations, taking a left turn or right, detecting a restaurant or motel on the travel path. The words can be stored in a database which can be used for future reference.

The words have to make a meaning when in sentence. For example "STOP" as a single word in traffic board requires to stop the vehicle, but when it come a sentence it should do the same meaning. For this the machine has to be trained for every new word or sentence.

3.2.1 WORD Localization

The newly grasped words stored in a database of the machine. These words are grouped based on their localization in their sentences. For detecting a complete sentence "." Or more than one whitespaces are considered as stopping criterion. The words are grouped based on the sentences they appear. If a name "Riya"s Inn" is detected it is stored in the database under hotels category. The traffic signals are stored under a separate category and the actions to be performed for their respective texts are also predefined.

Fig 6: Word Localization

So when a search for Food, Restaurant, Places to eat, Cuisine, Hotel are searched then it can locate the already plotted restaurant by which it has travelled before or even a new restaurant which is just been acquired by scanning the image seen at present.

For word localization the whole sentence is analyzed for noun and verbs. So that the noun and actions can be classified. Because if there is traffic signal saying to speed down the car on a bridge or speed up the car on a highway it has to drive fast or slow down based on the text scanned. This is like giving online commands to a robot in runtime through vision representation. The more the training sets are stored the accuracy or understanding of the system increases.

The training set consists of 16342 sentences with 97737 words stored in the database. The trained dataset only is considered for English Language. Other languages can also be included in the database

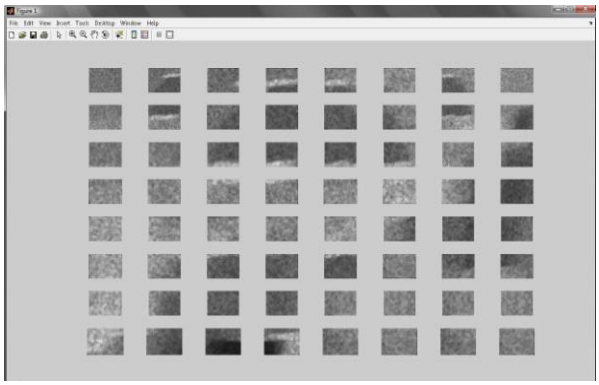

Fig 7: Sparse Filtering Deep Learning

4. RESULT AND COMPARISON

Table 1. OCR Vs Deep Learning

Algorithm	Nature of Image	Number of Input images	Precision
Optical Character Recognition	Plane image	6000	82.3%
Deep Learning	Plane image	6000	93%
Optical Character Recognition	Depth image	3874	34.9%
Deep Learning	Depth image	3874	84.1%

From the above Table 1. It can be inferred that the OCR performs well when a Plane image is given as input. The text extraction is simpler in a plane image where the edge detection and intensities are alone taken into account. When a depth image is given to OCR algorithm the text extraction is almost not efficient. The number of images as shown in the table is given as inputs and more number of words are added to the local database.

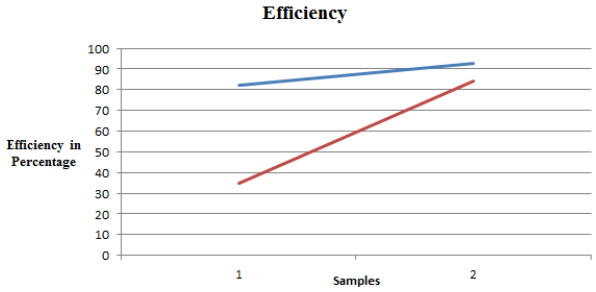

Fig 8: Efficiency Graph

From Fig 8 when the Deep algorithm is given with a depth image it is observed to have a efficiency of 84.1% tested with 14 images of traffic signals, Hotel names, Motel names, buildings and so on.

5. CONCLUSION

In this paper the Deep learning algorithm takes about 3.6 times processing speed higher than that of a OCR. OCR also cannot be used for multidimensional images. The more the dimension of the image the processing speed of Deep Learning increases in which is a tradeoff made for precision text extraction.

6. FUTURE WORK

The software based data processing can be used for any applications not depending upon the hardware of the application. The categorization of words is to be implemented in a robot to study about the live feed of the scanned images. More addition of words will increase the understanding knowledge of the system.

7. REFERENCES

[1] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, pp. 1527–1554, 2006.

[2] Y. Freund and D. Haussler, "Unsupervised learning of distributions on binary vectors using two layer networks," Technical Report UCSC-CRL-94-25, University of California, Santa Cruz, 1994.

[3] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in Advances in Neural Information Processing Systems 19 (NIPS'06), (B. Scholkopf, J. Platt, and T. Hoffman, eds.), pp. 153–160, MIT Press, 2007.

[4] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in Advances in Neural Information Processing Systems 19 (NIPS'06), (B. Sch¨olkopf, J. Platt, and T. Hoffman, eds.), pp. 1137–1144, MIT Press, 2007.

[5] M. Belkin, I. Matveeva, and P. Niyogi, Regularization and semi-supervised learning on large graphs," in Proceedings of the 17th International conference on Computational Learning Theory (COLT'04), (J. Shawe-Taylor and Y. Singer, eds.), pp. 624–638, Springer, 2004.

[6] O. Delalleau, Y. Bengio, and N. L. Roux, "Efficient non-parametric function induction in semi-supervised learning," in Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, (R. G. Cowell and Z. Ghahramani, eds.), pp. 96–103, Society for Artificial Intelligence and Statistics, January 2005.

[7] Pardis Noorzad, "Feature Learning from Deep Networks for Image classification," in Computer Vision Seminar.

[8] Ming Zhao, Shutao Li and James Kwok, "Text detection in images using sparse representation with discriminative dictionaries," Elsevier on Image and Vision Computing, 28, 2010.

[9] F. Chen, H. Yu and R. Hu, "Shape sparse representation for joint object classification and segmentation", IEEE Trans. image processing, 22(3):992-1004, 2013.

[10] D. Ciresan, J. Schmidhuber. "Multi-Column Deep Neural Networks for Offline Handwritten Chinese Character Classification", 1 Sep 2013.