RESEARCH ARTICLE OPEN ACCESS

Modelling Framework of a Neural Object Recognition

Aswathy K S*, Prof. (Dr.) Gnana Sheela K**
*(Department of Electronics and Communication, Toc H Institute of Science and Technology, Kerala, India)
** (Department of Electronics and Communication, Toc H Institute of Science and Technology, Kerala, India)

ABSTRACT
In many industrial, medical and scientific image processing applications, various feature and pattern recognition techniques are used to match specific features in an image with a known template. Despite the capabilities of these techniques, some applications require simultaneous analysis of multiple, complex, and irregular features within an image as in semiconductor wafer inspection. In wafer inspection discovered defects are often complex and irregular and demand more human-like inspection techniques to recognize irregularities. By incorporating neural network techniques such image processing systems with much number of images can be trained until the system eventually learns to recognize irregularities. The aim of this project is to develop a framework of a machine-learning system that can classify objects of different category. The framework utilizes the toolboxes in the Matlab such as Computer Vision Toolbox, Neural Network Toolbox etc.

Keywords: Artificial Intelligence, Neural Networks, Computer Vision, Learning, Bag of words, Scale Invariant Feature Transform.

I. INTRODUCTION

Today, machine vision applications crop up in many industries, including semiconductor, electronics, pharmaceuticals, packaging, medical devices, automotive and consumer goods. Machine vision systems offer a non-contact means of inspecting and identifying parts, accurately measuring dimensions, or guiding robots or other machines during pick-and-place and other assembly operations. In the near term, computer vision systems that can discern the story in a picture will enable people to search photo or video archives and find highly specific images. Eventually, these advances will lead to robotic systems able to navigate unknown situations. Driverless cars would also be made safer. However, it also raises the prospect of even greater levels of government surveillance. Two important specifications in any vision system are the sensitivity and the resolution. The better the resolution, the more confined the field of vision. Sensitivity and resolution are interdependent. All other factors held constant, increasing the sensitivity reduces the resolution, and improving the resolution reduces the sensitivity. In many industrial, medical and scientific image processing applications, various feature and pattern recognition techniques are used to match specific features in an image with a known template. Despite the capabilities of these techniques, some applications require simultaneous analysis of multiple, complex, and irregular features within an image as in semiconductor wafer inspection. In wafer inspection discovered defects are often complex and irregular and demand more human-like inspection techniques to recognize irregularities. By incorporating neural network techniques such image processing systems with much number of images can be trained until the system eventually learns to recognize irregularities. Object recognition is nothing but finding and identifying objects in an image or video sequence. Humans recognize a multitude of objects in images with little effort, despite the fact that the image of the objects may vary somewhat in different viewpoints, in many different sizes and scales or even when they are translated or rotated. Objects can even be recognized when they are partially obstructed from view. This task is still a challenge for computer vision systems. Many approaches to the task have been implemented over multiple decades.

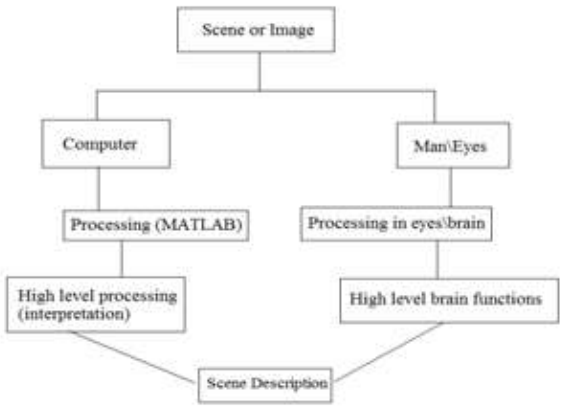


Fig 1.1 Vision - Human vs. Machine

II. LITERATURE SURVEY

Automatically generating captions of an image is a task very close to the heart of scene understanding. This requires, identifying and detecting objects,

people, scenes etc., reasoning about spatial relationships and properties of objects, combining several sources of information into a coherent sentence. Hence it is a complex task to define an image or a scene; which is an important problem in the field of computer vision. Even though it is a challenging one, a lot of research is going on which explores the capability of computer vision in the field of image processing and it helps to narrow the gap between the computer and the human beings on scene understanding. The purpose of this survey is to analyze various techniques used for an image caption generation using the neural network concepts.

Table 2.1 Comparative Analysis on various methods

Author	Year	Method	Remarks
Kelvin Xu et al	2015	Hard attention mechanism and Soft attention mechanism	• Three benchmark datasets: Flickr8k, Flickr30k and MS COCO dataset ; • Evaluated and obtained much better performance than the other methods;
Oriol Vinyals et al	2015	A generative model based on a deep recurrent architecture	• Accurate when verified both qualitatively and quantitatively; • This approach yields 59, to be compared to human performance around 69 and which far better than previous method which shows only a score of 25;
Jimmy Lei Ba et al	2015	An attention-based model for recognizing multiple objects in images	• Used deep recurrent neural network; • More accurate than the state-of-the-art convolutional networks and uses fewer parameters and less computation;
Dzmitry Bahdanau et al	2015	Soft attention based encoder–decoder architecture	• Qualitatively good performance, but lacks in quantitative analysis;
Kyunghyun Cho et al	2014	RNN model	• Qualitatively the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases ; • Maximize the conditional probability of a target sequence given a source sequence;
Jeff Donahue et al	2014	Long-term Recurrent Convolutional Networks for Visual Recognition and Description	• Evaluated on various dataset such as flicker320k, coco2014etc; • Architecture is not restricted to deep NN inputs but can be cleanly integrated with other fixed or variable length inputs from other vision systems;
Junhua Mao et al	2014	Deep Captioning with Multimodal Recurrent Neural Networks	• Validated on Four benchmark datasets : iapr tc-12, flickr 8k, flickr 30k and ms coco; • More improved performance than previous methods;
Andrej Karpathy et al	2014	Deep Visual-Semantic Alignments for Generating Image Descriptions	• Experimented on Flickr8K, Flickr30K and MSCOCO datasets; • Good performance;
Razvan Pascanu et al	2014	Deep Recurrent Neural Networks	• Evaluated on the tasks of polyphonic music Prediction and language modeling; • High performance than conventional RNN;
Bharathi S et al	2014	BoF framework for remote sensing image classification using RANSAC and SVM	• Time complexity of the classification is not very complex; • It took 3mins for a dataset; • One of the best methods for content based image classification;
Chih-Fong Tsai et al	2012	Bag-of-Words Representation in Image Annotation	• One of the most widely used feature representation methods ; • Good in performance;

Misha Denil et al	2011	Learning where to Attend with Deep Architectures for Image Tracking	• Good performance in the presence of partial information;
Siming Li et al	2011	Composing Simple Image Descriptions using Web-scale N-grams	• Viable to generate simple textual descriptions that are pertinent to the specific content of an image;
Yezhou Yang et al	2011	Corpus-Guided Sentence Generation of Natural Images	• Strategy of combining vision And language produces readable and descriptive sentences compared to naive strategies that use vision alone; • Sentences are the closest in agreement with the human annotated ones; • More relevant and readable output;
Stephen O'hara et al	2011	Bag of features paradigm for image Classification and retrieval	• Less quantization errors; • Improved feature detection, and speed up image retrieval;
Xiaoli Yuan et al	2011	A SIFT-LBP image retrieval model based on bag-of-features	• Better image retrieval even In the case of noisy background and ambiguous objects; • Average performance is lower than bof model;
Ahmet Aker et al	2010	Generating image descriptions using dependency relational patterns	• Better higher scores than former n-gram language models; • More readable summary obtained on output;
Juan C Caicedo et al	2009	Histopathology Image Classification using Bag of Features and Kernel Functions	• Tested six different codebook sizes starting with 50 code blocks and following with 150, 250, 500,50 and 1000; • The classification performance decreases while the codebook size increases; • Performance of the sift points decreases faster than the performance of raw blocks; • Sift-based codebook requires less code blocks to express all different patterns in the image collection; • A block-based codebook requires a larger size because it is representing the same visual patterns using different code blocks;
Eric Nowak et al	2006	Sampling Strategies for Bag-of-Features Image Classification	• Interest point based samplers such as harris-laplace and laplacian of gaussian each work well in some databases for small numbers of sampled patches;
Jim Mutch et al	2006	Biologically inspired model of visual object recognition to the multiclass object categorization	• Utilized neural network concepts; • Better in performance than any model without NN concepts.

III. METHODOLOGY

In order to decide about the apt feature to be extracted out of the input image I started ofF with various types of features of an image and experimented and analyzed various methods used to obtain those features. Out of these experiments the bounding algorithm and Bag of features functions were found to be useful for the purpose of this project.

3.1 BOUNDING BOX METHOD

In an image, the edge is a curve that follows a path of rapid change in image intensity. Edges are often associated with the boundaries of objects in a scene. Edge function looks for places in the image where the intensity changes rapidly, using one of these two criteria:

- Places where the first derivative of the intensity is larger in magnitude than some threshold
- Places where the second derivative of the intensity has a zero crossing

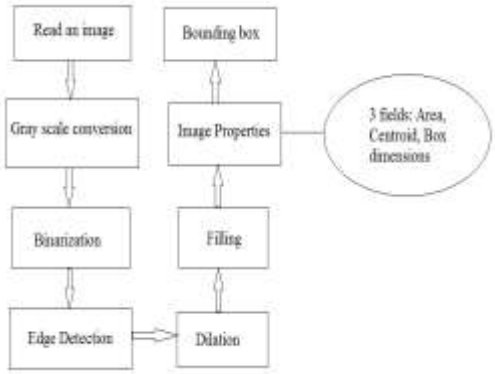


Fig 3.1 Flowchart of bounding box algorithm

The most powerful edge detection method that edge provides is the canny method. The Canny method differs from the other edge detection methods in that it uses two different thresholds (to detect strong and weak edges), and includes the weak edges in the output only if they are connected to strong edges. This method is therefore less likely than the others to be fooled by noise, and more likely to detect true weak edges. Dilation, a morphological operation adds pixels to the boundaries of objects in an image and the number of pixels added to the objects in an image depends on the size and shape of the structuring element used to process the image. The regional properties of the objects in binary image are obtained. The properties include three fields:

Area $A = \sum_{(r,c\in R)} 1$ (1)

Centroid $r = 1/A\sum_{(r,c\in R)} r$ (2)

$C = 1/A\sum_{(r,c\in R)} c$ (3)

Box dimensions

The smallest rectangle containing the region it can be specified by:
– the location of the upper left corner
– the width and height

3.2 BAG-OF-FEATURES METHOD

The bag-of-features (BoF) method is largely inspired by thebag-of-words. In the BoW model, each word is assumed to be independent. In the BoF model, each image is described by a set of order less local features, recent research has demonstrated its effectiveness in image processing. To extract the BoW feature from images involves the following steps:

- automatically detect regions/points of interest
- compute local descriptors over those regions/points
- quantize the descriptors into words to form the visual Vocabulary
- find the occurrences in the image of each specific word in the vocabulary for constructing the BoW feature (or a histogram of word frequencies)

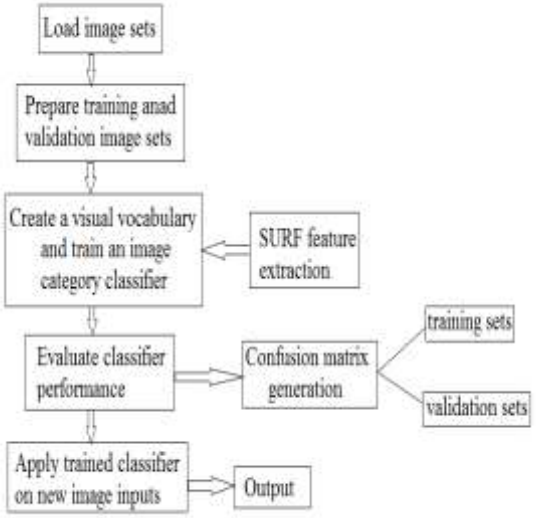


Fig 3.2 Flowchart of BOF algorithm

3.3 Combination of Bounding Box and BoF Method

The bounding box method was used to segment the objects on an image and then provided those objects to the bag of functions to recognize each object. The Figure 6.2 shows the flowchart for this combination of bounding box method and BoF method. Using this combination method I was able to recognize different objects on the same image. Again the degree of the correctness of the output is purely dependent on the images provided to the algorithm.

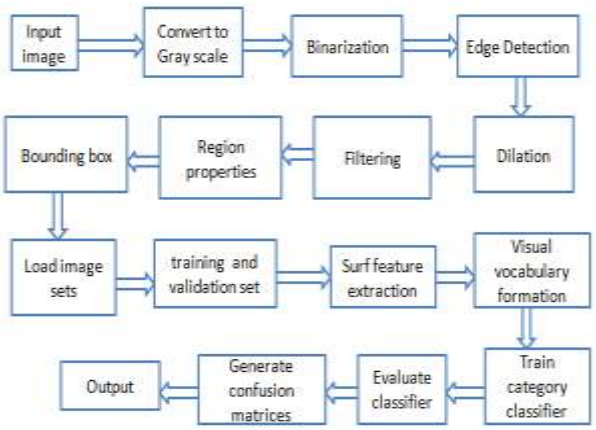


Fig 3.3 Flow chart of combination of bounding method and BoF method

3.4 Scale Invariant Feature Transform (SIFT)

Scale-invariant feature transform (or SIFT) is an algorithm in computer vision to detect and describe local features in images. The algorithm was published by David Lowe. SIFT key points of objects are first extracted from a set of reference images and stored in a database. An object is recognized in a new image by individually comparing each feature from

the new image to this database and finding candidate matching features based on Euclidean distance of their feature vectors. From the full set of matches, subsets of key points that agree on the object and its location, scale, and orientation in the new image are identified to filter out good matches. Once the features are obtained it is provided to the Neural Network Toolbox which utilizes the gradient descent with momentum and adaptive LR training network.

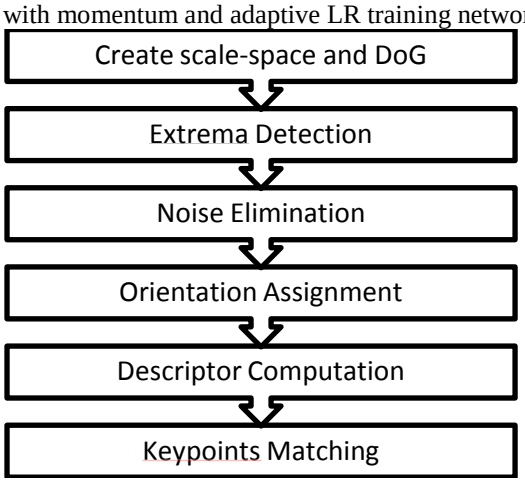


Fig 3.4 Flowchart of SIFT algorithm

IV. EXPERIMENT ANALYSIS

The various methods done are included in Table 4.1. Among these methods the combination of bounding and BoF was better. But the feature vector obtained here is not a constant one. It keeps on changing upon each run command. Hence decided to go for another method called SIFT which extracted the local features of the image and this feature vector was provided to the NN toolbox for recognition of new objects.

Table 4.1 Various methods to analyze feature extraction

Methods	**Purpose**
Analysis of different features	Color, texture, edges, corners, shapes analysis are done on an image set
Using Bounding box	To separate objects on an input image
Harris method	Corner detection based on intensity variations
SURF method	A comparison method to detect an intended portion on a given image
Bag of features method	A technique adapted to computer vision from the world of natural language processing
Combination of bounding box and bag of features	Recognizes different categories of objects on same image
SIFT method	Upon training NN extracted sift feature vectors and recognized the objects on a blank background correctly and generated output correctly.

V. SIMULATION

In order to analyze the performance of various methods experimented, a specific set of datasets are utilized. One dataset include different varieties of fruits. First of all an individual object or a single fruit was provided as input. For eg: an apple. The neural network recognized the given image of fruit correctly. Again different categories of objects were utilized such as chairs, cars, flowers, books etc and each of them were recognized correctly. This model even identifies the objects correctly that are not even present within the dataset. Afterwards my aim was to recognize objects of different categories present on a single image. Thus an image with different kinds of fruits, flowers etc was given as input and each of them were recognized correctly. Some simulation results and plots are included below.

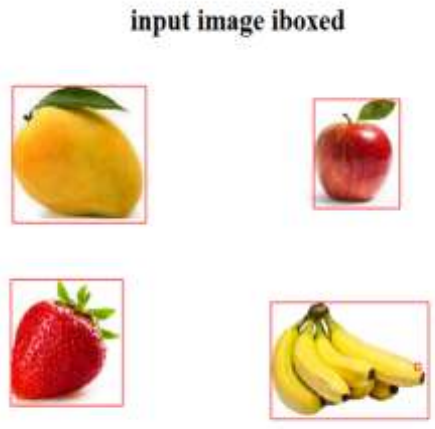


Fig 5.1 Boxed input image

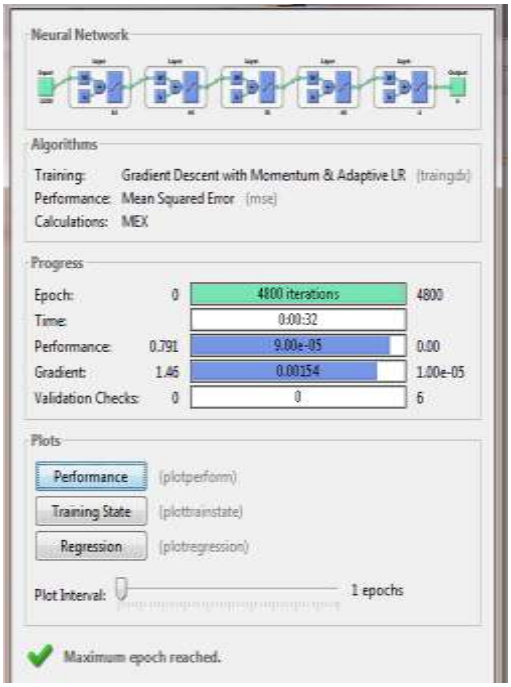


Fig 5.2 the Network used

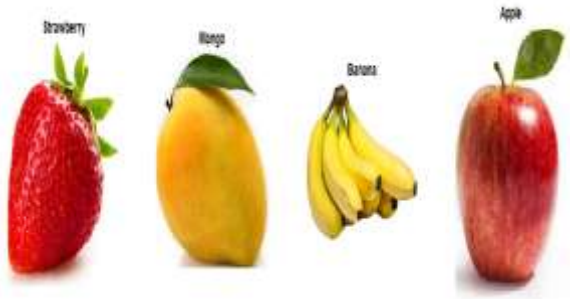
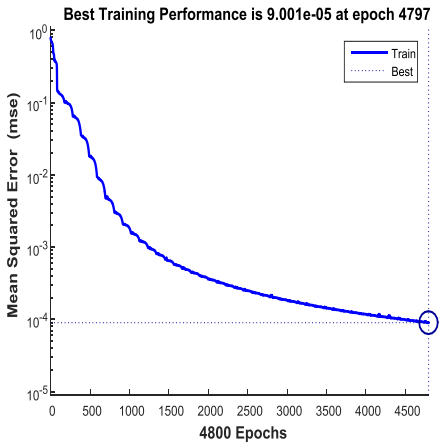


Fig 5.3 Output obtained



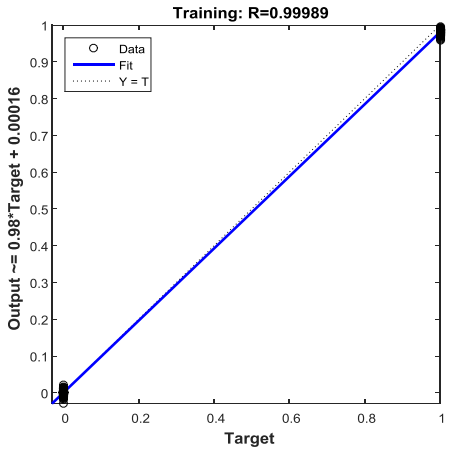


Fig 5.4 Performance Curve

VI. CONCLUSION

The major part lies in the extraction of the correct features of the given input image. Various methods of feature extraction are available. Upon survey it was found that most of the previous methods are concentrating on a single feature alone, which would not aid for my purpose. Hence after working on various available methods the SURF features were found to be better as it is independent of the scale and orientation of an image. But still it didn't serve my purpose. Thus decided to choose another feature extraction process called Bag-Of-Visual words, which is the better one so far. Finally utilizing the bounding method to identify objects in a single image and applied to the BoF method to recognize each of them. But still the presence of neural networks is not there as the feature matrix obtained out of BoF is not a stable one. Hence utilized the SIFT method - as the name indicates a method independent of scale and rotation changes.

REFERENCES

[1] Jim Mutch, David G. Lowe, "Multiclass Object Recognition with Sparse, Localized Features", *In Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06),* 2006.

[2] Eric Nowak, Frederic Jurie, Bill Triggs, "Sampling Strategies for Bag-of-Features Image Classification", Springer-Verlag Berlin Heidelberg, ECCV Part IV, LNCS 3954, pp. 490–503, 2006.

[3] Juan C. Caicedo, Angel Cru, Fabio A. Gonzalez, "Histopathology Image Classification using Bag of Features and Kernel Functions", *Bioingenium Research Group, National University of Colombia,* 2009.

[4] Ahmet Aker, Robert Gaizauskas¸ "Generating image descriptions using

dependency relational patterns", *In Proc of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1250–1258, Uppsala, Sweden*, July 2010.

[5] Xiaoli Yuan, Jing Yu, Zengchang Qin, "A SIFT-LBP image retrieval model based on Bag-Of-Features", *18th IEEE International Conference on Image Processing*, 2011.

[6] Stephen O'hara AND Bruce A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval", arXiv: 1101.3354v1 *[cs.CV]*, January 2011.

[7] Yezhou Yang , Ching Lik Teo, Hal Daume, Yiannis Aloimonos, "Corpus-Guided Sentence Generation of Natural Images", *In Proc of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 444–454,Scotland, UK*, July, 2011.

[8] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi, "Composing Simple Image Descriptions using Web-scale N-grams", *In Proc CoNLL'*, 2011.

[9] Misha Denil, Loris Bazzani, Hugo Larochelle, Nando de Freitas, "Learning where to Attend with Deep Architectures for Image Tracking", *arxiv:1109.3737*, September 2011.

[10] Chih-Fong Tsai, F. Camastra, "Bag-of-Words Representation in Image Annotation: A Review", *International Scholarly Research NetworkISRN Artificial Intelligence*, 2012.

[11] Bharathi S, Karthik Kumar S, P Deepa Shenoy, Venugopal K R, L M Patnaik, "Bag of Features Based Remote Sensing Image Classification Using RANSAC And SVM", *In Proceedings of the International Multi Conference of Engineers and Computer Scientists Vol I, IMECS*, March 2014.

[12] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Yoshua Bengio, "How to Construct Deep Recurrent Neural Networks", *arXiv: 1312.6026v5 [cs.NE]*, April 2014.

[13] Andrej Karpathy, Li Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions",*cs.stanford.edu/people/karpathy/deepimagesent/StanfordUniversity*, 2014.

[14] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan L. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)", *Published as a conference paper at ICLR 2015*,July 2014.

[15] Jeffrey Donahue, Lisa Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell "Long-term Recurrent Convolutional Networks for Visual Recognition and Description", *University of California at Berkeley, Technical Report No. UCB/EECS-2014-180*, November 2014.

[16] Kyunghyun Cho, Bart van, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation", *arXiv:1406.1078v3 [cs.CL]*,September 2014.

[17] Dzmitry Bahdanau, Kyung Hyun Cho, Yoshua Bengio, "Neural machine translation by jointly learning to align and translate", *arXiv: 1409.0473v6 [cs.CL],Google, Published on ICLR*, April 2015.

[18] Jimmy Lei Ba, Volodymyr Minho, Koray Kavukcuoglu, "Multiple object recognition with visual attention", *arXiv: 1412.7755v2 [cs.LG], Google*, April 2015.

[19] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator", *arXiv: 1411.4555v2 [cs.CV] , Google*, 2015.

[20] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", *In Proc of the 32nd International Conference on Machine Learning, France, JMLR: W&CP volume 37*, February 2015.