A Neural Network Architecture for Semantic Segmentation in Real Time Applications

V G Hamsaveni[1], Sreekaram Jagadeesh[2], S Mohan Rao[3],
[1]*Associate Professor*, [23]*Assistant Professor*,
[12] *SITAMS (AUTONOMOUS)*, [3]*SVCE JNTUA University*

Abstract- Interests in augmented reality wearables, automated devices used in home, self-driving vehicles are in high demand, which uses semantic segmentation. The methodology includes the consideration of each and every pixel i.e. every pixel may be belongs to any one category and can be operated in real-time applications on less-power mobile devices. Although, huge availability of data sets and many machine learning algorithms outstrip the performances of this kind of applications but as a penalty in floating point operators which increases the running time. In order to classify the spatial images along with the meaningful segmented regions numerous Neural Networks are proposed, includes SegNet or Fully CNN for multiple class classification. In this work, we used the newly introduced model ENet also called as Efficient Neural Network, which is specially designed for task requiring low latency operations. This model is 18x times faster, used less number flops i.e. 75x less and uses less number of parameter i.e.79X less. The proposed method uses Cityscapes database for the experiments and the results are compared with other conventional techniques. We also shown, the performance measurements using ENet on embedded systems, which is highly required for improving the software's that could make ENet faster. We have not used any post processing, since it may reduce the performance of CNN, but however you may include as one of the step to get even more accurate results.

Keywords- *Segmentation, Semantic Segmentation, Weakly supervised segmentation, neural networks.*

I. INTRODUCTION

In the domains such as image processing and computer vision, semantic segmentation is one of the major applications. It is being used in many other domains like medical and intelligent transportation. Many of the specified data sets are given to the researchers to examine their paradigms. This semantic segmentation is being studied from some decades. The arrival of Deep Neural Network (DNN), this segmentation made a great progress. Based on computer vision's previous history, a key challenge is a Semantic Segmentation which is able to perform segmentation process on any image. In this process it partitions an image into multiple sections and as well as objects such as ocean, cat, and person [1]. Additionally, segmentation is somewhat depth when compare to objects detection since detection is need not to be used in segmentation. Particularly, humans perform segmentation on an image by not knowing about the objects. It is a key part to understand the process visually which can yields a robust model for getting understand community and can also be utilized to enhance the prevailing techniques of computer vision [2].

Currently the computer vision domain is facing semantic segmentation problem. In broader case, this problem is high-level action that covers the way to understand the total situation. The significance behind understanding the situation regarding the issue in the computer vision is spotlighted by a fact that increase in count of applications leads to attain knowledge on imagery. Few of the applications comprises of vehicles with self-driving, interaction between the human and a computer and so on. With the significance of deep learning most of the semantic segmentation related issues are being dealt with the application of deep architectures, frequently Convolutional Neural Nets, which can extend the remaining models greatly with good accuracy rate and efficiency [3]. The process of grouping the image partitions of similar object class together called as Semantic Segmentation. Such kind of paradigm has many use cases like finding roads signs, identifying tumors [4], and finding medical tools in surgery [5], colon based cryptographic segmentation [6]. Many segmentation applications in medical domain are sorted in [7]. In the other hand, non-semantic segmentation only groups the image pixels together on the basis of general features of individual objects. Therefore a non-semantic segmentation is undefined properly like other. To have a concept of semantic segmentation is somewhat advantageous while attempting to track the instances of objects. During this it leads to occur two issues and those are a) the neighboring pixels of similar class may belongs to heterogeneous instances of objects and the regions are not belonging to the similar object instance. Semantic segmentation is so powerful compared to traditional segmentation. The main use case of semantic segmentation is self-driving cars, which we have chosen in this project. The other applications include Geo Sensing, Autonomous driving, Face segmentation, Fashion i.e. parsing cloths, Precision farming. The main idea of this paper is recognizing, understanding what's in the image at Pixel level shown in Figure 1. You may get the clear idea. The conventional algorithm just do the segmentation either binary or multi-segmentation, but won't label the classes of different categories. The input for this algorithm is Image or Video Stream and the output is the image with multiple regionsregions with different classes.observe the sample in the following Image. The summary of our paper is given here. In section 2, we are going to discuss the traditional works of the semantic segmentation to recent works which are used neural networks for identifying the objects in real time. Section 3 focused on identifying the features used for segmentation

pipeline for getting the segmentation results along with the methods used. In section 4, we have presented the deep learning work using E-net architecture for doing the segmentation task. Section 5 concludes the paper.

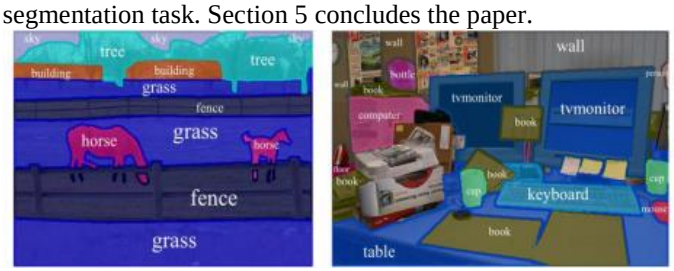


Fig.1: Semantic Segmentation

II. RECENT WORKS

It is highlighted that to make weakly-supervised learning available to the semantic segmentation [8]. It is specifically important, training set acquisition by manual labelling of images at a pixel level predominantly highly cost oriented that allocating class labels at an image level. Earlier approaches of segmentation used challenged annotations in many forms: enclosed things around the objects [9], labels og images in describing the existence of a type or combination of both. In general, extra cues will arrive easily to describe this challenging issue. In [10], movement is one of the cues to semantic segmentation, which enables us to find the objects extension together with its exteriors in a scene with high accuracy. In this paper work, we are intended to cover the gap through learning the accurate model of segmentation that helps in cues movements obtained from the videos which are weakly-annotated. Besides to the fully-supervised segmentation techniques like [11], many weakly-supervised approaches are proposed from many years. Few of those use bounding boxes [12] whereas others depend on the labels of an image. Conventional models to this task apply different kinds of visual features such as SIFT histograms, texture, integration of graphical and/or parametric modelled patterns and color. Such recent applications are out performed by a method called FCNN. Pathak et al. [14] elaborated the framework of MIL which is specialized to identify the objects and perform segmentation over it by considering a pixel with greater prediction outcome as its related sample during evaluation of loss. An alternate to this MIL approach is to introduce an aggregate function which converts pixel-level predictions of FCNN to a distribution of image labels. Practically this strategy works better than [14], but needs images under training which comprises of individual object and also clearly stated background images.

In [13], Weakly-supervised FCNNs defines boundaries over the forecasted pixel labels. Papandreou et al. [15] introduced an approach known as Expectation Maximization (EM). This approach alternates between pixel labels prediction (E-step) and forecasting the parameters of FCNN (M-step). Here in E-step there required minimum 20% of image pixels and these must be allocated to each of the image-level types and about 40% for the background. Both of these approaches yields best outcomes over a dataset of VOC 2012. We stated this constraint in our M-CNN framework. Methods following this recent trend [16] are kick-started with either a small number of manually annotated examples. Other algorithms concerning with the weakly-supervised learning like Co-localization and co-segmentation needs a video or an image to have a dominant object class.

The aim of Co-localization approaches is to make similar object available locally with the bounding boxes. In contrast, the objective of co-segmentation is estimating the segment labels pixel-wise. Such approaches [17] rely on already manually computed group of regions and select the optimal one with the application of optimization process. In [19], the researchers recommended M-CNN framework where they took a single frame from a sample video. The soft potentials (foreground appearance) computed from motion segmentation and the FCNN predictions (category appearance) together determined the latent segmentation (inferred labels) to evaluate the loss, and so that network get updated. The structure of their study is presented in Figure 2. We have also shown the architecture semantic segmentation from natural language expression in Figure 3.

III. BASIC STRUCTURE

The models of semantic segmentation are simply designed and represented in this paper. We are also presenting total state-of-art methods. So that it enables us to implement the rest of the models. Since almost those all have similar underlying infrastructure, settings and flow. Current state-of-art methods

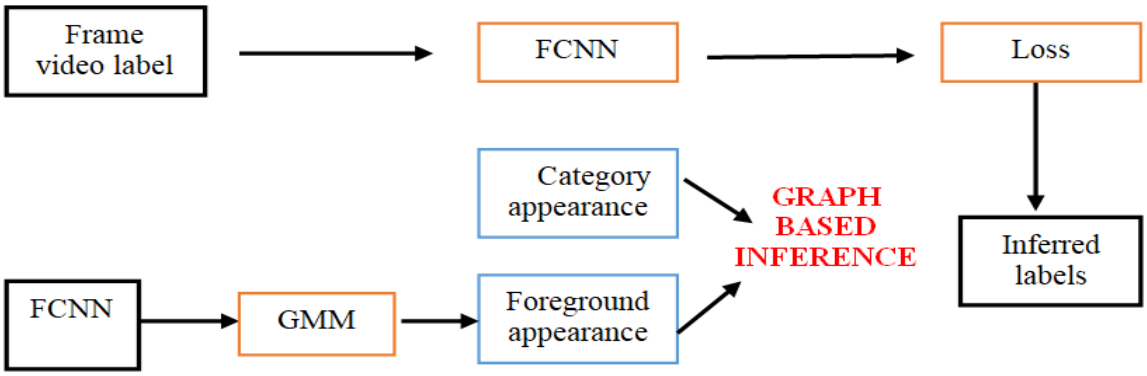


Fig.2: Overview of M-CNN framework.

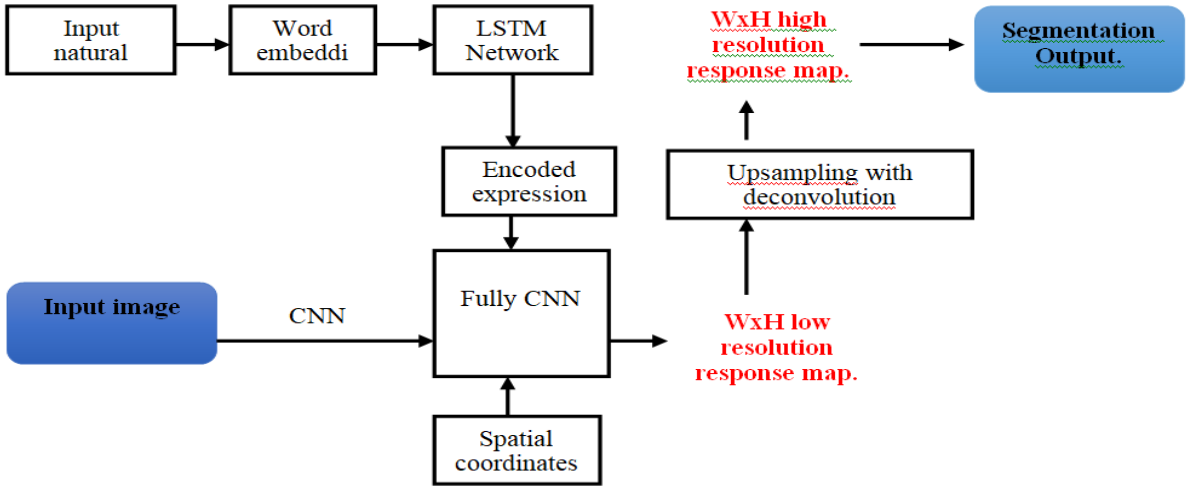


Fig.3: Segmentation from natural language expression.

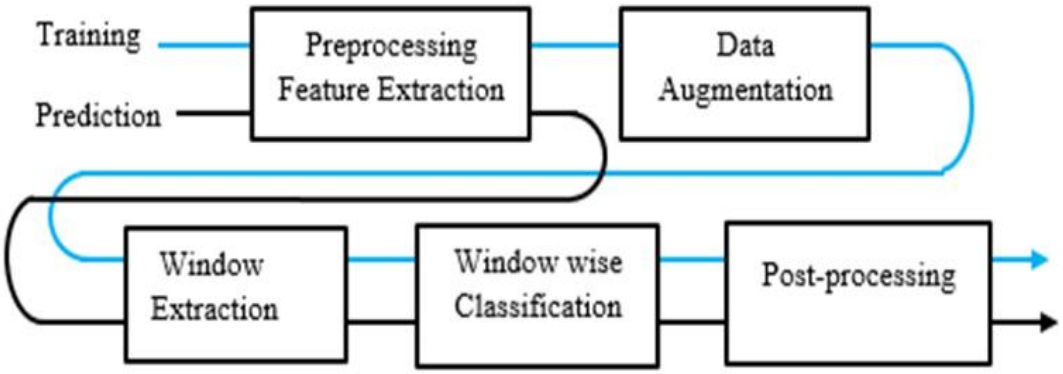


Fig.4: Basic Structure of Semantic Segmentation

Followed the similar scheme of extracting the features along with multi-scale processing. So that many becomes easier in implementing and training from end-to-end. Your choice of usage relies on your requirement for accuracy or speed or memory. Figure 4 shows the representative segmentation pipeline of conventional architecture

A. Input Data

The existing information that is used for segmentation inference changes by application. **Grayscale vs colored -** The images belonging to Grayscale are usually used in the medical imaging like MR scanning whereas the colored images are extended obviously. **Single image vs stereo images vs Co-segmentation-** The widely extended type of segmentation is Single image segmentation. By the usage of stereo images is boredom. It can be said as the most natural process of segmentation and it also said as it is concerned to have data in depth. Co-Segmentation is said to the issue during the detection of consistent segmentation of many images. We can see this problem in the two ways. One way is as the problem in detecting the similar objects from at least two images. Another way is as problem in adding an extra source of data to keep as a meaningful segmentation by adding each image one after the other. **2D vs 3D -** The process of segmenting the images is a task of 2D segmentation in which the minute part is known as a pixel. 3D data represents as X-ray CT images the way they used.

B. Features and Pre-processing methods

Some of the algorithms related to image segmentation will use conventional approaches so better not to apply the neural networks instead use domain data. In the following section, we presented our work's experimental study by applying neural networks. In the context of conventional models, feature selection plays an important role. The features which are frequently used either locally or globally or both are described in the below and also the reduction algorithms for feature dimensionality. **Pixel Color -** The extensively used features in terms of pixel colors of heterogeneous image spaces are 3 features for RGB, 3 features for HSV, 1 feature for the gray-value. An individual image is in RGB color space but on the basis of classifier and an issue another color space may yields better outcome. **Histogram of oriented Gradients (HOG) -** Features of HOG interprets an image as a discrete function I: $N2 \rightarrow \{0... 255\}$ which maps the position (a, b) to color. For every pixel, there will be two gradients: the partial derivative of a and b. Now the actual image can be transformed in to two feature based maps with the same size that denotes the

gradient. **SIFT -** These feature descriptors depicts important pixels in an image which is 16x16 patch size and a key point around it can be taken. This patch is partitioned in to 16 different regions with 4x4 sizes. For every part of those, a histogram with 8 orientations is evaluated in a same way as for the HOG features. It yields a 128-dimensional feature vector for every key point. **Textons -** A Textons is a basic building block of vision. The study of computer vision do not stated a better definition for Textons, but the edge detectors can be said as one example. One may argue that the deep learning approaches with Convolution Neuronal Networks (CNNs) learn Textons in the first filters.

C. Segmentation Algorithms

Clustering Algorithms - These can be directly applied on pixels. Two clustering paradigms are k-means and mean-shift algorithm. The former one is a general-purpose clustering algorithm which needs the set of clusters given earlier. Firstly, it positions a k centroids in a feature space randomly. Later it allocates every data point to a neighboring centroid and moves that centroid towards the center of cluster and proceeds the process till the desired criterion get satisfied [20]. **Graph Based Image Segmentation -** These types of algorithms interprets the pixels like vertices and also weight of edge is said as a measure of difference in the color [21]. **Random Walks-** This belongs to graph-based image segmentation paradigms. This type of image segmentation generally works as: seed pixels are located in image for varied objects in that. Through from each of the individual pixel, there might be a chance to meet various seed pixels with a random walk is evaluated. **Watershed Segmentation -** A watershed algorithm considers an image with Grayscale and interprets that image as height map. Low values are said as catchment basis and higher values that exists in between the dual neighboring catchment basins is the watershed. This represents that such regions must be kept dark on a Grayscale images. This algorithm initiates to complete the basins from least point. When the two basins are met, a watershed is detected. This algorithm quits when it meets the greatest point. **Random Decision Forests-** Firstly these are proposed in [22]. Such kind of a classifier uses some approaches known as ensemble learning where the training of various classifiers done along with their hypothesis integration. An ensemble learning approach is method of random subspace in which every classifier's trained over random subspaces of feature space. Bagging is another technique of ensemble learning. Here concerning Random Decision Forests, the classifiers are treated as decision tree. **SVMs -** These are said as well-examined classifiers which could be depicted by the five central notions. For those notions, the representation of training data is as (X_i, Y_i) where X_i is a feature vector and $Y_i \in \{-1, 1\}$ the binary label for training example $i \in \{1... n\}$.

D. Neural Networks

ANN (Artificial Neural Networks) is the classifiers which are derived from the biologic neurons. Every individual artificial neuron comprises of inputs which are weighted and also summed up. Later the neuron attempts to apply activation function to that weighted sum and yields an output. Those neurons could consider either a featured vector as an input otherwise as an output of the rest of the neurons. In such a way, they can construct feature hierarchies. Many neural networks related ideas on regularization, best reduction algorithms, and instant generation architectures and so on. The detailed explanation of it is not discussed here. But some of the key breakthroughs are highlighted. There will be major reduction in the parameters which are to be learned during the problem in the images field. It was stated by Alex Krizhevsky et al. in [23]. An important notion is a clever regularization is known as dropout training, which keeps the neurons outputs randomly to zero during training.

IV. EXPERIMENTAL WORK

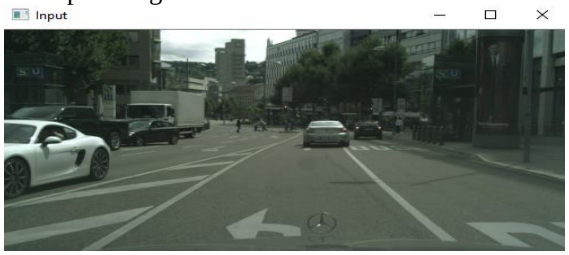
In our project, we are using ENet architecture of deep learning to do the semantic segmentation. This can be useful to apply for both images as well as for videos. The main advantage of ENet is that, it is 18 times faster, requiring only fewer parameters and giving better accuracy compared with other models. The size of the model is just around 4 megabytes only. Coming to the execution time, one pass may took 0.2 seconds in CPU and it will be even faster if you use the GPU for the same. The other architectures, which can do the same task are Alex Net, VGG-16, Google Net and ResNet. This data set includes examples of images, which can be used for urban scene understanding, including self-driving vehicles. This model is trained on the examples classes such as Road, Side walk, Buildings, Pole, Traffic Light, Train, Bus etc. (Contains around 20-30 classes). We used to support all the classes in a text file and will proceed further. The packages numpy, argparse, imutils, time and OpenCV are used in this work. ENet architecture uses very few parameters, required very less space around 1 MB of memory. This makes to fit the whole network and is so fast, used in on-chip embedded processors. The major steps include in semantic segmentation are classification, localization or detection and finally semantic segmentation. Here the localization provides some extra information about the spatial location of the classes. The method can be applied for both images and real time videos. Here I have shown, just one sample result for input image along with segmented image in Figure 5. In addition, the clustered colors or legend colors are also shown in Figure 6. In Figure 7, I have shown some sample frames in the video which is used in self-driving of car.

V. CONCLUSION

Semantic Segmentation is crucial to analyses the content in the input image and to detect the objects in the given image. The Difference between segmentation and semantic segmentation is that, Conventional algorithms used to segment the given images in to different parts for example, normal graph cuts, super pixels etc., but without knowing what each part denotes. The main use case of semantic segmentation is self-driving cars, which we have chosen in this project. The other applications include Geo Sensing, Autonomous driving, Face segmentation, Fashion i.e. parsing cloths, Precision farming. In this paper, we discussed the traditional works of the

semantic segmentation to recent works which are used neural networks for identifying the objects in real time. Further, we also focused on identifying the features used for segmentation pipeline for getting the segmentation results along with the methods used. Next, we have presented the deep learning work using E-net architecture for doing the segmentation task. We also, included the results of E-net Architecture for automated driving vehicles. Improving the efficiency and performance of these algorithms are considered as our future work.

a. Input Image



b. Segmented Image

Fig.5: Semantic Segmented Result

Fig.6: Legend Colors

Fig.7: Semantic segmented Frames from Self driving car

VI. REFERENCES

[1]. Thoma M (2016) A survey of semantic segmentation. arXiv preprint arXiv:1602.06541 VOC2010 preliminary results. http://host.robots.ox.ac.uk/pascal/VOC/voc2010/results/index.html.

[2]. WuZ, Shen C,Hengel A(2016a) High-performance semantic segmentation using very deep fully convolutional networks. arXiv preprint arXiv:1604.04339.

[3]. Xu J, Schwing AG, Urtasun R (2015) Learning to segment under various forms of weak supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3781–3790

[4]. N. Moon, E. Bullitt, K. Van Leemput, and G. Gerig, "Automatic brain and tumor segmentation," in Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002. Springer, 2002, pp. 372–379.

[5]. G.-Q. Wei, K. Arbter, and G. Hirzinger, "Automatic tracking of laparoscopic instruments by color coding," in CVRMed-MRCAS'97, ser. Lecture Notes in Computer Science, J. Troccaz, E. Grimson, and R. Mösges, Eds. Springer Berlin Heidelberg, 1997, vol. 1205, pp. 357–366. [Online]. Available: http://dx.doi.org/10.1007/BFb0029257.

[6]. C. Huang, L. Davis, and J. Townshend, "An assessment of support vector machines for land coverclassification,"InternationalJournalofremote sensing, vol. 23, no. 4, pp. 725–749, 2002.

[7]. D. L. Pham, C. Xu, and J. L. Prince, "A survey of current methods in medical image segmentation," Annual Review of Biomedical Engineering, vol. 2, no. 1, pp. 315–337, 2000,

pMID: 11701515. [Online]. Available: http:// dx.doi.org/10.1146/annurev.bioeng.2.1.315.
[8]. 3. Pinheiro, P.O., Collobert, and R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR. (2015).
[9]. Wu, J., Zhao, Y., Zhu, J., Luo, S., Tu, Z.: MILCut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In: CVPR. (2014)
[10]. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV. (2010).
[11]. Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: ECCV. (2012).
[12]. Monroy, A., Ommer, B.: Beyond bounding-boxes: Learning object shape by modeldriven grouping. In: ECCV. (2012.
[13]. Pathak, D., Kra¨henbu¨hl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: ICCV. (2015).
[14]. Pathak, D., Shelhamer, E., Long, J., Darrell, T.: Fully convolutional multi-class multiple instance learning. In: ICLR. (2015).
[15]. Papandreou, G., Chen, L.C., Murphy, K., Yuille, and A.L.: Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. In: ICCV. (2015).
[16]. Divvala, S.K., Farhadi, A., Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning. In: CVPR. (2014).
[17]. Joulin, A., Tang, K., Fei-Fei, L.: Efficient image and video co-localization with Frank-Wolfe algorithm. In: ECCV. (2014).
[18]. Liang, X., Liu, S., Wei, Y., Liu, L., Lin, L., Yan, S.: Towards computational baby learning: A weakly-supervised approach for object detection. In: ICCV. (2015).
[19]. Pavel Tokmakov, Karteek Alahari, Cordelia Schmid. Weakly-Supervised Semantic Segmentation using Motion Cues. ECCV-European Conference on Computer Vision, Oct 2016, Amsterdam, Netherlands. Springer, 9908 (Part IV), pp.388-404, 2016
[20]. J. A. Hartigan, Clustering algorithms. John Wiley & Sons, Inc., 1975.
[21]. P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," International Journal of Computer Vision, vol. 59, no. 2, pp. 167–181, 2004. [Online]. Available: http://link.springer.com/article/10.1023/ B: VISI. 0000022288.19776.77
[22]. T. K. Ho, "Random decision forests," in Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, vol. 1. IEEE, 1995, pp. 278–282. [Online]. Available: http://ect.bell-labs.com/who/ tkh/publications/papers/odt.pdf.
[23]. A. 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.

Author's Profile –

Mrs. V G Hamsaveni, currently working as Associate Professor at SITAMS (Autonomous), JNTU_A University, Anantapuram. She received her master's degree in Energy Management from SV University in the year 2005. She received her Bachelor's degree in EIE from JNTU University in the year 2002. She has published various national and international journals. Her research interests are Nano technology and image analytics, and image enhancement.

Mr. S. Jagadeesh, currently working as Assistant Professor at SITAMS (Autonomous), JNTU_A University, Anantapuram. He received his master's degree in VLSI Design from JNTU_A University in the year 2015. He received his Bachelor's degree in ECE from JNTU_A University in the year 2011. He has published various national and international journals. His research interests are VLSI systems and image analytics, and image enhancement.

Mr. S. Mohan Rao, currently working as Assistant Professor at SVCE, JNTU_A University, Anantapuram. He received his master's degree in VLSI System Design from JNTU_H University in the year 2013. He received his Bachelor's degree in ECE from JNTU University in the year 2009. He has published various national and international journals. His research interests are image analytics, and image enhancement and VLSI systems.

-