An Optimised Approach on Object and Text Detection from Real Time Data using Histogram Equalization Technique

Monika Kapoor
Research Scholar
Department of Computer Science,
Sri Sai University, Palampur, India

Saurabh Sharma
Assistant Professor
Department of Computer Science
Sri Sai University, Palampur, India

ABSTRACT
The initiation of artificial intelligence, gave a curious approach towards the artificial neural networks. The developers are trying to make computers think and perform tasks just like humans. Artificial Neural Networks was developed with the view in mind to make computers to do so. The machines that are to be used in the field of robotics, medicines, industry needs to be so smart that they should be able to perform the day to day tasks easily without needing the human help. Take the case of self-driven cars, it is very crucial for the AI to make out the scenario in the real time and act accordingly take the necessary actions like breaking or evading in the traffics. Since the introduction of the object detection by the AI Researchers at Facebook, it has become easier for us to identify the objects in the images. While the object identifiers can use the image identification in the still images but there is a need to identify the text in the images. The techniques can be used for identification of the objects and texts in the real time scenario also. The present work here shows that image identification can work not only for the object identification but also identifies the text in the images as well. A CLAHE algorithm is developed to identify these objects in the images and text and classify these entities into the categories as desired by the programmer. Here in this paper an attempt is made to show the working of an algorithm based on the EAST text detector and PMTD that can perform object as well as text identification in the real time itself. The algorithm achieved success in achieving the identification of the objects in under 5 seconds and better yet identifying text in under 1.5 seconds. The developed algorithm outperforms its earlier precursors. The Recall, Precision and F-measure values all were found better than both the previous algorithms.

Keywords
ANN, AI, PMTD, EAST, CLAHE, Text Classification, Image Identification, Object Classification.

1. INTRODUCTION
The chief aim of the Computer recognition is the recognition of visual scenes. Scene recognition includes various understanding of which items are there, localization of the 2D and 3D objects, determining what they are and scene characteristics, relation that exists between them and describing the scene. The present research paper helps us to develop an understanding of the objects and the relation that exists between them.[33],[36],[3],[5]

Take for example, the ImageNet database [6], that contains a large set of images has enabled breakthroughs in object and text identification and classification [22],[8],[9]. ImageNet has also created a dataset that consists of objects of common characters [19], key points [14], and 3D scene information [35]. That has led to a question: what are the datasets that will best suit a purpose that will guide towards the better understanding of the scene?

Fig 1: Object detection for detecting people and baggage claim in an airport [15]**.**

In the current research explores a new vast-scale dataset, and real time data, that will be addressing the core research problems in scene recognition detecting non-iconic views [28] of objects, contextual reading among the objects and the accurate 2D location of objects. Different groupings of objects, there will occur an iconic view. For instance, while performing an image search on internet like that of a "bike" the most searched bikes will occur on the top while the rest will follow. The proposed search algorithm performs very well for iconic views but it will be difficult for the algorithm to recognise the objects in other cases like that when there is occlusion in the background fogy conditions etc. [15] showing the composition of real everyday scenes. Here in this work it checks this work experimentally with an algorithm called as EAST text detection [4]. It is an accurate, faster object text detector and uses only two stages in its working. The working of this algorithm uses a fully convolutional network (FCN) model. FCN produces text in line predictions and will make the word appear in the text boxes. It reduces the redundant steps involved that were involved in the previous models. The produced text can be oriented in any direction based on the user requirements. On comparing the EAST to its precursors, it achieved significant improvements both in speed and qualitative and quantitative aspects, that when the images are

evaluated on everyday scenes the trained model whether outperforms or underperforms the algorithm. The primary objective is finding the real images that have multiple objects. The identity of different objects can only be solved using background, due to little pixel size or distorted appearance in the picture. To really take the algorithm to the limit it is necessary for the checking images in the real world situation rather than identifying the individual objects in isolation [16]. Finally, discussing the detailed spatial understanding of object spread will be a core component of scene analysis. An object's location in the array can be defined coarsely using a bounding box [17] or with an accurate pixel-level differentiation[20],[10],[22]. To measure any kind of localization it is essential for any dataset to have every case considered, labelled and fully differentiated. Here used Microsoft COCO dataset for training the algorithm. The dataset is unique in its annotation of instance-level segmentation masks.

Content-based multimedia database indexing and retrieval tasks require auto grabbing of the described feature that are related to the subject materials (images, video, etc.). The characteristic low-level features that are grabbed in images and video contains measures of colour [34], texture [31], or shape [11]. Though these topographies can easily be taken, they will not provide an actual data of the image content. Taking more descriptive features and higher-level contents, such as text [26] and human faces [23], is currently grabbing a growing interest in the research. Text contained in images and video, chiefly captions, provide quite a few but important content information, like that the name of players or speakers, the title, location, date of an event, etc. This text can be a keyword resource as powerful as the information provided by speech recognizers.

Text detection and recognition in images and video frames, that is provided by the optical character recognition (OCR) and text-based searching methodologies, is now identified as a chief part in the making of advanced image and video annotation and retrieval systems. But the grey scale imaging used in the images becomes to be troublesome for the OCR. Therefore, correct identification and segmentation of text characters present in the background is important to fill the gap between image and video documents and the input of a standard OCR system. Previously, proposed methods can be classified into bottom-up methods [29],[18],[7] and top-down methods [18], [27], [30], [2]. Bottom-up methods segment images into regions and then group "character" regions into words. The recognition performance relies on the segmentation algorithm and the complexity of the image content. Top-down algorithms first detect text regions in images and then segment each of them into text and background. The top down approach is faster and more precise as compared to the bottom up approach.

CLAHE [1] on the other hand is a histogram Equalisation method that is used to enhance the image information based on the contrast improvement and reducing the noise in the image. In this method the contrast enhancement is done in a cumulative manner where the adjacent pixels share the values with each other. A slope transformation enhance feature is required in the CLAHE method. This method limits the scatter of the pixels over a large area and concentrates the scatter to a relatively smaller region that can be read by the image procession software. CLAHE will limit the magnification by chopping the histogram at a pre-set value before calculating the CDF (cumulative distribution function). This limits the slope of the CDF and therefore of the transformation function. The mean value beyond a certain limit applied by the user also known as the clip limit, depends on the normalization of the histogram and thereby on the size of the neighbourhood region. Commonly the amplification is in the order of 3- 4 times the previous value.

2. RELATED WORK

Text can be detected by manipulating the distinct characters of text characters like that of the vertical edge density, the texture or the edge position variance. Previous approach of localizing text in covers of Journals or CDs [26] considered that text characters were in regions of more horizontal variance that would satisfy some spatial properties which can be broken in a connected component analysis process. Smith et al. [24] localized text by first identifying vertical edges in an already present model, then clustering vertical edges into text regions using a smoothing algorithm. These two methods were quick but also gave several faulty warnings because many background regions contained strong horizontal contrast. The method of Wu et al. [30] for text localization is based on texture segmentation. Texture features were computed at every pixel from the derivatives of the image at several scales. Using a K-means algorithm, pixels are classified into three classes in the feature space. The class with highest energy in this space indicates text while the two others indicate non-text and uncertainty. But the segmentation was very highly attentive to the background noise and image content and the feature extraction is computationally expensive. More recently, Garcia et al. [27] proposed a new feature referred to as variance of edge orientation. This relies on the fact that text strings contain edges in many orientations. Variation of edge orientation was computed in local area from image gradient and combined with edge features for locating text blocks. The method, however, may exhibit some problems for characters with strong parallel edges characteristics such as "i" or "1".

Besides the possessions of individual characters, Sobotka et al. [7] tells that baseline detection can be used for text string localization. More precisely, printed text strings are categorized by detailed top and bottom baselines, which can be detected in order to evaluate the presence of a text string in an image block.

As an alternative, a few systems considered machine learning tools to perform the text detection [29],[15],[7],[27],[30],[26]. These systems extracted wavelet [37] or derivative features [26] from fixed-size blocks of pixels and classified the feature vectors into text or non-text using artificial neural networks. Since the neural network-based classification was applied to all the possible positions of the whole image, the detector algorithm was not able to perform correctly and produced several false negative values.

3. CURRENT WORK

In the existing system the researchers worked on different classifier techniques such as PMTD, EAST text detectors, MASK RCNN among which ANN is the most efficient one with the help of which the accurate prediction can be made. However, there is still some lacking in the field of text detection. As these do not provide the efficient results for every type of data set that will affect the detection process.

Many researchers have been training and testing their classificatory on the well-known text detection datasets and others have used information from affordable regions.

Finally, machine-learning can actually be used in the field of healthcare, automobiles and many more. Most scientists have

their datasets from the same of the earlier researches. Also, these classification algorithms can be further improved by increasing the number of attributes in it with the help of which more accurate prediction can be done. Many possible improvements to enhance the scalability and precision of this forecast scheme could be studied.

4. PROPOSED WORK:

As per the statement defined in the aforementioned section, it is observed that there is a provision to update the existing system. In the existing system the existing classifiers are used for accurate prediction. However, it was observed that an alone classifier is not able to produce best result for every dataset and not a single algorithm can give consistent results for all types of text identification detection data.

Thus, an algorithm is required which can provide an optimal result. For this, the proposed methodology will have a classifier which will hybridize with the optimization technique named, firefly.

This is inspired from the literature survey in which it is explained that optimization algorithm can improve the factors of the classifiers to improve their classification or prediction property. For instance, in neural network if initial weights of the network are updated, the classification accuracy is also going to vary. So those weight values can be optimized using different optimization algorithms.

These findings motivate the proposed work to provide a hybrid model as best of classifier achieved (ANN) with traditional approach along with the CLAHE algorithm.

5. METHODOLOGY:

Explanation of the steps involved in the flow chart:

Step 1: Collection of input data from COCO database.

Step 2: Read the input data.

Step 3: Apply the CLAHE histogram equalization in the pre possessing stage of image processing.

Step 4: Apply the CLAHE algorithm to the images obtained from the database.

Step 5: Using the algorithm on a preloaded video for text detection using CLAHE in the EAST text detection method.

Step 6: Application of the Text detection on the live/ webcam.

Step 7: Compiling the Program.

Step 8: Cross checking the algorithm under different situations to ensure its working.

Step 9: Printing the results.

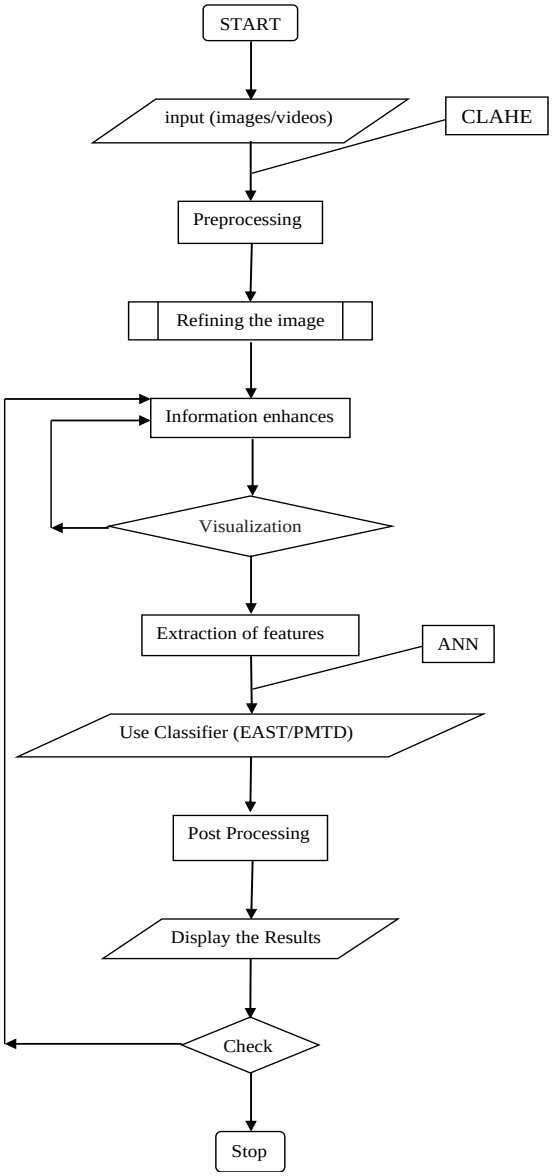

Fig 2: Flow chart of the proposed work.

6. EXPERIMENTAL PROCEDURE:

OpenCV's EAST text detector is a deep learning model, based on a novel architecture and training pattern. It is capable of:

- running at near real-time at 13 FPS on 720p images and
- obtains state-of-the-art text detection accuracy.

For this research work the main two algorithms used for the analysis are EAST text detector and PMTD method. Both the methods are robust and self-sustaining of their own but here we try to improve their functionality even further by employing a new method commonly known as CLAHE. It is a histogram equalization method that is used in this program for the noise removal from the images. Firstly, the Pre-processing on the images are done using the CLAHE technique then the image blocks are created by applying the image clipping limit of 8x8 grid. This is termed as the image preparation. Further this separates the three layers (RGB) in the images this exposes the image for further algorithm and reduces noise. Application of the Histogram Equalization makes it easier on the image to be read and processed easily. Then a new

algorithm is applied on the images and the desired outputs are discussed.

7. RESULTS:

In this section, the comparison results of various existing classifiers with the proposed one i.e. CLAHE are represented. The various classifiers i.e. PMTD, EAST are compared in terms of three parameters i.e. Precision, F-measure and Recall and the obtained results are discussed below:

Fig 3: Object detection with Python, OpenCV and the EAST text detector [15]**.**

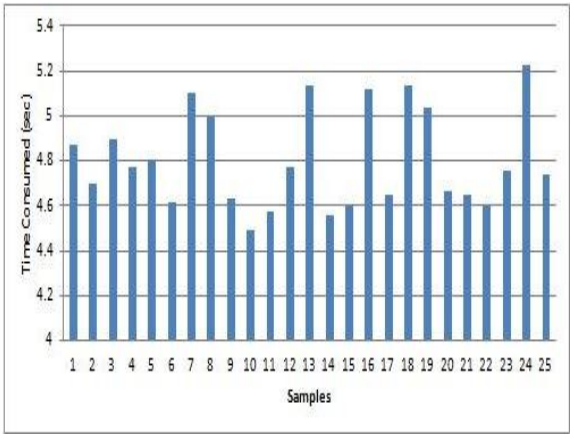

Fig 4: Time consumed per sample for object detection.

The graph in the figure 4 shows the time consumed by the new algorithm vs the number of samples it can take at a time. From the graph concurs with the algorithm actually stays close to a mean value of 4.5 secs irrespective of the number of images it is subjected to.

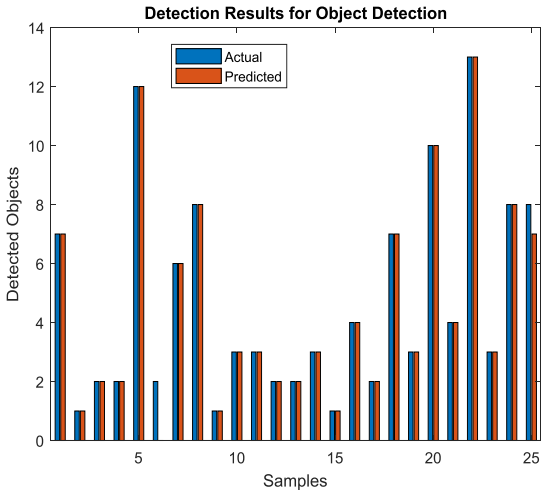

Fig 5: Bar graph showing the number of identified objects in the sample image.

Graph in the figure 5 shows the actual and predicted objects in the sample image. From the data it can be seen that algorithm works accurately for the predicted values. The predicted values those are observed by the eye and the identified objects that were returned by the program as output were a match. Numerically speaking from the Y-axis, we get to see the number of objects in image vs at X-axis the number of samples. So, there are 12 objects in an image and the algorithm returns the same values on execution then the actual and predicted values are shown same in case of 5, 10, 15, 20 samples.

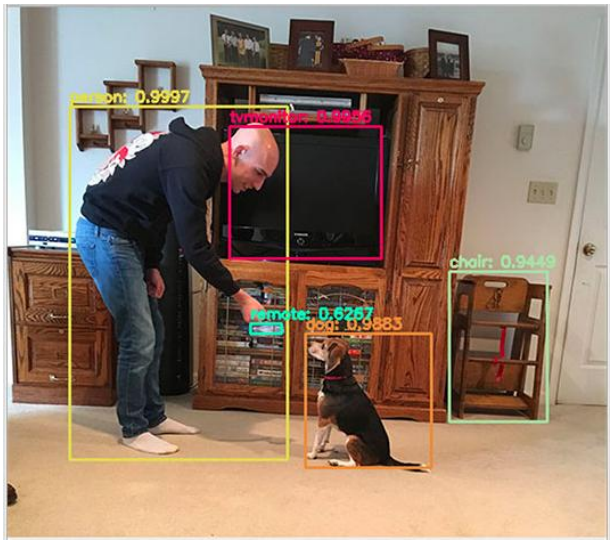

Fig 6: Example of object detection with 99.99% precision [15]**.**

The perfect working of the new CLAHE technique can be shown by an example in figure 6. In this figure it can be seen that the new and improved algorithm at work. It can be clearly seen that the new algorithm can identify the objects with 99.99 percent precision.

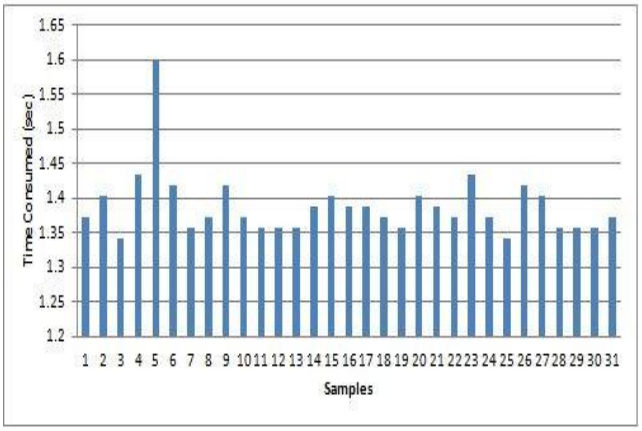

Fig 7: Time consumed per sample for text detection.

The graph in the figure 7 shows the time it takes for the new algorithm to identify the text in the image. From the graph it is seen that the new algorithm takes approximately 1.2 second for the detection of text in the group of samples.

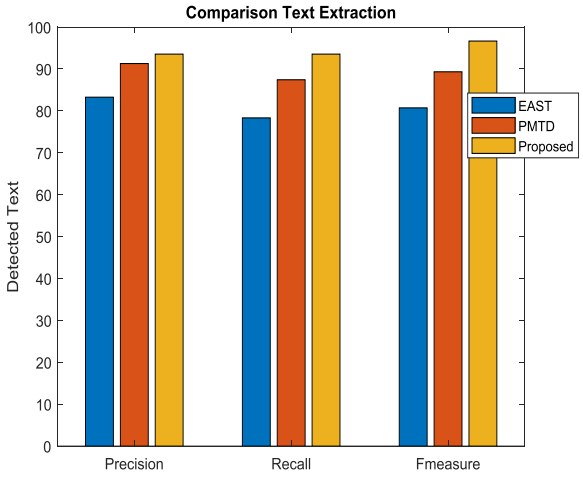

Fig 8: Comparison of the detected text for PMTD, EAST and proposed CLAHE algorithm.

The graph in the figure 8 clearly shows the new proposed CLAHE method outperforms both the previous methods chosen as the baseline research purposes.

The following table gives us the clear picture of the precision, recall values and F-measure values numerically.

Table 1: comparison of the different methods based on the Precision, recall and F-measure.

Methods	Precision	Recall	F-measure
EAST	83.2700	78.33	80.72
PMTD	91.3	87.43	89.33
Proposed	93.548	93.548	96.667

The table 1 and the figure 9 clearly shows that the Proposed method out performs the previous methods in all three respects which is an indicater that the new proposed methodology can be called better one among all the three algorithms.

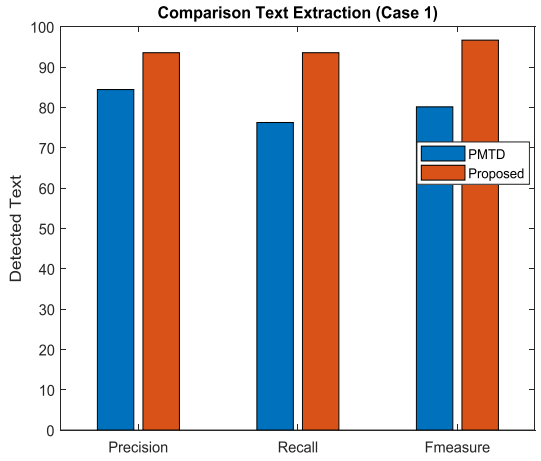

Fig 9: comparison of the precision, recall and F-measure between the PMTD and Proposed method.

The following formulas were used in the calculation of Precision, Recall and F-measure.

$$\% \text{ Precision} = {}^{TP}/_{(TP + FP)} \quad (1)$$

$$\% \text{ Recall} = \text{Sensitivity} \quad (2)$$

$$\% \text{ F-measure} = {}^{2 * TP}/_{(2 * TP + FP + FN)} \quad (3)$$

Where, TP = True positive.

FP = False Positive.

TN=True Negative

FN=False Negative

A True Positive is the value that is obtained when the model is rendered successfully. A True Negative is the value when the rendered model is unsuccessful.

A False Positive is the result where the actual value is negative but the Programme interprets it to be positive. A False Negative is results when the model is successful but the Program interprets it to be false.

8. CONCLUSION

In this research work, it has been tried to develop an algorithm to identify the text using OpenCV 4 on Python platform. The new algorithm is named after CLAHE technique, and takes EAST and PMTD method as a base line for the comparison of the precision, recall and F measure values. From the results it can be concluded that the new method CLAHE can outperform the rest of the two i.e. EAST and PMTD. The CLAHE algorithm was able to reach the levels of 93% precision values for the text detection in the video. These results are a significant improvement over the base work that was performed earlier. The time for the identification of the objects was significantly improved, and even time for the text detection was found to be just in the range of 1-1.5 seconds that is quite fast. The predicted values come close to one hundred percent while comparing to the actual cases. For the case of the text extraction, while comparing the three methods the CLAHE method outperforms in all the cases as can be seen from the figure 9. Now when it comes to the actual values of the precession, recall and F-measure values a significant improvement over the base work of PMTD method is seen. It will not be wrong to say that the CLAHE method

will outperform all the earlier text detection methods when put to test.

9. FUTURE SCOPE

Steps in the direction of object detection have been made by the Facebook AI researchers with the help of object identification algorithms that reached some precision in identifying the objects correctly. This object detection algorithm was extended for the text detection in the images/scenes. Here in this work with the help of python script were able to access the text detection by using the webcam in our pc the same can be done for the CCTV cameras, hawk-eye etc. systems for faster object identification. The object identification in real time at the phase is supposed to be helpful in practical life, take for instance, to resolve the conflicts, like that of fouls, or the winning cyclist in the race.

10. REFERENCES

[1] G. F. C. Campos, S. M. Mastelini, G. J. Aguiar, R. G. Mantovani, L. F. de Melo, and S. Barbon, "Machine learning hyperparameter selection for Contrast Limited Adaptive Histogram Equalization," *Eurasip J. Image Video Process.*, vol. 2019, no. 1, 2019.

[2] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[4] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 2642–2651, 2017.

[5] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[6] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[7] M. K. Chauhan and G. Kumar, "Automatic Text Detection and Information Retrieval on Mobile," vol. 5, no. 6, pp. 8285–8292, 2014.

[8] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," 2013.

[9] R. Socher, "Recursive Deep Models for Semantic Compositionality," no. October, pp. 1631–1642, 2013.

[10] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "OPENSURFACES: A richly annotated catalog of surface appearance," *ACM Trans. Graph.*, vol. 32, no. 4, 2013.

[11] F. Mokhtarian, S. Abbasi, and J. Kittler, "Robust and Efficient Shape Indexing through Curvature Scale Space," pp. 33.1-33.10, 2013.

[12] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2012.

[13] A. Krizhevsky and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks (presentation)," *ImageNet Large Scale Vis. Recognit. Chall. 2012*, p. 27, 2012.

[14] G. Patterson and J. Hays, "SUN attribute database: Discovering, annotating, and recognizing scene attributes," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2751–2758, 2012.

[15] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7574 LNCS, no. PART 3, pp. 340–353, 2012.

[16] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3485–3492, 2010.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[18] P. Nagabhushan and S. Nirmala, "Text Extraction in Complex Color Document Images for Enhanced Readability," *Intell. Inf. Manag.*, vol. 02, no. 02, pp. 120–133, 2010.

[19] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes_2009 IEEE Conference on Computer Vision and Pattern Recognition_2009_Farhadi et al.pdf."

[20] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 88–97, 2009.

[21] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, 2008.

[22] K. Jansen and H. Zhang, "Scheduling malleable tasks," *Handb. Approx. Algorithms Metaheuristics*, pp. 45-1-45–16, 2007.

[23] N. Chen, "A Survey of Indexing and Retrieval of Multimodal Documents : Text and Images," no. February, p. 40, 2006.

[24] D. Das, D. Chen, and A. G. Hauptmann, "Improving Multimedia Retrieval with a Video OCR," 2006.

[25] D. Chen, J. M. Odobez, and H. Bourlard, "Text detection and recognition in images and video frames," *Pattern Recognit.*, vol. 37, no. 3, pp. 595–608, 2004.

[26] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 4, pp. 256–268, 2002.

[27] C. Garcia and X. Apostolidis, "Text detection and segmentation in complex color images," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 4, pp. 2326–2329, 2000.

[28] S. E. Watson and A. F. Kramer, "Object-based visual selective attention and perceptual organization," *Percept.*

Psychophys., vol. 61, no. 1, pp. 31–49, 1999.

[29] K. Sobottka, H. Bunke, and H. Kronenberg, "Identification of text on colored book and journal covers," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 57–62, 1999.

[30] H. Li and D. Doermann, "Text enhancement in digital video using multiple frame integration," *Proc. ACM Int. Multimed. Conf. Exhib.*, pp. 19–22, 1999.

[31] B. S. Manjunath, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, 1996.

[32] L. Eikvil, "Line Eikvil, OCR- Optical Character Recognition., December 1993..pdf," no. December, 1993.

[33] M. H. Brill, "Computer Vision and Pattern Recognition: CVPR 92," *Color Res. Appl.*, vol. 17, no. 6, pp. 426–427, 1992.

[34] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

[35] I. Bose, A. R. Jana, and S. Chatterjee, "A BCS Theory ot Superconductivity in Heavy Fermion Systems," *Phys. Status Solidi*, vol. 136, no. 1, pp. 387–392, 1986.

[36] L. S. Vishnu and S. Rao, "Object Recognition and Object Counting using CNNs," no. 12376, 1237.