"The contributed chapters in the book written by the faculties of science stream in the light of the recent thinking and developments in the field of science and education. Science& Technology is now dominates almost every field of our activities in summary . The faculties (Science stream) of GEMS Arts & Science college have made an excellent attempt to bring about this book 'Homo Scientia' covering almost all the important areas from biological sciences to artificial Intelligence. Every article has its own merits in both academic and research fronts. I read my grateful appreciation and thanks to the contributors of this book for their untiring efforts."

Dr. Balagopalan Unni

GEMS

HOMO SCIENTIA

Dr. Balagopalan Unni,
Dr. Naveen Mohan,
Dr. K Gopalakrishnan,
Smitha Pramod V
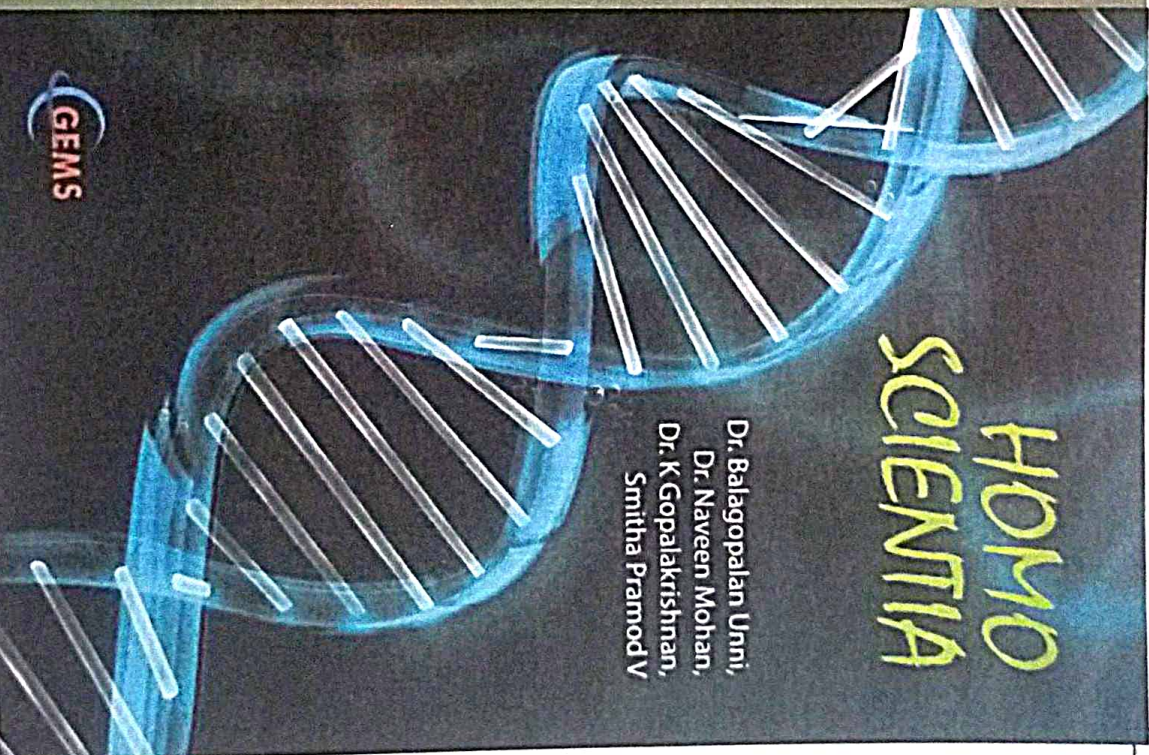
HOMO SCIENTIA

Dr. NAVEEN MOHAN
PRINCIPAL
GEMS ARTS AND SCIENCE COLLEGE
KADUNGAPURAM (PO), RAMAPURAM
MALAPPURAM DT., KERALA-679 321

Brief Biography

Dr. B.G.Unni, (Balagopalan Unni) Ph.D
(Allahabad central University)
FRES (London), FIANSc , FISAgBc, FICCE

Former Chief Scientist and Area Coordinator (Biotechnology & Biological Sciences) DADD and Fulbright Fellow retired from CSIR service in 2015 after 38 years of research career at CSIR North East Institute of Science & Technology Jorhat Assam. Appointed at Assam down town University as Director-Research in March 2015 and continued up to June 2019 and then re-designated as Adviser Research in August 2019). Back in Kerala, Dr.Unni is appointed as Director Academic & Research at GEMS College of Arts & Science affiliated to University of Calicut from August 2019. Both the positions are on honorary basis to strengthen the institutions in research areas. He did his BSc Biology (1972-74, Ewing Christian College, Alld University), MSc in Biochemistry(1974-76)(Second Rank) and Ph.D in Biochemistry from Allahabad University(1976-80) and PDF in Molecular Biology from Texas A&M University, USA(1988-91). Dr. Unni is specialized in Biochemistry, Molecular Biology, and Biotechnology and well established in his area of research and completed more than 40 years of research in both basic and applied fields of research. Dr.Unni got more than 130 research papers, 190 abstracts, 35 papers in proceedings, 7 patents,1 technology.18 chapters in books, edited 3 books and 29 students

Dr. NAVEEN MOHAN
PRINCIPAL
GEMS ARTS AND SCIENCE COLLEGE
KADUNGAPURAM (PO), RAMAPURAM
MALAPPURAM DT., KERALA-679 321

received PhD degrees under his guidance and supervision. Dr. Unni had completed more than 20 projects sponsored by Commonwealth Science Council, London, Ministry of Non conventional Energy Sources, Department of Non conventional Energy Sources Govt of India, North Eastern Council Govt of India, Department of Science &Technology, Department of Biotechnology, Central Silk Board,GB Pant Institute of Himalayan Environment and Development, CSIR and DRDO, Ministry of Defense, Govt of India during his scientific tenure at CSIR NEIST. Dr Unni received- Fulbright Travel Award/ Fellowship (USA) Dr. B.M. Das Memorial Science award, Hebrew University Award , H.R. Cama Memorial Travel Award,COSTED Travel Award,. DAAD- fellowship-Germany, Well Mark International Scholarship (USA) & Technology award in life sciences by CSIR,Govt of India . Best Fulbright Alumni Chapter Leader-South Asia Selected by the United States Education Foundation In India (USIEF), New Delhi .Nominated to represent India at the International Fulbright Scholars meet at Marrakech, Morocco- Nominated by United States Education Foundation In India, New Delhi . Dr. Unni is in the editorial board of more than eight indexed journal in the country .Dr.Unni was nominated to various state and central committees such as High power committee for development of sericulture activities Muga, Eri,Tassar and Mulberry in Assam nominated by Governor of Assam, .Expert in the area of non mulberry sericulture, Ministry of Textiles, Advisory Board, Post graduate Biotechnology programme, Academic Council, Assam Agricultural University, Research Council, Central Silk Board, Ministry of Textiles , DBT's Nominee for Biosafety Committee ,Vice President SBC (India) Indian Institute of Science Bangalore, Vice President Indian Academy of Neuro-sciences, Member Fulbright Academy of Science & Technology, USA, Board of studies- Botany Nagaland University and Biotechnology Saugar University Madhya Pradesh., Fellow, Indian Academy of Neurosciences & Indian Society of Agricultural Biochemists, Fellow Royal Entomological Society, London UK and Scientific

Dr. NAVEEN MOHAN
PRINCIPAL
GEMS ARTS AND SCIENCE COLLEGE
KADUNGAPURAM (PO), RAMAPURAM
MALAPPURAM DT., KERALA-679 321

Advisor International Foundation of Science, Sweden, Member, Board of Studies Raiganj University (2017----), Member Research Review committee Tea Board of India (2016-2019), Member Advisory Committee Cancer Research Advisory Board, North East Cancer Hospital & Research Institute (2017- -) President, Tea Improvement Consortium, Ltd, Tocklai Assam (2018-2020) .

Dr.Unni visited USA, Germany, Israel, Jordan, France, Morocco ,UK, Thailand ,Jordan, Singapore , China and UAE under various exchange program.

Dr. NAVEEN MOHAN
PRINCIPAL
GEMS ARTS AND SCIENCE COLLEGE
KADUNGAPURAM (PO), RAMAPURAM
MALAPPURAM DT., KERALA-679 321

Preface

I am very happy to learn that, the GEMS Arts & Science College is bringing out a series of books written by the faculty in this academic year. The college is occupying a very important position among the colleges in Kerala, the same way the college is having unique standing in both academic and research fronts too. This is because of the excellent management, faculties and the best performances of the students.. I have full confident that in the course of time, and with the sincere commitment and dedication of the faculties , students and with management , the college will attain high level perfection and excellence and became a model college in the state of Kerala

This book entitled " Homo Scientia" had comprehensive research topics in various aspects in the topics of cyber security, biotechnology, microbiology and geology.A brief description about the cybersecurity, the protection of computer set up such as hardware, software data from several threats have been described in the chapter The best practices for deploying and managing IPS network security tools have been explored. The integration of intrusion prevention system (IPS) solutions, adherence to security policies, regular updates, monitoring and the implementation of incident response procedures are considered to be the essential components of a comprehensive network security framework. The risk management in cyber security, various cyber-attack kinds, malware, and some strategies to tackle these attacks are also explained by the authors.   A comprehensive overview of the evolution of computer graphics, exploring the advancements in hardware, software, algorithms, and techniques that have propelled the field from its early pixel-based beginnings to the current state of realism etc also described. Optical character recognition has been extensively investigated in the past few years, and has been proven that high recognition rates can be achieved in specific

Dr. NAVEEN MOHAN
PRINCIPAL
GEMS ARTS AND SCIENCE COLLEGE
KADUNGAPURAM (PO), RAMAPURAM
MALAPPURAM DT., KERALA-679 321

application scenarios using some standard and well-studied methods such as neural network, support vector machine (SVM), etc. The possibility of learning an appropriate set of features for designing optical character recognition (OCR) has been investigated

Biotechnology is an interdisciplinary science using modern technologies to construct biological processes in research, agriculture, formulation of pharmaceutical products and other related fields. The better understanding of advances in plant genetic resources, genome modifications, omics technologies to generate new solutions for food security under changing environmental scenarios etc have been discussed in this chapter. The increasing demand for food had a great impact on the agriculture sector to address the various challenges associated with crop productivity. The tremendous advancement in plant research helps in understanding plant biology for sustainable food security, functional ecosystems, crop improvement and human health. One of the sustainable farming techniques is the use of fertilizer at nano level. Nanomaterials that enhance plant nutrition could be considered as an alternative to the conventional chemical fertilizers. one chapter covered the importance of nano fertilizer to enhance metabolic processes in plants and reviewed the concerns in developing nanotechnological methods in the future. Metabolomics has now emerged as a powerful tool for the comprehensive analysis of metabolites within biological systems. One of the chapters provides a review on metabolomics, encompassing its methodologies, applications, potential impact on personalized medicine ,and discusses further the need for advancements in analytical technologies. The antifungal activity of mangroves, particularly Rhizophora species are one of the main sources for fungicidal compounds due to the presence of high concentration of phenols. The antifungal activity of Rhizophora species has been elucidated, and could be further utilized as biocontrol agents for fungal disease in agricultural crops. One of the chapters discussed the species identification and its impact on economical and ecological level in the species like Nutmeg, one of the important medicinal plants that had a greater attention ,however, it was very difficult to differentiate the sexual identity

in the seedling stages. But the protein content screening among the studied plantlets had differentiated the sexes in the species as explained by the author.

AI (Artificial Intelligence) or machine intelligence enables farmers to enhance the quality and ensure a quick go-to market strategy for crops, and adoption of these algorithms to improve food industries. Artificial intelligence (AI) has also the potential to revolutionize education, from personalized learning to assessment and grading. Additionally, AI-powered tools can provide greater accessibility to students with disabilities, while also enabling more engaging and interactive content. AI continues to develop and become more prevalent in education, towards responsible and equitable implementation. However the negative and positive part of the AI may also be looked into.

The chapters related to microbiological aspects have also been incorporated in this book . Carbapenem-resistant A. baumannii (CRAb), bacteria that cause multi-infections in humans and resistant to multiple drugs too. The study attempted to isolate and characterize the bacterial species from the clinical specimens using biochemical techniques. The enzyme, carbapenemase produced by the bacteria was isolated and determined by different assays. Another study identified the antibacterial, antioxidant and anticancer activities of Ganoderma lucidum by various chromatographic techniques. Anticancer activity was also assessed on HeLa cell lines using MTT assay and DPPH assay. In one of the chapters,the author discussed L-asparaginase, one of the widely exploited enzymes for the treatment of acute lymphoblastic leukemia (ALL). Also attempted to isolate and characterize the enzyme from soil samples collected from different locations at Kerala. The study indicated that soils can provide a rich source for L-asparaginase which has got ample application in pharmaceutical industries.

The studies on various geological aspects with respect to different geographical areas in Kerala soil has been included in the book. The vertical geochemical variation and elemental mobility of the lateritic terrain in the Makkaraparamba of Malappuram District, Kerala has been very well investigated. Under extremely oxidizing and leaching conditions, laterite

Dr. NAVEEN MOHAN
PRINCIPAL
GEMS ARTS AND SCIENCE COLLEGE
KADUNGAPURAM (PO), RAMAPURAM
MALAPPURAM DT., KERALA-679 321

soil transformed into a variety of rocks and further developed into stable secondary product in the existing humid tropical and subtropical environments. The hydrogeological conditions in Kumbala- Kaliyar river basin, Kasaragod district, Kerala was assessed by means of Vertical Electrical Sounding (VES). The digital spatial data output of the present study would be much helpful for planning and management of surface and sub-surface water resources of Kasaragod River basin in which the Kasaragod township is centrally located

The contributed chapters in the book written by the faculties of science stream in the light of the recent thinking and developments in the field of science and education.Science& Technology is now dominates almost every field of our activities. In summary , The faculties (Science stream) of GEMS Arts & Science college have made a n excellent attempt to bring about this book Homo Scientia".covering almost all the important areas from biological sciences to artificial intelligence. Every article has its own merits in both academic and research fronts..I record my grateful appreciation and thanks to the contributors of this book for their untiring efforts.

Dr.Balagopalan Unni

Ph.D (Allahabad Central University), FRES (London)
Director Academic & Research
GEMS Arts & Science College, Malappuram Kerala
(Former Chief Scientist, CSIR-DST, Govt of India)
dir.ac.res@gemscollege.in

Dr. NAVEEN MOHAN
PRINCIPAL
GEMS ARTS AND SCIENCE COLLEGE
KADUNGAPURAM (PO), RAMAPURAM
MALAPPURAM DT., KERALA-679 321

Index

Dr. NAVEEN MOHAN
PRINCIPAL
GEMS ARTS AND SCIENCE COLLEGE
KADUNGAPURAM (PO), RAMAPURAM
MALAPPURAM DT., KERALA-679 321

OPTICAL CHARACTER RECOGINTION USING HOG AND DBN LEARNING

Dr. Sandhya Balakrishnan P K

Assistant Professor

Department of Computer Application

ABSTRACT

Optical character recognition has been extensively investigated in the past few years. It has been proven that high recognition rates can be achieved in specific application scenarios using some standard and well-studied methods such as neural network, support vector machine, etc. However, since their primary focus is usually fast or real-time recognition in the fairly controlled environments, the major issues related to the system's adaptivity are not fully investigated in the literature. Mainly, the recognition rate of the OCR systems will then be affected by the image noise and the difference between the training databases. In this paper, we investigate the possibility of learning an appropriate set of features for designing OCR. We have proposed a deep belief network for the recognition of text. Histogram of oriented gradients is used as the features. Better recognition accuracy is achieved when experimented with a dataset consisting of more than 10,000 samples.

INTRODUCTION

OCR is the recognition of printed or written text characters by a computer. This involves photo scanning of the text character-by-character, analysis of the scanned-in image, and

then translation of the character image into character codes, such as ASCII, commonly used in data processing. In OCR processing, the scanned-in image or bitmap is analyzed for light and dark areas in order to identify each alphabetic letter or numeric digit. When a character is recognized, it is converted into an ASCII code. Special circuit boards and computer chips designed expressly for OCR are used to speed up the recognition process.

The Deep Belief Network (DBN) and Deep Boltzmann Machine (DBM) are two popular deep probabilistic generative models that provide state-of-the-art results in many problems. These models contain many layers of hidden variables, and utilize an undirected graphical model called the Restricted Boltzmann Machine (RBM) as the building block. A nice property of the RBM is that gradient estimates on the model parameters are relatively quick to calculate, and stochastic gradient descent provides relatively efficient inference. However, evaluating the probability of a data point under an RBM is nontrivial due to the computationally intractable partition function, which has to be estimated, for example using an annealed importance sampling algorithm .

The paper is divided into five sections thus: (1) Introduction, (2) Pre-processing, (3) Segmentation and Feature Extraction, (4) Template matching and (5)Conclusion.

PRE-PROCESSING

Preprocessing is the first step in the development of any OCR system which prepares images for the subsequent phases. Depending upon the application and the type of input images, preprocessing may involve binarization, noise removal and, skew and slant detection and correction . In our study, we intend to work on contemporary images which are not likely to suffer from noise or degradations. The preprocessing in our case, therefore, comprises binarization of image to segment text from the background. In our implementation, we have employed the well-known Otsu's global thresholding to binarize the text image.

39

SEGMENTATION AND FEATURE EXTRATION

In any OCR system, character segmentation and feature extraction are essential stages after the image is being acquired. They constitute two major phases upon which largely OCR performance depends on as shown in the figure 1.
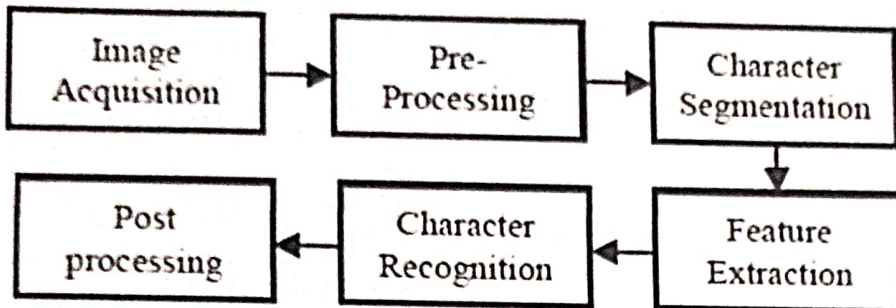
```
┌──────────────┐    ┌──────────────┐    ┌──────────────┐
│    Image     │──▶ │    Pre-      │──▶ │  Character   │
│ Acquisition  │    │  Processing  │    │ Segmentation │
└──────────────┘    └──────────────┘    └──────────────┘
                                                │
                                                ▼
┌──────────────┐    ┌──────────────┐    ┌──────────────┐
│     Post     │◀── │  Character   │◀── │   Feature    │
│  processing  │    │ Recognition  │    │  Extraction  │
└──────────────┘    └──────────────┘    └──────────────┘
```

Fig. 1: Overview of OCR.

SEGMENTATION

Segmentation processes, including following processes:

» Line Segmentation

Line segmentation is the process in which from the image, we extract only lines or differentiate the lines. Horizontal projection of a document image is most commonly used to extract the lines from the document. The horizontal projection will have separated peaks and valleys for the lines that are well separated and are not tiled, which serve as the separators of the text lines. These valleys are easily detected and used to determine the location of boundaries between the lines.

» Word Segmentation

Word segmentation is a process of dividing a string into its component words. Word splitting is the process of parsing concatenated text to infer where word breaks exist. By using vertical projection profile, one can get column sums. By looking for minima in horizontal projection profile of the page, we can separate the lines and then separate words by looking at minima

40

in the vertical projection profile of a single line. By using the valleys in the vertical projection of a line image, one can extract words from a line and also extracting individual characters from the word .

» Character Segmentation

In character segmentation, we extract only characters from word. Character segmentation is a difficult step of OCR systems as it extracts meaningful regions for analysis. This step decomposes the images into classifiable units called character.

FEATURE EXTRACTION

Histogram of oriented gradients algorithm is implemented for feature extraction. Histogram Of Oriented Gradients is a feature descriptor used to detect objects in computer vision and image processing. The HOG descriptor technique counts occurrences of gradient orientation in localized portions of an image – detection window, or region of interest.

Implementation of the HOG descriptor algorithm is as follows:

1. Divide the image into small connected regions called cells, and for each cell compute a histogram of gradient directions or edge orientations for the pixels within the cell.

2. Discretize each cell into angular bins according to the gradient orientation.

3. Each cell's pixel contributes weighted gradient to its corresponding angular bin.

4. Groups of adjacent cells are considered as spatial regions called blocks. The grouping of cells into a block is the basis for grouping and normalization of histograms.

5. Normalized group of histograms represents the block histogram. The set of these block histograms represents the descriptor.

41

TEMPLATE MATCHING

In this module, deep belief network algorithm is implemented for template matching. DBNs are graphical models which learn to extract a deep hierarchical representation of the training data. They model the joint distribution between observed vector x and the l and the hidden layers h^k as follows:

$$P(x, h^1, \ldots, h^l) = \left(\prod_{k=0}^{l-2} P(h^k|h^{k+1}) \right) P(h^{l-1}, h^l)$$

where, $x=h^u$, $P(h^{k-1}/h^k)$ is a conditional distribution for the visible units conditioned on the hidden units of the RBM at level k, and $P(h^{l-1}/h^l)$ is the visible-hidden joint distribution in the top-level RBM. This is illustrated in the figure below.
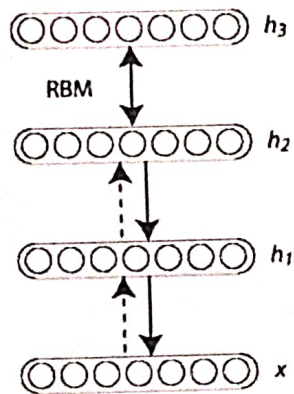
The principle of greedy layer-wise unsupervised training can be applied to DBNs with RBMs as the building blocks for each layer. The process is as follows:

1. Train the first layer as an RBM that models the raw input $x=h^{(0)}$ as its visible layer.

2. Use that first layer to obtain a representation of the input that will be used as data for the second layer. Two common solutions exist. This representation can be chosen as being the mean activations $P(h^{(1)}= 1/h^{(0)}$ or samples of $P(h^{(1)}/h^{(0)})$

3. Train the second layer as an RBM, taking the transformed data (samples or mean activations) as training examples (for the visible layer of that RBM).

42

4. Iterate (2 and 3) for the desired number of layers, each time propagating upward either samples or mean values.

5. Fine-tune all the parameters of this deep architecture with respect to a proxy for the DBN log- likelihood, or with respect to a supervised training criterion (after adding extra learning machinery to convert the learned representation into supervised predictions, e.g. a linear classifier).

We focus on fine-tuning via supervised gradient descent. Specifically, we use a logistic regression classifier to classify the input based on the output of the last hidden layer of the DBN. Fine-tuning is then performed via supervised gradient descent of the negative log-likelihood cost function. Since the supervised gradient is only non-null for the weights and hidden layer biases of each layer (i.e. null for the visible biases of each RBM), this procedure is equivalent to initializing the parameters of a deep MLP with the weights and hidden layer biases obtained with the unsupervised training strategy.

EXPERIMENTAL RESULTS

The performance of the proposed method is found by experimenting on the natural image samples. 400 samples for each character are collected from different images and with different background. The accuracy is calculated as the average of recognition rates for the five test subsets. Studied the effect of training sample set size on the recognition accuracy and found that using 80% of the data for training results in improvement of accuracy.

CONCLUSION

A new approach for OCR was proposed using modified template matching. The deep belief network is implemented to learn the HOG descriptors features. The deep learning network algorithm contributed to increase in recognition accuracy. The algorithm will have impressive results with the high training database.

43