Doctoral Consortium

(in conjunction with ICPR 2024)

Kolkata, India

Date: 1:30 pm – 5:30 pm, December 1st, 2024
Location: Biswa Bangla Convention Centre, Kolkata, India.

Chairs:

Daniel Lopresti, Lehigh University, Pennsylvania, USA
Mayank Vatsa, Indian Institute of Technology Jodhpur, India
Véronique Eglin, INSA Lyon, France

INTRODUCTION

The first edition of the Doctoral Consortium at ICPR2024 provides a unique opportunity for Ph.D. students to test their research ideas, present their current progress and future plans, and receive constructive feedback and insights regarding their future work and career perspectives.

The ICPR 2024 Doctoral Consortium has accepted 17 students spanning an impressive range of topics within diverse and dynamic fields of pattern recognition, illustrating the richness of the research landscape, like Deep Learning and Uncertainty, 3D Imaging and Profiling, Network Analysis, Handwriting Analysis, Medical Imaging, Image and Video Processing, Bias and Ethics in AI, Fake Detection, Financial Forecasting and Environmental Monitoring, Audio Forensics. During the DC, each research project is presented through a teaser/poster session, focusing on the outline of the objectives, the methodology, the expected results, the state of the art in their area, and the current stage of their research.

During the teaser (introductory) session, each student makes a brief presentation of their research to the audience, inviting attendees to visit their poster at the poster session. An award for the best poster will be delivered at the end of the doctoral consortium.

This half-day event marks a significant milestone for the ICPR community, and we are confident it will foster meaningful interactions and collaborations, enhancing the experience and careers of our PhD students. We hope it will encourage them to stay engaged with the ICPR research community, both during their studies and as they advance in their professional careers after graduation. We extend our gratitude to all contributors and organizers who made this initiative possible. We look forward to the exciting exchanges and insights that this session will undoubtedly bring.

PROGRAM

Start at 1:30 pm (end at 5:30 pm)

1 : 30 – 1 : 40	Opening - Introduction to ICPR Doctoral Consortium 2024 *D Lopresti, M Vantsa & V Eglin*
1 : 40 – 2 : 30	Teasers presentation of each PhD project
2 : 30 – 4 : 00	Poster session and discussions
4 : 00 – 4 : 30	*Coffee Break*
4 : 30 – 5 : 20	*"Ask Me Anything"* session
5 : 20 – 5 : 30	Concluding remarks and Best Poster Award

OVERVIEW OF CONTRIBUTIONS

The 17 PhD projects selected for DC-ICPR2024 showcase a diverse range of innovative approaches addressing societal and technological challenges. These contributions span various topics, leveraging advanced machine learning techniques, particularly deep neural networks, to advance the field of pattern recognition. This first edition of DC-ICPR offers a valuable platform to discuss these projects, foster collaboration, and explore the future of pattern recognition research.

Sharat Agarwal	Exploiting Contextual Uncertainty of Deep Models for Data Annotation	India
Vaishnavi Ravi	Novel Deep Learning Methods for 3D Profiling of Objects using Fringe Projection	India
Joyita Chakraborty	Time Evolving Citation Networks: Modeling, Analysis, Mining, and Applications	India
Jaya Paul	Development of a Writer Verification System- A Case Study on Bangla Scripts	India
Karri Karthik	Classification of Retinal Diseases from Enhanced Optical Coherence Tomography Images using Artificial Intelligence	India
Payel Rakshit	Development of Bangla Handwritten Text Recognition System	India
Nabajyoti Das	Polarimetric Synthetic Aperture Radar (PolSAR) Image Classification using Deep Learning Techniques	India
Honghui Yuan	Style Transformation for Text Image with Generation Model	Japan
Abhishek Tiwari	Deep Learning-Based Framework for DTI Parameters Estimation and Analysis for Sparse Diffusion MRI Data	India
Manali Patel	Towards Efficient Machine Learning Approach for Stock Price Movement Prediction	India
Tanusree Ghosh	GAN and DM Generated Synthetic Image Detection in the Age of Misinformation	India
Rajat Chakraborty	Identifying Deepfakes based on Human Physiological Signals	India
Siddharth Jaiswal	Audit & Mitigation of Gender Biases in Human-AI Platforms	India
Francisco Raverta Capua	Development of a forest mapping system using UAVs and AI models	Argentina
Taiba Majid Wani	Audio Forensics: Leveraging Deep Learning and Adaptive Learning Methods for Audio Deepfake Detection	Italy
Aniket Gurav	Handwritten Text Recognition for Historic Documents with Norwegian Perspective	Norway
Elena Bueno-Benito	Unsupervised Video Representation Learning: developments and applications	Spain

Exploiting Contextual Uncertainty of Deep Models for Data Annotation

Shart Agarwal[1]

Indraprastha Institute of Information Technology,Delhi, India
sharata@iiitd.ac.in

PhD context: I am pursuing my doctoral research under the joint supervision of Dr. Saket Anand from Indraprastha Institute of Information Technology Delhi (IIITD) and Dr. Chetan Arora from Indian Institute of Technology Delhi (IIT Delhi). My PhD journey began in August 2017, and I anticipate completing my research work by August 2025. This research combines deep academic insights with real-world applicability. We develop new theoretical frameworks to advance the field with approaches that can be readily deployed in real-world scenarios.

Abstract. Objects, in the real world, rarely occur in isolation and exhibit typical arrangements governed by their independent utility, and their expected interaction with humans and other objects in the context. For example, a chair is expected near a table, and a computer is expected on top. Humans use this spatial context and relative placement as an important cue for visual recognition in case of ambiguities. Similar to human's, DNN's exploit contextual information from data to learn representations. Our research focuses on harnessing the contextual aspects of visual data to optimize data annotation and enhance the training of deep networks. Our contributions can be summarized as follows: (1) We introduce the notion of contextual diversity for active learning CDAL [2] and show its applicability in three different visual tasks semantic segmentation, object detection and image classification, (2) We propose a data repair algorithm [3] to curate contextually fair data to reduce model bias, enabling the model to detect objects out of their obvious context, (3) We propose Class-based annotation [1], where contextually relevant classes are selected that are complementary for model training under domain shift. Understanding the importance of well-curated data, we emphasize the necessity of involving humans in the loop for accurate annotations. Contextual understanding by humans can be leveraged not only for data curation but also for intelligent data augmentation, enabling the creation of more comprehensive training datasets that capture real-world complexity. Future directions focus on developing novel interactive HITL systems enabling strategic expert intervention while leveraging model uncertainty for informed data selection.

Keywords: Active Learning · Human in the Loop · Semantic Segmentation · Object Detection · Data Fairness · Domain Adaptation.

1 Introduction

"For me context is the key - from that comes the understanding of everything." - Kenneth Noland.

Objects in the actual world exhibit typical arrangements, giving information of their interaction with other objects and the context of the overall scene. Humans use this spatial context as an essential cue for visual recognition in natural settings [4]. When we look at a complex scene, we perceive it effortlessly and identify the objects even without recognizing them. Context of the objects in the real world helps us solve perceptual inference faster and more accurately. It's generally observed that objects appearing in a consistent or familiar background are detected more accurately than objects in an inconsistent environment. For instance, we perceive the round disc-like object on the dining table from a distance as a diner plate.

Spatial context is an essential factor in facilitating scene understanding and object recognition for both machines and humans. Similar to how humans learn concepts and objects by observing their surroundings, deep learning models leverage large amounts of data, often annotated with labels, to learn diverse representations that can generalize to unseen data. One of the reasons why deep neural networks have done exceptionally well in the past decade is the availability of large diverse datasets. While data is crucial for deep learning, it is often overlooked in a learning setting. The more quality data a model is exposed to, the better it can capture intricate patterns, relationships, and nuances, leading to higher accuracy and robustness. While this works in theory, the sheer scale of data for practical applications comes at a high labor cost to label, especially in very specialized fields like medical or autonomous driving domains where the cost of running simulations to produce ground truth is very expensive.

Possible algorithmic solutions like active learning comes to let the algorithm iteratively pick most *informative* data examples to be labeled from unlabeled datasets in a manner such that it is representative of the underlying data distribution to a near-optimal learner [11]. Traditional AL techniques [7,8,6] have mostly been based on *uncertainty* and have exploited the ambiguity in the predicted output of a model. Existing approaches that leverage these cues are still insufficient in adequately capturing the spatial and semantic context within an image and across the selected set. The contextual uncertainty, which accounts for the relationships and dependencies between different elements in the data remains largely unexplored despite of its potential to significantly improve the selection of informative samples.

Considering the critical role that data plays in model training, *"we argue that along with the quantity of data, the quality of data needs attention."* To this end, we propose that model training should be designed so that models are trained using *contextually* diverse data to ensure they are accurate and unbiased while being efficiently trained. We thus investigate different aspects of contextual information from the available data and need of human in the loop for efficient annotation. The complexity of large-scale data annotation necessitates a human-in-the-loop framework that balances automation efficiency with human expertise.

2 Progress and Results

As motivated in previous section contextual information is a crucial part for humans visual understanding and inspired deep networks. In this section, we focus on the main contribution of this manuscript, discussing the importance of contextual information in selecting data for visual recognition in different applications.

Contextual Diversity for Active Learning [2],ECCV20: State of the art Active Learning approaches typically rely on measures of visual diversity or prediction uncertainty, which are unable to effectively capture the variations in spatial context. On the other hand, modern CNN architectures make heavy use of spatial context for achieving highly accurate predictions. Since the context is difficult to evaluate in the absence of ground-truth labels, we introduce the notion of *contextual diversity* that captures the confusion associated with spatially co-occurring classes. Contextual Diversity (CD) hinges on a crucial observation that the probability vector predicted by a CNN for a region of interest typically contains information from a larger receptive field. Such a measure would help select a training set that is diverse enough to cover a *variety of object classes* and their *spatial co-occurrence* and thus improve generalization of CNNs. The objective of this paper was to achieve this goal by designing a novel measure for active learning which helps select frames having objects in diverse contexts and background.

Contextually Fair Data To Reduce Model Bias [3], WACV22: Co-occurrence bias in the training dataset may hamper a DNN model's generalizability to unseen scenarios in the real world. For example, in COCO[9] dataset, many object categories have a much higher co-occurrence with men compared to women, which can bias a DNN's prediction in favor of men. Recent works have focused on task specific training strategies to handle bias in such scenarios, but fixing the available data is often ignored. We propose a novel and more generic solution to address the contextual bias in the datasets by selecting a subset of the samples, which is fair in terms of the co-occurrence with various classes for a protected attribute. We introduce a data repair algorithm using the coefficient of variation, which can curate fair and contextually balanced data for a protected class(es). This helps in training a fair model irrespective of the task, architecture or training methodology. Proposed solution is simple, effective, and can even be used in an active learning setting where the data labels are not present or being generated incrementally.

Contextual Class for Active Domain Adaptation [1], WACV23: In Active Domain Adaptation (ADA), one uses Active Learning (AL) to select a subset of images from the target domain, which are then annotated and used for supervised domain adaptation (DA). Given the large performance gap between supervised and unsupervised DA techniques, ADA allows for an excellent trade-off between annotation cost and performance. Prior art makes use of measures of uncertainty or disagreement of models to identify *regions* to be annotated by the human oracle. However, these regions frequently comprise of pixels at

object boundaries which are hard and tedious to annotate. Hence, even if the fraction of image pixels annotated reduces, the overall annotation time and the resulting cost still remain high. In this work, we propose an ADA strategy, which given a frame, identifies a set of classes that are hardest for the model to predict accurately, thereby recommending semantically meaningful regions to be annotated in a selected frame. We show that these set of *hard* classes are context-dependent and typically vary across frames, and when annotated help the model generalize better. We propose two ADA techniques: the `Anchor-based` and `Augmentation-based` approaches to select complementary and diverse regions in the context of the current training set.

3 Future Plan

Our research explores how contextual information within available data can be leveraged for effective annotation and DNN training. In an era where foundation / zero-shot models continue to advance and shape the future of AI, one cannot overstate the importance of high-quality labeled data. Our vision is to design systems that enable flexible human participation across different stages of AI pipeline while optimizing human effort and time. This direction is being successful in healthcare, where human-AI collaborative systems have achieved better accuracy through decision-support tools for clinicians [10], to wildlife conservation, where participatory systems have demonstrated robust performance in categorizing wildlife images [5]. In future, I plan to investigate the following research directions:

Context Aware Augmentation with Human In the Loop: Datasets often lack diverse contextual variation, thus leveraging human understanding to create new, realistic scenarios missing from current datasets. The system will suggest potential augmentations while humans validate their alignment with real-world possibilities, particularly valuable in applications like autonomous driving where standard datasets often miss critical edge cases

Neural Collapse for Data Selection: Neural Collapse is a phenomenon that describes how deep neural networks organize their features and class representations at the end of training, forming a simplex equiangular tight frame (ETF). Exploring this can provide insights into epistemic uncertainty and differentiating hard and easy samples. While NC is currently being studied in image classification, extending to more complex tasks like object detection, where features capture both semantic and contextual information is an exciting research direction.

Active Learning on Large Scale Datasets: Traditional AL approaches struggle to scale effectively to millions of images due to computational constraints. With the rise of foundation models and the increasing need for quality data, there is a critical need to develop scalable active learning approaches that can efficiently identify the most valuable samples from massive datasets while maintaining selection diversity and computational feasibility.

4 Conclusion

This thesis investigates the contextual uncertainty of deep models for data annotation. We proposed techniques to efficiently annotate and utilize data across various computer vision applications, significantly reducing annotation and training cost. Future directions focus on developing novel interactive HITL systems enabling strategic expert intervention while leveraging model uncertainty for informed data selection. Through this research, we aim to advance our knowledge of understanding deep models and data requirements to make real-world systems more `reliable`, `efficient`, and `trustworthy`.

References

1. Agarwal, S., Anand, S., Arora, C.: Reducing annotation effort by identifying and labeling contextually diverse classes for semantic segmentation under domain shift. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5904–5913 (2023)
2. Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 137–153. Springer (2020)
3. Agarwal, S., Muku, S., Anand, S., Arora, C.: Does data repair lead to fair models? curating contextually fair data to reduce model bias. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3298–3307 (2022)
4. Biederman, I., Mezzanotte, R.J., Rabinowitz, J.C.: Scene perception: Detecting and judging objects undergoing relational violations. Cognitive psychology **14**(2), 143–177 (1982)
5. Bondi, E., Koster, R., Sheahan, H., Chadwick, M., Bachrach, Y., Cemgil, T., Paquet, U., Dvijotham, K.: Role of human-ai interaction in selective prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 5286–5294 (2022)
6. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2372–2379. IEEE (2009)
7. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine learning proceedings 1994, pp. 148–156. Elsevier (1994)
8. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR'94. pp. 3–12. Springer (1994)
9. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
10. Peiffer-Smadja, N., Rawson, T.M., Ahmad, R., Buchard, A., Georgiou, P., Lescure, F.X., Birgand, G., Holmes, A.H.: Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clinical Microbiology and Infection **26**(5), 584–595 (2020)
11. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **6**(1), 1–114 (2012)

Novel Deep Learning Methods for 3D Profiling of Objects using Fringe Projection

Vaishnavi Ravi
vaishnavi1712@gmail.com

Indian Institute of Technology, Tirupati.

1 PhD Context

This PhD research is conducted under the supervision of Prof. Rama Krishna Gorthi. The program commenced in July 2019 and is completed by October 2024. The scope of this PhD falls within an academic context, focusing on advancing the field of 3D surface reconstruction. The work aims to contribute to both theoretical foundations and practical applications in this field with the help of recent advances in computer vision and deep learning.

2 Main Topic and Problem Statement

Traditional fringe projection profilometry (FPP) methods often struggle with complexities such as noise, discontinuities, and shadows, which can compromise the accuracy of 3D surface reconstruction. This thesis focuses on the development of lightweight, end-to-end deep learning (DL) frameworks to operate effectively with a single fringe image, specifically in real-world settings characterized by dynamic scenarios.

The main objectives of this thesis are:

- To explore recent advancements in deep learning and their applicability to FPP, focusing on developing lightweight, end-to-end DL frameworks that operate with single fringe images in real-world settings.
- To design and train DL models using synthetic data, minimizing the need for expensive and time-consuming data collection and annotation.
- To develop robust reconstruction algorithms capable of addressing challenging cases such as noise, discontinuities, and shadows in fringe images.

3 State-of-the-Art Methods

Fringe Projection Profilometry (FPP) is the most commonly used structured-light approach for the 3D surface profiling of objects. It has various advantages like non-contact high speed measurement, simple experimental setup and accurate reconstruction results. The advancements in this domain is of high importance in many applications like industrial design, virtual reality, 3D animation,

preservation of cultural heritage, medicine, human-computer interaction, security, quality control, etc. [1]

A typical fringe projection set-up consists of a light source that projects a sinusoidal light pattern onto the object of interest and a camera that captures the reflected deformed patterns due to the object from a different view. The recorded deformed patterns consist of phase shifts which contain information related to the height profile of the object. Hence, the main task at hand is to estimate these phase shifts, thereby estimating the height of the objects.

Traditional fringe projection profilometry (FPP) algorithms in literature follow a multi-step process, including fringe denoising, fringe analysis for phase extraction, phase unwrapping, and calibration. Each step typically involves distinct algorithms which are either multishot and singleshot approaches to determine the heights. Multishot techniques) [2] are known for their high precision in reconstruction, whereas singleshot approaches [3], [4] excel in dynamic situations. This creates a tradeoff between accuracy and speed. To extract height, all of these algorithms developed for each of the steps must be carefully integrated to achieve accurate reconstruction. The complexity of this process leaves room for errors, potentially leading to inaccurate surface height predictions.

Given, there are numerous traditional methodologies proposed to estimate precise depth information, their practical applicability is often limited, as real-world conditions deviate from the idealized assumptions made during their development. This mismatch introduces unwanted artifacts in the reconstructed results, caused by factors related to the object, the light source (projector), camera configuration, or the scene itself. Objects of interest exhibit diverse properties that complicate depth extraction. Factors such as complex structures that cast shadows [5], surface reflectivity that leads to intensity variations, and dynamic range that causes sudden discontinuities all contribute to artifact generation. Additionally, projector-related errors, such as synchronization issues or defocusing, and camera-related issues, like noise from hardware, dynamic conditions, or aberrations from exposure time and focus, further exacerbate these challenges.

Recent advancements in deep learning (DL) have led to a distinct class of algorithms for FPP [6,7]. This thesis emphasizes on handling real-world objects through synthetic data training and addressing challenges such as noise, discontinuities, and shadows.

4 Methodology and Contributions

In this thesis, we have explored the development of efficient preprocessing algorithms to handle different types of noise occuring in fringes and wrapped phase images by modeling fringe noise in deformed fringe images as Poissonian-Gaussian distributed for the first time and phase noise in wrapped phase images as Gaussian distributed. Further, lightweight DL model (LUNet) is developed to denoise the fringes for precise 3D reconstruction and a multi-task deep learning model, TriNet, is proposed for the simultaneous denoising and unwrapping of

the wrapped phase, enabling improved surface profile estimation of real objects in FPP.

- **Robust Fringe Denoising in FPP: A Poissonian-Gaussian Approach**
  A Poissonian-Gaussian model for camera sensor noise in contrast to conventional Gaussian is proposed for FPP. A customized lightweight encoder-decoder network (LUNet++) with just 20K parameters to perform fringe denoising is proposed.
- **A Multi-task Learning for 2D Phase Unwrapping in Fringe Projection**
  A multi-task learning-based phase unwrapping method for simultaneous denoising and wrap-count prediction is proposed for FPP. The proposed network, referred to as TriNet, has nested pyramidal architecture with a single encoder and two decoders, all connected through skip connections.

Conventional methods for FPP have many steps to obtain the final surface profile. To mitigate error propagation inherent in these conventional methods, recent advancements in DL have been leveraged. This framework is trained on synthetic data, reducing the cost and labor associated with data collection and annotation.

- **Lightweight Learning Framework for 3D Reconstruction using FPP (LiteF2DNet)** In this work, a lightweight DL framework to estimate depth profile of objects given linear reference and deformed fringes is developed. The proposed framework has dense connections in the feature extraction module to aid better information flow and has 40% lesser parameters than the base model, which provides lesser inference time, making it suitable for memory-constrained scenarios and real-time 3D reconstruction.

However, LiteF2DNet encounters limitations in handling critical cases involving discontinuities due to the many-to-one mapping problem arising in linear fringe patterns due to the phase shifts and pattern periodicity occurring in the same direction. To overcome this challenge, a radially symmetric circular pattern is introduced to record the underlying phase shifts in a one-to-one manner.

- **Single Shot Circular Fringe Projection for Profiling of Discontinuous Objects**
  This work proposes using a radially symmetric circular fringe as the structured light pattern for accurate unambiguous surface profiling of sudden height-discontinuous objects. Compared to the well-known fringe projection methods, the results depict that for a tolerable range of error, the proposed method can be applied for the reconstruction of objects with four times higher dynamic range and even at much lower fringe frequencies.
- **CF3DNet: A Circular Fringe Approach for Single-shot 3D Reconstruction**
  Dealing with discontinuities using single linear fringe is still an unsolved problem. This is due to the fact that the mapping is not one-to-one for discontinuous objects as the deformations and the periodicity of linear fringes

are in the same direction. To address this, CF3DNet is introduced in this work, which offers a one-to-one mapping between phase deformations and absolute phase shifts in 3D reconstruction from circular fringes.

Finally, to make this framework applicable to larger objects, an automated transformer-based fringe restoration network is developed to mitigate shadows caused by them in the fringe images. Blender, a computer graphics tool, is utilized to generate a realistic dataset with shadow effects. Through rigorous evaluation, the proposed model showcases precise object segmentation in fringe restoration, as quantified by metrics such as Mean Absolute Error (MAE), Intersection over Union (IoU), and Dice score. Overall, this thesis presents a promising approach to enhancing the accuracy and robustness of depth estimation in the presence of shadows in FPP, with potential applications in various fields requiring precise 3D surface measurements.

– **Transformer-based Fringe Restoration for Shadow Mitigation in FPP**
  This work introduces an algorithm to identify and repair the images with shadows created by complex objects with the help of single deformed fringe and provides better 3D profile estimation. In addition, this work introduces a new data generation procedure using blender.

List of Publications

1. **Vaishnavi Ravi**, and Rama Krishna Gorthi."*CF3DNet: A learning-based approach for single-shot 3D reconstruction from circular fringes.*" Optics and Lasers in Engineering 167 (2023): 107597.
2. **Vaishnavi Ravi**, and Rama Krishna Gorthi."*LiteF2DNet: A lightweight learning framework for 3D reconstruction using fringe projection profilometry.*" Applied Optics 62.12 (2023): 3215-3224.
3. K. S. Vengala, **Vaishnavi Ravi**, and Rama Krishna Gorthi."*A Multi-Task Learning for 2D Phase Unwrapping in Fringe Projection.*" IEEE Signal Processing Letters 29 (2022): 797-801.
4. Jagadeesh M, **Vaishnavi Ravi**, Sai Siva Gorthi, Subrahmanyam Gorthi, and Rama Krishna Gorthi, "*Single-shot circular fringe projection for the profiling of surface discontinuous objects.*" JOSA A 38, no. 10 (2021): 1471-1482.
5. **Vaishnavi Ravi**, Siddharth P, Sameer Ranjan and Rama Krishna Gorthi "*Transformer-based Shadow repairment framework for Fringe Projection Profilometry.*" International Conference on Pattern Recognition(ICPR), 2024 **[Accepted]**
6. **Vaishnavi Ravi** and Rama Krishna Gorthi "*Robust Denoising in Fringe Projection Profilometry: A Poissonian-Gaussian Approach with Lightweight Neural Networks.*" International Conference on Computer Vision & Image Processing, 2024. **[Accepted]**
7. **Vaishnavi Ravi**, and Rama Krishna Gorthi."*Robust Reconstruction of Objects using Deep Learning in Fringe Projection Profilometry*" IEEE Transactions on Instrumentation and measurement. **[Under Preparation]**

5 Action Plan & Future directions

There are still numerous challenges existing in FPP which pave way to several directions for pursuing further research in this field. We present some of the possible directions below.

- Investigating and mitigating inherent issues like saturation and color effects in Fringe Projection Profilometry to enhance accuracy and reliability.
- Extending the single-view surface profiling approach to reconstruct complete 3D CAD models by capturing deformed fringe images from multiple views, enabling applications in reverse engineering.
- Developing physics-based deep learning models that integrate traditional methods with deep learning techniques, leveraging their respective strengths to improve accuracy, robustness, and enable advanced applications in 3D surface measurement.

6 Career Plan

I am currently a **Postdoctoral Research Associate** in the Medical Imaging Group (MIG) at the Department of Computational and Data Sciences, Indian Institute of Science (IISc), Bangalore. I work under the guidance of Prof. Phaneendra Yalavarty. My current research is centered on the integration of radiomics-based information for enhancing treatment planning in Precision Oncology.Building on this experience, I aim to expand my research into developing robust AI-driven frameworks for real-time clinical decision support.

References

1. S. S. Gorthi and P. Rastogi, "Fringe projection techniques: Whither we are?" *Optics and Lasers in Engineering*, vol. 48, pp. 133–140, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:10361527
2. C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, "Phase shifting algorithms for fringe projection profilometry: A review," *Optics and lasers in engineering*, vol. 109, pp. 23–59, 2018.
3. M. Takeda and K. Mutoh, "Fourier transform profilometry for the automatic measurement of 3-d object shapes," *Applied optics*, vol. 22, no. 24, pp. 3977–3982, 1983.
4. Q. Kemao, "Windowed fourier transform for fringe pattern analysis," *Applied Optics*, vol. 43, no. 13, pp. 2695–2702, 2004.
5. R. Zhang, Y. Xiao, J. Cao, and H. Guo, "Rapid matching of stereo vision based on fringe projection profilometry," *in Proceedings of 8th International Symposium on Advanced Optical Manufacturing and Testing Technologies: Optical Test, Measurement Technology, and Equipment*, vol. 9684, pp. 139–148, 2016.
6. G. Spoorthi, R. K. S. S. Gorthi, and S. Gorthi, "Phasenet 2.0: Phase unwrapping of noisy data based on deep learning approach," *IEEE transactions on image processing*, vol. 29, pp. 4862–4872, 2020.
7. S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo, "Fringe pattern analysis using deep learning," *Advanced Photonics*, vol. 1, no. 2, pp. 025 001–025 001, 2019.

Time Evolving Citation Networks: Modeling, Analysis, Mining, and Applications

Joyita Chakraborty

National Institute of Technology (NIT), Durgapur, India
joyita.ckra@gmail.com

1 PhD context

PhD supervisors: Prof. *Subrata Nandi*, NIT Durgapur, India and Dr. *Dinesh K. Pradhan*, Dr. B.C. Roy Engineering College, Durgapur, India
Start and expected completion date of the PhD: 26/07/2019 to 19/02/2025
Context: Academic
Brief overview: Citation count is frequently used as a proxy for measuring research impact, leading to several performance indicators. Cumulative network growth models are mostly proposed for several modeling and prediction tasks. However, they have significant limitations, primarily because they consider growth only in a fixed time window. Hence, they fail to capture heterogeneous patterns exhibited over time, leading to inaccurate predictions. Moreover, citations can be easily manipulated and the easiest way is by adding excessive self-citations. Additionally, past literature predominantly focuses on microscopic analysis of individual scientific entities, which is becoming impractical today due to exponential growth in the size of scholarly databases. In contrast, analyzing citation time series offers a more effective alternative to capture temporal and topic-related information and the complex inter-dependencies among them [5].

This thesis aims to study time-evolving citation networks with several key objectives. Specifically, it involves understanding citation patterns, classifying individual articles based on their time-varying impact, predicting future trends, mapping fields, identifying citation gaming patterns with abnormal impact inflations, and developing tools and policies for accelerating quality research. In doing so, we employ advanced machine learning and deep learning-based techniques.

2 Problem statement

Broadly, we address four research questions in our thesis– **RQ1:** *Clustering citation time-series of individual articles to unfold generalized patterns that illustrate their natural cycles of rise and fall*, **RQ2:** *Can we develop a model that autonomously identifies significant features or segments from the citation time series and predicts these generalized classes?*, **RQ3:** *Scientific fraud detection by identifying anomalous trends from time-series journal impact factors. We aim to investigate methods for detecting such anomalous instances from a large-scale*

scholarly dataset, **RQ4:** *Can we leverage insights gained from citation time series to develop a practical academic search tool that is more personalized and user-customizable?*

3 State-of-the-art

Existing literature [4] on clustering citation time series of individual articles mostly defines arbitrary features, intuitive thresholds, and heuristic or fixed rule-based techniques, leading to subjective and redundant interpretations and ambiguity in the type and characteristics of identified trajectory patterns. There are specific properties of citation time series that pose modelling challenges [6]. It includes non-linearity, non-stationarity, anomalous diffusion, long-ranged correlations, and high variability. As a result, there is a lack of automated classification systems.

Moreover, in recent years, several anomalous citation practices have been reported [2], such as *excessive self-citations*, *citation stacking*, *cartels*, *cabals*, *rings*, etc. This leads to an abrupt inflation in the impact factor of individual journals and authors. Other recently reported incidents of citation anomalies include *index-jacking*, in which hijacked journals penetrate into academic databases; *sneaked references*, which refer to the manipulation of reference lists; *citation purchasing and generating fake Google Scholar profiles [3]*; *AI generated papers*, and *biased peer review process* [1] etc. Besides, popular research evaluation metrics consider citations received in a fixed window length adding to such biases. It favours papers from disciplines receiving early impact. Such manipulative or biased citation patterns are prevalent at all levels including individual authors, journals, editors, and publishers. Moreover, there is a lack of comprehensive studies on detecting scientific fraud from large-scale datasets.

In addition to it, traditional academic search systems do not disclose retrieval algorithms running in the back end leading to inefficient search processes. Therefore, personalized and user-customizable academic search systems are needed. Finally, handling such large-scale scholarly databases of the order of 10^8 entities requires high-end computing systems. The lack of annotated datasets specific to research problems leads to redundant data pre-processing steps and duplication of efforts each time researchers extract data from such databases. The *broad objective* of this thesis is to do macroscopic level modelling and analysis of citation data primarily as a time series and address the above issues.

4 Methodology & Contributions

First, for addressing the clustering citation time series problem, we define a standard feature set and propose a multiple k-means cluster ensemble algorithm. We validate the final cluster sets on varying lengths and identify three distinct clusters for short-term lengths– Early Rise-Rapid Decline (2.2%), Early Rise-Slow Decline (45%), and Delayed Rise-No Decline (53%). Further, for long-term lengths, we obtained three distinct clusters– Early Rise-Slow Decline (42%),

Delayed Rise-No Decline (57%), and Delayed Rise-Slow Decline (0.8%). The Microsoft Academic Graph (MAG) dataset is used for this study. We conduct analyses on three different lengths: 10-year, 20-year, and 30-year, consisting of 195,783, 56,380, and 41,732 papers, respectively. Finally, we define detailed temporal, citation, and field-related properties based on empirical analysis. This work also helps us to get annotated labels of trajectory patterns against raw time series. No significant cluster change was obtained by clustering articles with a citation time-series length of 20 years. Most articles fall into the Early Rise-Slow Decline (ER-SD) and Delayed Rise-Not Yet Declined (DR-ND) clusters. Besides, delayed-rise papers accumulate more citations over time than early risers, although early risers exhibit higher citation intensity with multiple peaks. This approach effectively captures all random groups and sub-groups of trajectory patterns, addressing inconsistencies and ambiguities in prior literature.

Second, we use this annotated data and propose an end-to-end explainable deep learning-based framework (*DeepTimeCiteX*) for automated classification and interpretation of citation time series. We generally find that transfer learning models can classify with up to 77% accuracy. We find ResNet and AlexNet perform best for long-term and short-term trajectory data, respectively. Besides, noise reduction, treating class imbalance, and data pre-processing are crucial steps for improving the accuracy of our framework. We employ the Empirical Mode Decomposition technique to remove noise without which accuracy significantly drops to up to 30%. Moreover, the LIMESegment model generates explanations– for papers which are early risers 5^{th} and 6^{th} years are important time segments. Similarly, for articles with delayed impact, 7^{th} and $23^{rd} - 25^{th}$ years are important time segments. Such automated inferences can help academic search engines and literature maintenance softwares manage information overload and help a user identify an article during its peak citation activity.

Third, to tackle the issue of scientific fraud detection, we systematically identify macroscopic features (at individual journal level) and microscopic features (at individual author, publisher, field level and temporal aspects) and use a simple k-means clustering algorithm to detect Collective Outliers (CO) and Point Outliers (PO) from large-scale journal citation datasets, respectively. Out of a total of 2,621 journals, 3% of them are identified as Collective Outliers (CO). Next, a microscopic time series analysis of journal impact factor data is done to identify Point Outliers (PO). 15% of CO are identified as PO.

Fourthly, we develop a personalised prototype graph-based academic search system, R^4-LitGraphs. It retrieves similar/relevant articles based on a single query paper given in a search and displays them as a force-directed graph. Unlike existing systems, the search or retrieval parameters are not kept in a black box. They are made available to users in the front-end interface through multiple filter options, allowing users to infer article relevance through intuitive infographics. It gives users flexibility and direct control over the search algorithm, enabling multidimensional comparison of articles and helping cross-discipline researchers identify relevant papers from other domains. Our proposed system

https://rb.gy/90i28t provides user customization and enhances interactiveness, diversity, and flexibility.

Our research establishes that time-series citation data is a better alternative for solving several issues of the research community. Search engines and academic recommender systems can use such automated citation trend classification and predictive models for efficient search. This can also help algorithms operating in literature maintenance software. Comprehensive evaluation metrics can be formulated considering variable-length windows based on distinct patterns. Moreover, our curated annotated datasets, developed models, scientific fraud early detection strategies, and prototype academic search systems will be helpful in future research in this domain.

4.1 Publications (progress made so far)

In this section, we list all our published and communicated works so far.

Journal Publications

1. **Joyita Chakraborty**, Dinesh K. Pradhan, Subrata Nandi (2024). A multiple k-means cluster ensemble framework for clustering citation trajectories. **Journal of Informetrics (Elsevier)**, DoI: https://doi.org/10.1016/j.joi.2024.101507 **(SCIE Indexed)**
2. **Joyita Chakraborty**, Dinesh K. Pradhan, and Subrata Nandi (2021). On the identification and analysis of citation pattern irregularities among journals. **Expert Systems (Wiley),** DoI: https://doi.org/10.1111/exsy.12561 **(SCIE Indexed)**
3. Dinesh K. Pradhan, **Joyita Chakraborty**, Prasenjit Choudhary, and Subrata Nandi (2020). An automated conflict of interest based greedy approach for conference paper assignment system. **Journal of Informetrics (Elsevier)**, DoI: https://doi.org/10.1016/j.joi.2020.101022 **(SCIE Indexed)**
4. **Joyita Chakraborty**, Biswajit Maity, Dinesh K. Pradhan, Subrata Nandi. DeepTimeCiteX: An explainable deep learning-based framework for automated classification and interpretation of citation time series. **Under Review (SCI Indexed)**

Conference Publications

1. **Joyita Chakraborty**, Biswajit Maity, Dinesh K. Pradhan, Subrata Nandi (2024). CiteDEK: A hybrid EMD-KNN-DTW model for classification of paper citation trajectories. **In 11th ACM IKDD CODS and 29th COMADS (CODS-COMADS), January 4–7, 2024, Bangalore, India,** DoI: https://doi.org/10.1145/3632410.3632481
2. **Joyita Chakraborty**, Dinesh K. Pradhan (2022). Citation Biases: Detecting Communities from Patterns of Temporal Variation in Journal Citation Networks. **In 6th International Conference of Data Management,**

Analytics, and Innovation (ICDMAI), January 14 - 16, 2022 (online mode), Proceedings published as part of Lecture Notes on Data Engineering and Communications Technologies book series (Springer), DoI: https://doi.org/ 10.1007/978-981-19-2600-6_42
3. **Joyita Chakraborty**, Dinesh K. Pradhan, and Subrata Nandi (2021). Research misconduct and citation gaming: A critical review on characterization and recent trends of research manipulation. **In 5th International Conference of Data Management, Analytics, and Innovation (ICDMAI), January 15 - 17, 2021 (online mode),** Proceedings published in Advances in Intelligent Systems and Computing (AISC) book series (Springer), DoI: https://doi.org/10.1007/978-981-16-2937-2_30
4. Dinesh K. Pradhan, **Joyita Chakraborty**, and Subrata Nandi (2019). Applications of Machine Learning in Analysis of Citation Network. **In 6th ACM IKDD CODS and 24th COMADS (CODS-COMADS), January 3–5, 2019, Kolkata, India,** DoI: https://doi.org/10.1145 /3297001.329 7053
ieeexplore.ieee.org/document/10725057

5 Planned actions planned before finishing the Ph.D.

Some of the potential future research plans before finishing the Ph.D. includes exploring new large-scale databases such as OpenAlex and the SciSciNet data lake. Further, we aim to work by fusing multi-modal paper meta-data information. We also aim to conduct a retraction study, and develop explainable citation time-series models. Actively looking out for post-doctoral research positions in several industry labs and academia.

References

1. Abalkina, A.: Challenges posed by hijacked journals in scopus. Journal of the Association for Information Science and Technology **75**(4), 395–422 (2024)
2. Haghighat, M., Hayatdavoudi, J.: How hot are hot papers? the issue of prolificacy and self-citation stacking. Scientometrics **126**, 565–578 (2021)
3. Ibrahim, H., Liu, F., Zaki, Y., Rahwan, T.: Google scholar is manipulatable. arXiv preprint arXiv:2402.04607 (2024)
4. Jiang, S., Koch, B., Sun, Y.: Hints: Citation time series prediction for new publications via dynamic heterogeneous information network embedding. In: Proceedings of the web conference 2021. pp. 3158–3167 (2021)
5. Yin, Y., Wang, D.: The time dimension of science: Connecting the past to the future. Journal of Informetrics **11**(2), 608–621 (2017)
6. Zamani, M., Aghion, E., Pollner, P., Vicsek, T., Kantz, H.: Anomalous diffusion in the citation time series of scientific publications. Journal of Physics: Complexity **2**(3), 035024 (2021)

Development of a Writer Verification System- A Case Study on Bangla Scripts

Jaya Paul[1][0009−0004−6303−2939]

Jadavpur University,Department of Computer Science and Engineering, West Bengal, India , Starting date (17/02/2017)- Finalization date(01/08/2024)
jayapl2005@gmail.com

Abstract. The primary focus of my thesis is writer verification within the context of Bangla scripts, a challenging area due to the intricate nature of the script and the wide variability in individual writing styles. These complexities make it difficult to achieve high accuracy in writer verification, as traditional methods struggle to account for the nuances in handwriting. The problem lies in effectively distinguishing between genuine and forged handwriting across different users while accommodating the unique features of Bangla script. To address this, my research introduces a novel approach that combines advanced image pre-processing techniques with machine learning algorithms. This approach leverages a newly developed dataset of handwritten Bangla samples, specifically designed to improve the accuracy of writer verification systems. The ultimate goal is to enhance the reliability of these systems, particularly for Bangla script, and contribute to related fields such as document analysis, forensic science, and biometric authentication. By advancing writer verification methodologies, this study aims to provide practical solutions for real-world applications where accurate identification is critical.

Keywords: Writer verification · Bangla script · Multi-level scripting · Genetic Algorithm · Majority voting · Tri-script.

1 PhD Context

The thesis, supervised by Dr. Anasua Sarkar (Jadavpur University) and Dr. Kaushik Roy (West Bengal State University), began on February 17, 2017, and was completed on August 1, 2024. This industrial-focused research at Jadavpur University aims to advance writer verification systems for Bangla scripts, addressing challenges in distinguishing genuine from forged handwriting through methods in computer science, biometrics, and forensic science.

2 Problem statement

The main topic of my thesis is writer verification within the context of Bangla scripts, which is a complex and challenging area due to the intricate structure of the script and the variability in writing styles. The problem statement revolves

around the difficulty in achieving accurate writer verification due to these complexities. To address this, my study proposes a novel approach that combines advanced image pre-processing techniques with machine learning algorithms, utilizing a newly developed dataset of handwritten Bangla samples to enhance verification accuracy. The study's goal is to improve the efficacy of writer verification systems, particularly for the Bangla script, and to contribute to fields such as document analysis, forensics, and biometrics.

3 State-of-art

The word 'biometrics' comes from' bios' (mean life) and 'metrics' (mean measure). In the last few years, behavioural biometrics had several applications for personal authentication and are widely used in forensics to detect identity and security applications. Handwriting biometric systems are typically divided into two components: verification and identification. In the verification process, an individual asserts their identity, and the system determines whether the handwriting matches the claimed identity. In contrast, the identification process involves the system recognizing the writer by comparing a provided handwriting sample against all previously enrolled users [1]. Online writer verification and identification systems implemented using spatial, temporal and pressure information [2] of the writing. An offline system implemented only using spatial information [3]. These types of authentication systems can also have used in hand mobile devices, historical document analysis [4], writer identification and verification systems [5] and security applications [6]. The flow of this literature survey is presented Offline (pen & paper) and online modes (writing on electronic devices) are two main categories. Document, paragraph, line, word, character, and hybrid level tasks are available in handwritten biometric analysis, which can be specific to any script or even multi-scripts. Offline (pen & paper) and online modes (writing on electronic devices) are two main categories. Document, paragraph, line, word, character, and hybrid level tasks are available in handwritten biometric analysis, which can be specific to any script or even multi-scripts. I have summed up the methods for feature extraction and classification problems in this area.

So, different types of surveys [7], published on the different segments (related to writer recognition) to cover this domain in past years. The utilization of deep learning approaches in the writer identification and verification domain has been notably limited [8].In most cases, deep learning approaches are too computationally expensive for real applications, while the traditional counterparts have much lower performance. This thesis focuses on different scripts (non-Indic and Indic) at various levels (document, paragraph, line, word, and character) and online and offline techniques for writer identification and verification tasks. Reviewed the literature on various handcrafted features used in handwriting script identification and verification tasks. Different types of handcrafted feature extraction techniques are used in several languages, and discusses the results of these techniques [9].

4 Methodology and contributions

In this thesis, introduce two new datasets aimed at addressing gaps in the availability of publicly accessible Bangla script databases and multilingual tri-script datasets for writer verification. The JUDVLP-BLWVdb dataset [10], collected from 101 native Bengali writers, addresses the absence of a benchmark dataset for writer verification in the Bangla language. The dataset comprises 488 pages of Bangla script, featuring variations in handwriting styles across 90 writers contributing 5 pages each, 6 writers contributing 4 pages each, 4 writers contributing 3 pages each, and 1 writer contributing 2 pages. For experiments, 3416 lines containing 20,778 words of Bangla script were selected. This dataset, categorized at the page, block, line, and word levels, offers a valuable resource for research in deep learning for vision and language processing.

Recognizing the absence of a publicly available benchmark for tri-language writer verification, we present the JUDVLP-TLWVdb dataset. Collected from 31 Indian writers proficient in Bengali, Hindi, and English, the dataset comprises 443 pages with samples written in multiple languages. Each writer contributed five reproductions of the same content, resulting in a total of 148 pages in Bangla, 147 pages in Hindi, and 151 pages in English. The dataset offers a unique resource for studying tri-script writer verification.

In the first experiment [10], the study focuses on offline Bangla handwriting content and evaluates the approach using specific hand-crafted features with Simple Logistic and RBF networks, SMO , and auto-derived features [11] using a CNN architecture. The hand-crafted feature set outperformed auto-derived features, achieving 94.54% average verification accuracy on a 100-writer database. The hand-crafted features utilized in the study comprised Radon Transform, Histogram of Oriented Gradient, Local Phase Quantization, and Local Binary Pattern, all of which were extracted from both inter- and intra-writer data. A Genetic Algorithm [12] was employed to reduce the dimensionality of the features and identify the most significant ones, using a Support Vector Machine as the fitness function. The top five experimental results were achieved through a consensus-based selection of the optimal feature set. Comparisons with alternative methods and features demonstrated promising performance.

The second experiment presents a comprehensive methodology that integrates techniques at the page, line, and word levels to verify the identity of the writer. A method is developed, leveraging the newly created dataset, JUDVLP-BLWVdb, which significantly enhances page-level writer verification performance. By employing the ensemble technique of majority voting, three classifiers (Support Vector Machine , Multilayer Perceptron, and Simple Logistic) are amalgamated, yielding a notable 12% enhancement in writer verification accuracy at the page level. This achievement reaches an impressive 97.62% accuracy across 101 diverse writers. This paper compares results with state-of-the-art writer verification approaches and explores deep learning-based methods, including VGG16 [13], ResNet34 [14], and AlexNet [11].

Final and last experimental results indicate that the SMO classifier outperforms other classifiers such as simple logistics and KNN [15]. A novel dataset for

writer verification systems using the tri-script approach is introduced, achieving a peak verification accuracy of 91.50% through a combination of Radon Transform, HOG, LBP, and LPQ features. The overall performance of the tri-script approach reaches 91.80%. Furthermore, this study employs the Vision Transformer (ViT) model [16] for writer recognition, demonstrating the superior performance of ViT when using tri-level block images of the page.

My contributions are in this thesis-

In this thesis, the challenges and opportunities of writer verification and authentication for Bangla scripts were thoroughly explored and addressed. The main objective was to create a reliable and efficient system capable of accurately verifying and authenticating authors of handwritten Bangla documents.

Throughout the research, diverse methodologies, algorithms, and techniques in pattern recognition, machine learning, and image processing were investigated. Special attention was given to the distinct characteristics of Bangla scripts, which were leveraged to develop an effective writer verification system. The collection of a comprehensive dataset of handwritten samples from numerous writers facilitated extensive experiments and evaluations.

The findings highlighted that the proposed writer verification system achieved impressive accuracy and performance. This success was attributed to the integration of the advanced machine learning models and feature extraction techniques. The thesis also emphasized the significance of dataset quality in determining the overall performance of the verification system.

Moreover, the study's practical implications extended beyond writer verification, with potential applications in forensic document analysis, authorship attribution, and data security. Furthermore, the advancements made in this research offer a foundation for future studies on other Indic scripts, contributing to the improvement of writer verification systems in multiple languages.

5 The actions planned after finishing the PhD:

After completing my PhD, I plan to extend my research into more practical applications of writer verification systems, particularly in forensic document analysis and digital security. I aim to collaborate with industry and academic partners to integrate these verification models into real-world systems, such as secure authentication methods for governmental and financial institutions. Additionally, I will explore the application of cross-script verification techniques to other Indic languages, expanding the scope of my work beyond Bangla, Hindi, and English. Future research will also involve refining the models for scalability and computational efficiency, ensuring they can be deployed in resource-constrained environments, such as mobile devices. Finally, I intend to pursue interdisciplinary collaborations, working with linguists, forensic experts, and cybersecurity professionals to explore new challenges and applications for writer verification technologies.

References

1. Abdeljalil Gattal, Chawki Djeddi, Faycel Abbas, Imran Siddiqi, and Brahim Bouderah. A new method for writer identification based on historical documents. *Journal of Intelligent Systems*, 32(1):20220244, 2023.
2. Rebecca Johinke, Robert Cummings, and Frances Di Lauro. Reclaiming the technology of higher education for teaching digital writing in a post—pandemic world. *Journal of University Teaching & Learning Practice*, 20(2):01, 2023.
3. Janek Bevendorff, Ian Borrego-Obrador, Mara Chinea-Ríos, Marc Franco-Salvador, Maik Fröbe, Annina Heini, Krzysztof Kredens, Maximilian Mayerl, Piotr Pęzik, Martin Potthast, et al. Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection: Condensed lab overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 459–481. Springer, 2023.
4. Gattal Abdeljalil, Chawki Djeddi, Imran Siddiqi, and Somaya Al-Maadeed. Writer identification on historical documents using oriented basic image features. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 369–373, 2018.
5. Chandranath Adak, Bidyut B Chaudhuri, and Michael Blumenstein. An empirical study on writer identification and verification from intra-variable individual handwriting. *IEEE Access*, 7:24738–24758, 2019.
6. Vikas Hassija, Vinay Chamola, Vikas Saxena, Divyansh Jain, Pranav Goyal, and Biplab Sikdar. A survey on iot security: Application areas, security threats, and solution architectures. *IEEE Access*, 7:82721–82743, 2019.
7. Heena Girdher and Harmohan Sharma. A survey on writer identification system for indic scripts. *Available at SSRN 3538594*, 2020.
8. Verónica Aubin, Marco Mora, and Matilde Santos Peñas. Off-line writer verification based on simple graphemes. *Pattern Recognition*, 79:414–426, 02 2018.
9. Jaya Paul, Kalpita Dutta, Anasua Sarkar, Nibaran Das, and Kaushik Roy. A survey on different feature extraction methods for writer identification and verification. *International Journal of Applied Pattern Recognition*, 7(2):122–144, 2023.
10. Jaya Paul, Kalpita Dutta, Anasua Sarkar, Kaushik Roy, and Nibaran Das. Writer verification using feature selection based on genetic algorithm: A case study on handwritten bangla dataset. *ETRI Journal*, page 1–12, 2024.
11. Siyuan Lu, Zhihai Lu, and Yu-Dong Zhang. Pathological brain detection based on alexnet and transfer learning. *Journal of Computational Science*, 30:41–47, 2019.
12. W. Siedlecki and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, 10(5):335 – 347, 1989.
13. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
14. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
15. Shaveta Dargan, Munish Kumar, Anupam Garg, and Kutub Thakur. Writer identification system for pre-segmented offline handwritten devanagari characters using k-nn and svm. *Soft Computing*, 24(13):10111–10122, Jul 2020.
16. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Classification of Retinal Diseases from Enhanced Optical Coherence Tomography Images using Artificial Intelligence

Karri Karthik

School of Medical Science and Technology
Indian Institute of Technology, Kharagpur, India
karrikarthik@gmail.com

Abstract. Optical coherence tomography (OCT) is a noninvasive imaging modality using interferometry to produce a cross-section of the retina. It provides anatomical information that assists ophthalmologists in diagnosing retinal diseases and planning treatment. The suppression of speckle noise becomes crucial for OCT image analysis studies. This thesis presents a speckle suppression technique based on Shannon coding for enhancing images. A convolutional neural network architecture is designed to classify retinal diseases using a custom residual module and an activation function. This system improves classification accuracy while simultaneously enhancing the retinal boundaries. OCT images possess a distinctive characteristic where each neighboring retinal layer exhibits a discernible pixel intensity range. This facilitates the design of an image representation technique that relies on Zeckendorf's theorem, which uses Fibonacci numbers and adaptive convolution layers, which utilize local texture information. The thesis also presents a network training optimization technique specifically designed for early stopping, which uses the entropy derived from the network's weights. The optimization technique does not necessitate separate validation data simultaneously, resulting in increased efficiency compared to scenarios where validation data is utilized.

Keywords: Optical Coherence Tomography (OCT) · Convolution Neural Networks (CNN) · Retinal Diseases · Image Classification.

1 PhD context

My PhD supervisor is Prof. Manjunatha Mahadevappa in the School of Medical Science and Technology (SMST) at the Indian Institute of Technology Kharagpur. My PhD registration data is 03-10-2019, and the expected completion date is 01-02-2025. The PhD has been carried out in an academic context.

2 Problem Statement

Retinal diseases present a considerable risk of permanent blindness when not addressed appropriately. The lower availability of retinal specialists exacerbates

the already precarious situation. Artificial intelligence (AI) in clinical evaluation has great potential to impact the healthcare delivery for patients with retinal diseases. This motivates the selection of the primary research problem the thesis addresses, which aims to explore and investigate the different AI techniques in classifying retinal diseases using optical coherence tomography images.

3 State-of-art

Traditional machine learning (ML) methods such as random forest and decision trees have been used for OCT image classification, but they require labeled data [6, 7]. For medical image segmentation and noise reduction, unsupervised methods that use ML techniques offer an excellent computational solution that only needs unlabeled data [2, 9]. Convolutional neural networks (CNNs) have played a significant role in image processing tasks since the 1980s [11]. CNNs employ multiple layers of convolution operations that autonomously learn important features during the training process. The introduction of non-linearity is achieved through the implementation of activation functions. The features learned by stacked convolution layers are known as feature maps, which are subsequently integrated into a fully connected neural network, facilitating image classification and segmentation tasks [8]. Image noise refers to the stochastic fluctuation in image intensity, manifesting as either multiplicative or additive in nature [1]. Noise constitutes an inherent component of any biological signal or image, and its presence invariably imposes challenges upon rule-based algorithms devised to assist medical practitioners in disease identification. Therefore, the reduction of speckle noise is essential.

4 Methodology and Contributions

4.1 Enhancement and Segmentation of OCT Images

This study presents a methodology for reducing speckle noise in OCT images. This independent approach aims to enhance the quality of the OCT images. The noise reduction algorithm is founded on the principles of Shannon-Fano coding. Shannon-Fano coding is a probabilistic decoding technique for information digits [3]. This study additionally describes an unsupervised image segmentation technique combined with a preprocessing technique for segmenting the retinal layer region as a foreground from the background. The retinal labeling algorithm utilizes a hybrid approach that combines the self-organizing map (SOM) [10] and K-means [4] clustering techniques. The preprocessing step significantly improves the performance of unsupervised clustering methods by eliminating any unwanted edges and regions that could impact the accuracy of the clustering result.

4.2 Deep Learning Methodology for Retinal Disease Classification

The study proposes a residual module to substitute the residual module in existing ResNet architectures [5]. The proposed module consists of an EdgeEn block

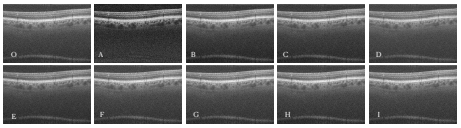


Fig. 1: Comparison of Denoising Techniques Applied to a High-Noise OCT Image: (O) Original, (A) Proposed Method, (B) Gaussian Filter, (C) Low-Pass Filter, (D) Wavelet Transform, (E) Lee Filter, (F) Anisotropic Diffusion, (G) Bilateral Filter, (H) Total Variation Denoising, and (I) BM3D

and a batch normalization (BN) layer (Figure 2). EdgeEn block also works on the derivative matrix, which increases the rate of change of stronger weights while suppressing the rate of change of smaller weights, corresponding to the noise in the image. The study also introduces a new activation function. The proposed activation function is positioned immediately after the fully connected layer, replacing the ReLU activation function. The derivative of the proposed activation function plays a crucial role in determining how the weights are updated. The primary purpose of the activation function is to retain the smaller negative weights and ensure a high rate of change for the smaller weights, preventing their loss during the backpropagation of errors.

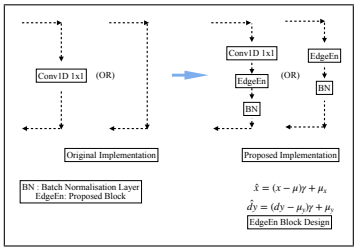
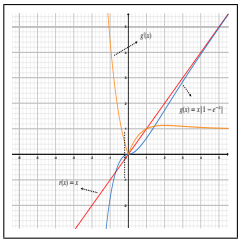
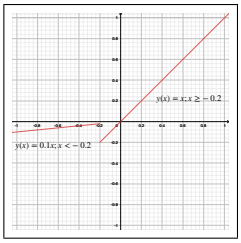


Fig. 2: Proposed Block Design



(a) Proposed Activation Function



(b) Augmented LeakyReLU

Fig. 3: Graphical Plots

4.3 Adaptive Convolutions for Retinal Diseases Classification Using Alternative OCT Image Representation

Convolution layers that lack explicit texture dependence may fail to capture intricate OCT image features, resulting in potentially less than optimal performance. Thus, driven by the necessity to overcome these limitations and ensure precise disease classification while preserving vital diagnostic information, we present an image representation technique and a specialized adaptive convolution layer tailored for the intricate features of OCT images. The proposed method operates in the spatial domain without directly manipulating the pixel value. This approach ensures the preservation of texture information in the processed image while minimizing blurring. The preprocessing technique based on Fibonacci numbers aims to enhance OCT images by reducing speckle noise. The adaptive convolution layer facilitates the learning of texture-sensitive feature maps. The clustered heatmap in Figure 5 showcases the correlation values among the evalu-

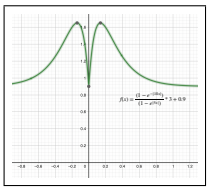


Fig. 4: Graphical Plot for Adaptation Function.

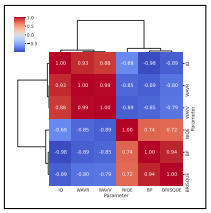


Fig. 5: Clustered Heatmap Analysis of Noise Evaluation Parameters.

ation parameters. The dendrogram identifies the two separate groups of parameters. The proposed adaptive convolution layer and its corresponding activation function were evaluated with seven OCT image classification CNN architectures. The overall improvement in accuracy varied across different architectures when implementing the proposed techniques, with a range of 0.44% to 2.44%.

4.4 Epoch Optimization for Efficient Training of Convolution Neural Networks

The validation data-based method is the most commonly used early-stopping technique because it's intuitive and simple to use. This technique allows for estimating a network's performance to unseen data. Entropy, a concept derived from information theory, quantifies the variability present in data, serving as a reflection of its informational content [12]. This serves as a unit of measurement for information, choice, and uncertainty [12]. The proposed approach involves the computation of the entropy of the weight matrix in the first fully connected layer. The first fully connected layer is an essential component in every CNN architecture, making it an ideal choice.

5 Future Work and Career Plan

Future studies could be categorized into four domains: AI model design research, image data processing techniques, multi-modal analysis, and research on the deployment of AI models.

The short-term career plan involves engaging in multimodal data research, mainly focusing on the integrated analysis of neuroimaging and ophthalmic imaging with collaborations for validation with animal studies to develop neurological disease models. The long-term career objective involves the development of smart devices that incorporate AI models for real-time monitoring, support, and predictive analytics, utilizing virtual models of diseases and human systems.

6 Dissemination of Research Work

Patent (Under Review)

– **Karri Karthik** and M. Mahadevappa, "A SYSTEM FOR PROGNOSIS OF RETINAL DISEASES FROM OPTICAL COHERENCE TOMOGRAPHY," Application Number:- 202431037312 dated: 2024/05/11 21:38:46

Journals

- **Karri Karthik** and M. Mahadevappa, "Entropy-based deep neural network training optimization for optical coherence tomography imaging," in *Applied Artificial Intelligence*, 38.1 (2024): 2355760, DOI: 10.1080/08839514.2024.2355760.
- **Karri Karthik** and M. Mahadevappa, "Deep learning with adaptive convolutions for classification of retinal diseases via optical coherence tomography," in *Image and Vision Computing*, 146 (2024): 105044, DOI: 10.1016/j.imavis.2024.105044
- **Karri Karthik** and M. Mahadevappa, "Enhancement and labelling of OCT images," in *Current Directions in Biomedical Engineering*, 9.1 (2023): 105044, DOI: 110.1515/cdbme-2023-1137
- **Karri Karthik** and M. Mahadevappa, "Convolution neural networks for optical coherence tomography (OCT) image classification," in *Biomedical Signal Processing and Control*, 79 (2023): 105044, DOI: 10.1016/j.bspc.2022.104176

References

1. Boyat, A.K., Joshi, B.K.: A review paper: noise models in digital image processing. arXiv preprint arXiv:1505.03489 (2015)
2. Eybposh, M.H., Turani, Z., Mehregan, D., Nasiriavanaki, M.: Cluster-based filtering framework for speckle reduction in oct images. Biomedical optics express **9**(12), 6359–6373 (2018)
3. Fano, R.: A heuristic discussion of probabilistic decoding. IEEE Transactions on Information Theory **9**(2), 64–74 (1963)
4. Gonzales, R.C., Wintz, P.: Digital image processing. Tech. rep., Addison-Wesley (1987)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hussain, M.A., Bhuiyan, A., D. Luu, C., Theodore Smith, R., H. Guymer, R., Ishikawa, H., S. Schuman, J., Ramamohanarao, K.: Classification of healthy and diseased retina using sd-oct imaging and random forest algorithm. PloS one **13**(6), e0198281 (2018)
7. Koprowski, R., Teper, S., Wróbel, Z., Wylegala, E.: Automatic analysis of selected choroidal diseases in oct images of the eye fundus. Biomedical engineering online **12**(1), 1–18 (2013)
8. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436 (2015)
9. Mayer, M.A., Tornow, R.P., Hornegger, J., Kruse, F.E.: Fuzzy c-means clustering for retinal layer segmentation on high resolution oct images. In: 19th Biosignal Conf (2008)
10. Natowicz, R., Sokol, R.: Self-organizing feature maps for image segmentation. In: New Trends in Neural Computation: International Workshop on Artificial Neural Networks, IWANN'93 Sitges, Spain, June 9–11, 1993 Proceedings 2. pp. 626–631. Springer (1993)
11. Rawat, W., Wang, Z.: Deep convolutional neural networks for image classification: A comprehensive review. Neural computation **29**(9), 2352–2449 (2017)
12. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)

Development of Bangla Handwritten Text Recognition System

Payel Rakshit[1][0000−0002−9006−2437]

West Bengal State University, Barasat, Kolkata-126. prmylife20@gmail.com

1 PhD Context

The PhD on the above mentioned topic was an academic science PhD, commenced on 6^{th} February, 2018 under the supervision of Prof. Kaushik Roy, Department of Computer Science, West Bengal State University, Barasat, Kolkata-126. It has been completed on 10^{th} May, 2024.

2 Problem statement

The recognition of Bangla handwritten text has always been a great challenge for researchers mostly due to the complex nature of the script along with the challenges of higher handwriting variations. The script has a wide range of structurally complex compound and modified characters and also upper and lower modifiers which makes the problem highly challenging compared to other scripts. The current study has emphasized the development of a Bangla handwritten text recognition system by means of segmentation-based methods. In this endeavour, tri-level (line, word, and character) segmentation and recognition at each segmented level have produced some significant methods for achieving Bangla text OCR.

3 State of the art methods

In the context of handwritten Bangla scripts, Das et al. [3] explored a novel approach for isolated compound character recognition, along with the basic characters. Similar type of works are also found in [5],[8], [15]. Some Bangla handwritten character recognition systems are also developed by Sazal et al. [16], Dutta et al. [4], and Shelke and Apte[17]. Bhattacharya et al. [1] explored a system for recognition of handwritten Bangla basic characters where local chain code histograms are computed for input character shape to obtain the features from the characters. Chaudhury et al. [2] proposed a deep OCR model for degraded Bangla documents whereas Shuvo et al. [18] proposed the 'MATHNET' to recognize Bangla numerals as well as mathematical symbols. A deep learning model namely 'RATNet' is introduced by Islam et al. [7] for handwritten Bangla isolated character recognition. The model is experimented on multiple standard Bangla isolated datasets.

4 Methodology and contributions

The proposed method for the development of a Bangla handwritten text recognition system consists of multiple phases. Firstly, an input grayscale image is converted to its binary form and the tri-level i.e. line, word, and character segmentation is performed. Once the segmentation is complete, the resultant segmented characters are fed to the proposed pre-trained CNN model for feature extraction and classification purposes. After classification, the segmented characters are converted to their editable character form. Once all the characters of a word image are converted as per their label, the whole word will be converted to its editable form and all words will form the editable text. The following Figure 1 gives a clear idea about the proposed handwritten text recognition system.

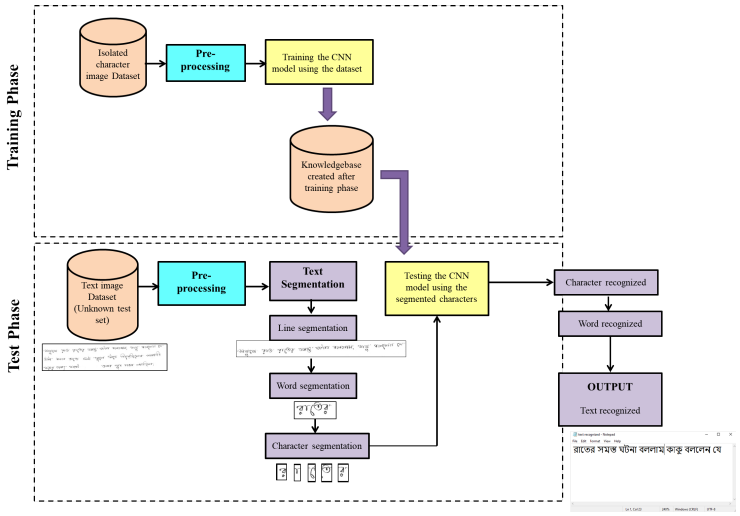


Fig. 1. Graphical representation of the proposed handwritten text recognition system

4.1 Tri-level Segmentation

The tri-level segmentation method contains line, word, and character segmentation. The text document is segmented into lines then the lines into words and the words into characters. The attempt of tri-level segmentation is first proposed in our work [10]. The enhanced version of line segmentation using a novel light projection method is proposed in our work [12].

4.2 Isolated Character Recognition

The chain code-based feature is extracted from the binarized images of the isolated characters and then different classifiers (Random Forest, Instance-Based K-

NN, Simple Logistic, and Multi-Layer Perceptron) are applied. The whole experiment is performed on multiple standard Bangla handwritten isolated datasets like CMATERdb, parts of WBSUBNdb_Isolated [13].

In another experiment three types of experiments are applied where the 'Ekush' dataset is subdivided into a subset of 10 digit classes, a subset of 50 basic characters, and a total dataset of 122 classes. For all three experiments images are divided into training, validation, and test set in a ratio of 70:10:20. Recognition performance of eleven SOTA CNN models (InceptionResNetV2, Vgg-19, Xception, ResNet50, MobileNetV2, EfficientNetB0, ResNet152V2, DenseNet201, InceptionV3, NASNetMobile, and EkushNet) are compared in this study [9].

4.3 Word Recognition

This step can be used as the precursor tool for the whole system. The proposed word recognition system is composed of two major phases: pre-processing and recognition. The recognition phase consists of two tasks: feature extraction and classification. In the current study two different CNN models are uses to recognize the segmented characters from the words. One of them is the existing CNN model 'MobileNetv2' and another one is the proposed CNN model.

4.4 Text Recognition

After successful completion of line and word segmentation the word recognition system is applied. Once the segmented words are recognized by the system, the words are directly pushed into list of **P** words with their corresponding labels. After **P** entries of the recognized words, the set of words from the list is written into a text file by maintaining the order of the recognized words to generate the lines of the text. The lines are also indexed sequentially. Now, by maintaining the line indexing, the full text is generated in its editable form.

4.5 Recognition Result

The text recognition system is applied on standard text dataset WBSUBNdb_text [6], [12]. The text accuracies are calculated based on two different methods: (i) word-wise text recognition (WTR) and (ii) char-wise text recognition (CTR). The recognition accuracies achieved for WTR and CTR are 72.27% and 58.32% respectively. A resultant text image with its editable form is illustrated in Figure 2. The WTR and CTR calculated for this example image are 73.39% and 60.76% respectively.

The publications related to the PhD work are provided in the reference as: [10, 13, 11, 12, 9, 14]

5 Future Perspective

The proposed system is the very first attempt to develop a full system of Bangla handwritten text OCR. The tri-level segmentation approaches are very efficient

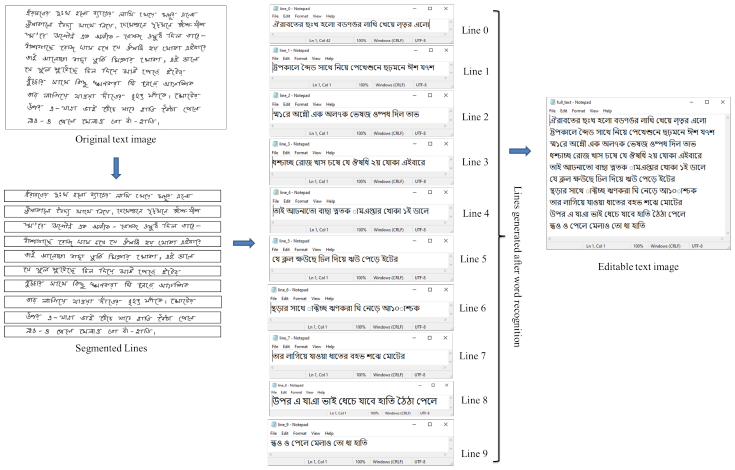


Fig. 2. Sample text recognition result generated by the system. The original text image, line-wise recognized editable text, and the final generated editable full text are presented with the required labels.

yet effective due to its domain-specific design. The proposed CNN model for the recognition of characters after segmentation has a good generalization capability and comparatively lower computational overhead. In the future, the proposed system can be compiled together as a complete application for Bangla handwritten text OCR. Additionally, the system performance can be improved by applying different transformers and multi-scale models. The system can be extended for other Indic scripts as well.

References

1. Bhattacharya, U., Shridhar, M., Parui, S.K.: On recognition of handwritten bangla characters. In: Computer vision, graphics and image processing, pp. 817–828. Springer (2006)
2. Chaudhury, A., Mukherjee, P.S., Das, S., Biswas, C., Bhattacharya, U.: A deep ocr for degraded bangla documents. Transactions on Asian and Low-Resource Language Information Processing (2022)
3. Das, N., Sarkar, R., Basu, S., Saha, P.K., Kundu, M., Nasipuri, M.: Handwritten bangla character recognition using a soft computing paradigm embedded in two pass approach. Pattern Recognition **48**(6), 2054–2071 (2015)
4. Dutta, S., Banerjee, S.: Isolated handwritten bangla character recognition using 1d discrete wavelet transform. In: Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012). pp. 266–267. IET (2012)
5. Gaur, A., Yadav, S.: Handwritten hindi character recognition using k-means clustering and svm. In: 2015 4th international symposium on emerging trends and technologies in libraries and information services. pp. 65–70. IEEE (2015)

6. Halder, C., Obaidullah, S.M., Santosh, K.C., Roy, K.: Content independent writer identification on bangla script: A document level approach. International Journal of Pattern Recognition and Artificial Intelligence **32**(09), 1856011 (2018). https://doi.org/10.1142/S0218001418560116
7. Islam, M.S., Rahman, M.M., Rahman, M.H., Rivolta, M.W., Aktaruzzaman, M.: Ratnet: A deep learning model for bengali handwritten characters recognition. Multimedia Tools and Applications **81**, 10631–10651 (2022). https://doi.org/10.1007/s11042-022-12070-4
8. Maitra, D.S., Bhattacharya, U., Parui, S.K.: Cnn based common approach to handwritten character recognition of multiple scripts. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1021–1025. IEEE (2015)
9. Rakshit, P., Chatterjee, S., Halder, C., Sen, S., Sk, O., Roy, K.: Comparative study on the performance of the state-of-the-art cnn models for handwritten bangla character recognition. Multimedia Tools and Applications pp. 1–22 (11 2022). https://doi.org/10.1007/s11042-022-13909-6
10. Rakshit, P., Halder, C., Ghosh, S., Roy, K.: Line, Word, and Character Segmentation from Bangla Handwritten Text—A Precursor Toward Bangla HOCR, pp. 109–120. Springer (2018). https://doi.org/10.1007/978-981-10-8180-4_7
11. Rakshit, P., Halder, C., Obaidullah, S., Roy, K., et al.: A survey on line segmentation techniques for indic scripts. In: International Conference on Recent Trends in Image Processing and Pattern Recognition. pp. 511–522. Springer (2020)
12. Rakshit, P., Halder, C., Obaidullah, S.M., Roy, K.: A generalised line segmentation method for multi-script handwritten text documents. Expert Systems with Applications (2022)
13. Rakshit, P., Halder, C., Roy, K.: An Approach toward Character Recognition of Bangla Handwritten Isolated Characters, pp. 15–28. Chapman and Hall/CRC (2019)
14. Rakshit, P., Mukherjee, H., Halder, C., Obaidullah, S.M., Roy, K.: Historical digit recognition using cnn: a study with english handwritten digits. Sādhanā **49**(39) (2024). https://doi.org/10.1007/s12046-023-02322-w
15. Sarkhel, R., Das, N., Saha, A.K., Nasipuri, M.: A multi-objective approach towards cost effective isolated handwritten bangla character and digit recognition. Pattern Recognition **58**, 172–189 (2016)
16. Sazal, M.M.R., Biswas, S.K., Amin, M.F., Murase, K.: Bangla handwritten character recognition using deep belief network. In: 2013 International Conference on Electrical Information and Communication Technology (EICT). pp. 1–5. IEEE (2014)
17. Shelke, S., Apte, S.: A multistage handwritten marathi compound character recognition scheme using neural networks and wavelet features. International Journal of Signal Processing, Image Processing and Pattern Recognition **4**(1), 81–94 (2011)
18. Shuvo, S.N., Hasan, F., Ahmed, M.U., Hossain, S.A., Abujar, S.: Mathnet: using cnn bangla handwritten digit, mathematical symbols, and trigonometric function recognition. In: Soft Computing Techniques and Applications, pp. 515–523. Springer (2021)

Polarimetric SAR (PolSAR) Image Classification using Deep Learning Techniques

Nabajyoti Das

Tezpur University, Tezpur, Assam, 784028, India
nabajd@tezu.ernet.in

1 PhD context

My PhD journey is being guided by the esteemed supervision of Dr. Swarnajyoti Patra, whose expertise and mentorship have been invaluable. I embarked on this academic pursuit in October, 2020 with the anticipation of completing it by September, 2025. My PhD is primarily focused on an academic context.

2 PolSAR Image Classification

A PolSAR image is constituted with all the four polarisation channels which preserves the full vector nature of the electromagnetic radiations. The information about the target surface can be retrieved on the basis of its response in different polarisation states. When a polarized radar wave interacts with earth's surface, the polarization of the wave is modified depending upon the specific characteristics of the surface. This includes its geometrical structure, shape, reflectivity, orientation as well as the geophysical properties such as moisture content, surface roughness etc. Due to this reason, earth features give different response in different polarization channels and on the basis of this response, various earth features can be identified from the radar image. Features such as ice, ocean waves, soil moisture, vegetation, geological features and man-made structures are better detected in Polarimetric SAR images as compared to images acquired by optical sensors. Hence, PolSAR image has garnered lots of interest for earth monitoring.

PolSAR image classification is a pixel-wise classification problem where we intend to assign a label to each pixel of the given PolSAR image. In the case of the supervised classification, we require training samples with known labels. Such training samples can be obtained by field survey, during which parameters such as type of crop, height, biomass, water content, etc., are collected at different points over the area of interest synchronous with satellite pass. A GPS location of such a sample is also collected, which allows us to register the surface observation with the image's pixel. A ground truth map can be generated from the field survey data or by combining it with visual and statistical analysis of the SAR image data or the optical image data. Such a ground truth map allows us to map the pixel of the image with the ground reality (label of a pixel).

Over the years, different methods for PolSAR image classification have been proposed. In the early days, parametric classifiers that model the probability distribution of the data were popular. Recently, due to advancements in the field of machine learning and deep learning such methods have been widely researched and more commonly in use today for PolSAR image classification.

2.1 Benefits of PolSAR Image Classification

Being an active remote sensing system, it doesn't depend on any external source of illumination like the sun. So, PolSAR system can monitor earth surface both during day as well as night time. PolSAR system can also penetrate through clouds and form an image of the earth's surface even for cloud covered region which is beneficial for scene classification task. Moreover, the image formed by PolSAR data is a very high resolution image and is formed by the back-scatter information that contains significantly higher information as compared to optical images which makes it suitable for many applications of remote sensing like disaster monitoring, forest monitoring, crop monitoring and land use monitoring.

2.2 Objectives

The main objectives of this research work is to develop deep learning based techniques for PolSAR Image classification.

1. To develop a robust method that incorporates textural feature for PolSAR image classification.
2. To incorporate spatial information and implement it using CNN architecture for PolSAR Image Classification.
3. To implement advance deep learning based model like Transformers and Explainable AI(XAI) by incorporating Semantic information for PolSAR Image Classification.

3 Literature review and brief state-of-art

1. Traditional textural measures like GLCM, MRFs and Gabor wavelets are often used to incorporate textural information to classifiers but these traditional textural measures are often mostly suitable for optical images rather than PolSAR images. Although, these traditional measures increases the information but it is still dependant on the selection on polarimetric features and requires domain expertise. [1], [8], [9], [11].
2. Incorporating spatial information are utilized with deep learning based technique like CNN but mostly such techniques rely on the automatic extraction of spatial information using CNN which makes it dataset dependant and as such it is difficult for the same technique to work well with different PolSAR dataset. [3], [4], [5], [7], [10].

3. The advanced deep learning based techniques which are being developed and used recently for PolSAR image classification although provides better classification accuracy but such methods are very complex and time consuming and also relies mainly on large collection of labelled training samples. [2], [6], [12].

4 Proposed Methodology

The proposed framework of PolSAR image classification using deep learning techniques is shown in Figure 1. As seen from the figure, the input will be the PolSAR image. Then in the next step a decision will be made if there is a need for pre-processing. The pre-processing step may involve applying of filtering strategy to remove the speckle noise and converting the imaginary values into decibels. The next step will be polarimetric feature selection which depends upon the approaches used. As mentioned, the three underlying core elements controlling the performance of PolSAR image classification are the the unlabeled pixels, polarimetric features being used and complexity of the classifiers. As such, to develop a better PolSAR image classification using deep learning based technique, the first approach is to incorporate textural information to the PolSAR image to enhance the classification performance. The second approach is to incorporate

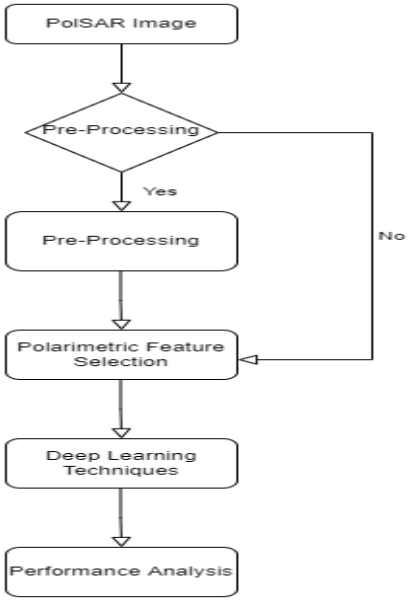


Fig. 1. Proposed Framework for PolSAR Image Classification using deep learning techniques

spatial information using CNN architecture and third approach is to develop advance deep learning based technique that incorporates semantic information. Hence, the outcome will be development of deep learning based techniques that improve PolSAR image classification. The methodology of the proposed research work is broken down into goals and tasks as follows:

- **Goal 1: To develop a method that incorporates robust textural feature for PolSAR image classification.**
  - **Task 1:** Extensive survey on literature and techniques focusing in addition of Textural information to PolSAR data.
  - **Task 2:** Development of method to incorporate textural feature in PolSAR image for improving classification accuracy.
  - **Task 3:** Performance analysis.
- **Goal 2: To incorporate spatial information and implement it using CNN architecture for PolSAR Image Classification.**
  - **Task 4:** Extensive survey on literature and techniques focusing on the incorporation of spatial information for PolSAR data.
  - **Task 5:** Development of method to incorporate spatial information and CNN model to enhance PolSAR image classification.
  - **Task 6:** Performance analysis.
- **Goal 3: To implement advanced deep learning based model like Transformers and Explainable AI(XAI) by incorporating Semantic information for PolSAR Image Classification**
  - **Task 7:** Extensive survey on literature and deep learning based techniques focusing in semantic information based PolSAR image classification .
  - **Task 8:** Implement advanced deep learning models by incorporating semantic information to enhance PolSAR image classification.
  - **Task 9:** Performance analysis.

4.1 Goals completed

Goal 1 and Goal 2 are completed so far.

4.2 List of publications

1. Classification of Polarimetric SAR Image using JS-Divergence Profile, 2022 IEEE Calcutta Conference (CALCON)
2. Dual-Branch CNN Incorporating Multiscale SVD Profile for PolSAR Image Classification, IEEE Transactions on Geoscience and Remote Sensing, Vol. 62, 2024
3. PolSAR Image Classification Using Superpixel Profile and CNN, ICPR-2024

References

1. Chen, S.W., Tao, C.S.: PolSAR Image Classification Using Polarimetric-Feature-Driven Deep Convolutional Neural Network. IEEE Geoscience and Remote Sensing Letters **15**(4), 627–631 (2018). https://doi.org/10.1109/LGRS.2018.2799877
2. Hua, W., Wang, X., Zhang, C., Jin, X.: Attention-Based Multiscale Sequential Network for PolSAR Image Classification. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022). https://doi.org/10.1109/LGRS.2022.3164464
3. Hua, W., Xie, W., Jin, X.: Three-Channel Convolutional Neural Network for Polarimetric SAR Images Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 4895–4907 (2020). https://doi.org/10.1109/JSTARS.2020.3018161
4. Hua, W., Zhang, C., Xie, W., Jin, X.: Polarimetric SAR Image Classification Based on Ensemble Dual-Branch CNN and Superpixel Algorithm. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 2759–2772 (2022). https://doi.org/10.1109/JSTARS.2022.3162953
5. Jamali, A., Mahdianpari, M., Mohammadimanesh, F., Bhattacharya, A., Homayouni, S.: PolSAR Image Classification Based on Deep Convolutional Neural Networks Using Wavelet Transformation. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022). https://doi.org/10.1109/LGRS.2022.3185118
6. Jamali, A., Roy, S.K., Bhattacharya, A., Ghamisi, P.: Local Window Attention Transformer for Polarimetric SAR Image Classification. IEEE Geoscience and Remote Sensing Letters **20**, 1–5 (2023). https://doi.org/10.1109/LGRS.2023.3239263
7. Marpu, P.R., Chen, K.S., Chu, C.Y., Benediktsson, J.A.: Spectral-spatial classification of polarimetric SAR data using morphological profiles. In: 2011 3rd International Asia-Pacific Conference on Synthetic Aperture Radar (APSAR). pp. 1–3 (2011)
8. Mullissa, A.G., Persello, C., Stein, A.: PolSARNet: A Deep Fully Convolutional Network for Polarimetric SAR Image Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(12), 5300–5309 (2019). https://doi.org/10.1109/JSTARS.2019.2956650
9. Pingxiang, L., Shenghui, F.: SAR image classification based on its texture features. Geo Spat. Inf. Sci. **6**(3), 16–19 (Jan 2003)
10. Tan, X., Li, M., Zhang, P., Wu, Y., Song, W.: Complex-Valued 3-D Convolutional Neural Network for PolSAR Image Classification. IEEE Geoscience and Remote Sensing Letters **17**(6), 1022–1026 (2020). https://doi.org/10.1109/LGRS.2019.2940387
11. Uhlmann, S., Kiranyaz, S.: Classification of dual- and single polarized SAR images by incorporating visual features. ISPRS Journal of Photogrammetry and Remote Sensing **90**, 10–22 (2014). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2014.01.005, https://www.sciencedirect.com/science/article/pii/S0924271614000094
12. Zhang, L., Xie, W., Zhao, F., Liu, H., Duan, Y.: Deep Learning Based Classification Using Semantic Information for Polsar Image. In: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium. pp. 196–199 (2020). https://doi.org/10.1109/IGARSS39084.2020.9323161

Style Transformation for Text Image with Generation Model

Honghui Yuan[0009−0001−4334−9363]

The University of Electro-Communications, Chofu, Tokyo, JAPAN
yuan-h@mm.inf.uec.ac.jp

1 PhD Context

In my PhD research, I focus on the style transformation of text images with reference images or prompts. Recently, with the success of generation models such as the stable diffusion model, style transformation of text images has achieved promising results. However, style transformation specifically targeting the text parts in scene images while preserving the original content and background has not yet been fully realized. To address this, I propose a novel style transformation method that leverages prompts for more intuitive control over the transfer process. In addition, Existing methods have managed to transfer object font images into specific fonts using reference font images. However, these approaches have primarily focused on modern fonts and have struggled with generating ancient fonts, such as the Japanese Kuzushiji font. To overcome this limitation, I propose a diffusion model-based approach that facilitates the generation of ancient fonts. My PhD under the supervision of Professor Keiji Yanai started in April 2023 and is expected to be completed by March 2026 in an academic context.

2 Main Topic

The primary issue I am addressing in my PhD research is the style transfer of text images. Most existing style transformation methods focus on the entire image, with relatively little research dedicated to transforming specific parts of an image, particularly the text. Due to the unique structure of the text, it is crucial to maintain a balance between readability and stylistic features during the style transfer process. Current research does not adequately address style transformation for the text portion of an image. In addition, in the realm of scene text editing, existing methods are capable of effectively replacing text content without altering the style or background. However, these approaches fail to only change the style of the text while preserving the text content. To tackle this issue, I propose a model that uses prompts to alter the style of the text in scene images, while keeping both the text content and image background unchanged.

Furthermore, most current methods for font image generation primarily focus on modern fonts and use font images as reference styles to convert the target font to the desired font. Unlike modern texts, ancient scripts often feature unique

structures with larger deformations and continuous stroke patterns. As a result, existing methods struggle with the generation of ancient fonts. To address this gap, I propose a diffusion model [6]-based network to overcome the limitations of current methods in ancient font generation.

3 Related Work

Scene text editing has made significant progress in editing text within an image. For example, TextStyleBrush [4] incorporated text image style vectors into the generator, leveraging StyleGAN [3] to guide the generation of final images. TextDiffuser [2] used diffusion models for natural scene text generation and editing in high quality but lacked control over the text's style.

The development of font generation models has achieved promising results in generating various fonts. In recent years, several font generation methods have been introduced to change the font based on reference font images such as FS-Font [7]. However, these methods struggle with issues such as incomplete font strokes and distortion of the font, particularly in fonts with complex structures, such as Chinese. As a result, they have not performed well in generating fonts with complex designs and numerous strokes, such as the Kuzushiji font. FontDiffuser [8] has achieved excellent results in generating over 100 fonts and has obtained better results for handling large deformation styles.

4 Methodology

4.1 Scene text style transfer

Existing scene text editing methods enable the editing of text parts in images in replacement of text content while preventing the style and background unchanged. However, these methods require the reference image and cannot change the style of text arbitrarily. In contrast, our proposed method does not rely on the style reference image and allows users to specify the style of scene text through prompts.

I designed a network [A3] to address the limitations of existing text image editing methods, which cannot perform style transfer on a specific part of the image. An overview of the proposed network is shown in Fig. 1. The proposed network mainly consists of a MaskNet network that extracts the mask image of the text part of the scene text image and a StyleNet network that performs style transformation. Using the pre-trained text image embedding model CLIP [5] and the loss function proposed in this method, the parameters of StyleNet are optimized to apply the style features based on the input prompts, generating the desired styled output.

Loss Function To transform the style in the text region of an image, I introduced the Distance Transform Loss [1] in this study. Specifically, a distance

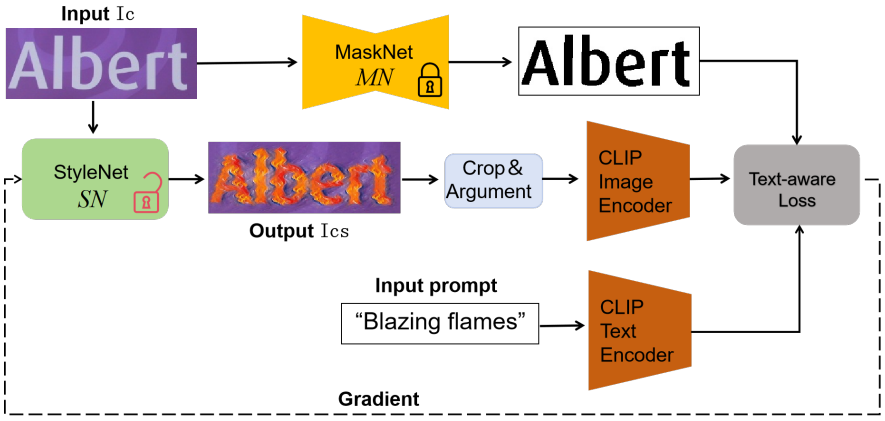


Fig. 1. The framework of our scene text style transfer method.

transform map is generated from the mask image of the text region in the scene text image. Both the input and output images are multiplied by this distance transform map, and the mean squared error (MSE) is calculated.

I also modified the Patch CLIP Loss used in CLIPstyler. Specifically, the background region of the input image is extracted using the text mask, and the cosine similarity between the patches of the generated image and the original background is computed. Patches with significantly different similarities are identified as belonging to the text region, and the Patch CLIP Loss is then calculated only for those patches.

To minimize the influence of the original text style on the generated image and to ensure that the background remains unaltered, I introduced a Background Reconstruction Loss. Specifically, VGG Loss is calculated for the patches corresponding to the background region.

4.2 Kuzushiji font generation

Recent font generation methods have achieved great results in generating various fonts. However, these methods struggle with generating ancient fonts due to the significant differences with modern fonts, such as complex, consecutive stroke structures. To solve this problem, I propose a font generation model [A4] based on the diffusion model that specifically focuses on ancient Japanese Kuzushiji characters.

The overview of our proposed model is shown in Fig. 2. Specifically, the model is divided into three sub-networks, which are the style model, the content model, and the conditional diffusion model. The diffusion model is used to generate the font images, and the style and content models are used as conditions to control the style features and content features during the generation process. Specifically, based on the network of FontDiffuser, I applied a new multiple-heads stroke encoder and the MLP module in the style model to enable our

model to achieve Kuzushiji-style transfer at the stroke level. I also added a patch content discriminator in the content model to ensure the stability and readability of the font structure. Our methods could transform the modern text into corresponding Kuzushiji text in real time, helping to address the digitization challenges of Kuzushiji font characters.

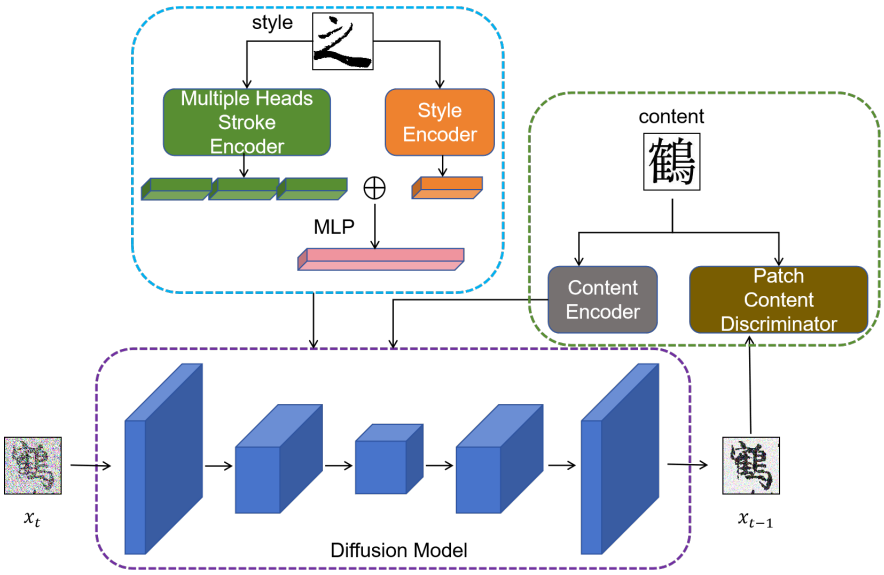


Fig. 2. The network overview of our Kuzushiji font generation method.

Publication list

[A1] Yuan, Honghui, and Keiji Yanai. "Multi-style transfer generative adversarial network for text images." 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2021.

[A2] Yuan, Honghui, and Keiji Yanai. "Multi-Style Shape Matching GAN for Text Images." IEICE TRANSACTIONS on Information and Systems 107.4 (2024): 505-514.

[A3] Yuan, Honghui, and Keiji Yanai. "Font Style Translation in Scene Text Images with CLIPstyler." International Conference on Pattern Recognition (ICPR), 2024.

[A4] Yuan, Honghui, and Keiji Yanai. "KuzushijiDiffuser: Japanese Kuzushiji Font Generation with FontDiffuser." International Conference on Multimodal Modeling (MMM), 2025.

5 Future Work

With the success of Stable Diffusion, using text prompts or reference images to generate high-quality images has yielded impressive results. Recently, several

methods have been developed for generating scene text images using detailed prompts, successfully producing accurate text generation in images. However, these methods are limited to specifying the content, font, and basic color of the text, without the ability to arbitrarily control the style during the generation process. Although our proposed scene text style transfer method can alter the style of the text within an image, it relies on a multi-stage image editing approach. Therefore, in our future work, I aim to focus on text-to-image generation that can generate scene text images with stylized text directly with prompts during the image generation process.

Additionally, existing image editing methods have effectively achieved operations such as object removal, duplication, transfer, enlargement, and reduction. However, performing transformations like scaling and rotating text is often more complex due to the stroke structure of the text. In future research, I plan to explore ways to implement various editing operations specifically for text within images.

References

1. Atarsaikhan, G., Iwana, B.K., Uchida, S.: Contained neural style transfer for decorated logo generation. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 317–322 (2018)
2. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. Advances in Neural Information Processing Systems **36** (2024)
3. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
4. Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., Hassner, T.: Textstylebrush: Transfer of text aesthetics from a single example. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
5. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
7. Tang, L., Cai, Y., Liu, J., Hong, Z., Gong, M., Fan, M., Han, J., Liu, J., Ding, E., Wang, J.: Few-shot font generation by learning fine-grained local styles. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 7895–7904 (2022)
8. Yang, Z., Peng, D., Kong, Y., Zhang, Y., Yao, C., Jin, L.: Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In: Proc. of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6603–6611 (2024)

Deep Learning-Based Framework for DTI Parameters Estimation and Analysis for Sparse Diffusion MRI Data

Abhishek Tiwari[1,2]

[1] Bennett University (The Times Group) India
[2] Shiv Nadar University (Institution of Eminence) India
abhiphd02@gmail.com

Abstract. Today, mental health issues are prevalent, which makes the diagnosis and prognosis of neurological diseases crucial from a clinical perspective. Addressing mental health on a broad scale presents challenges in terms of both cost and time. Although psychiatrists typically address mental health through therapy and counseling, the effectiveness of these approaches varies from person to person. Therefore, noninvasive techniques such as diffusion tensor imaging (DTI) play a vital role in providing quantitative measurements that assist in assessing mental health. Understanding the structure of white matter is key to diagnosing and predicting mental health conditions accurately. Moreover, it is essential to have quantitative measurements that are unbiased and can be deployed on a large scale. These DTI quantitative parameters can be acquired in large scale in less amount of time using sparse diffusion MRI. Sparse diffusion MRI, which can be acquired by small diffusion measurements, presents challenges due to limited diffusion directions and inherent noise. Deep learning has emerged as a promising approach to resolved these problem compare to traditional methods. Introduces a novel Deep Learning Based Framework for DTI Parameter Estimation and Analysis tailored to sparse diffusion MRI data. This framework, incorporating Transformer Neural Network and Convolutional Neural Network (CNN), aims to overcome the limitations of traditional DTI reconstruction methods. We conducted experiments on various datasets, including the Human Connectome Project (HCP) which is high resolution, the MICCAI Quad22 Migraine dataset, the National Institute of Mental Health Data (NIFD), and the Alzheimer's Disease Neuroimaging Initiative (ADNI) which are mental health diseases with lower resolution. Our findings show that our framework effectively improves DTI parameter estimation and analysis for sparse diffusion MRI data. These results contribute to advancing our understanding of brain connectivity and neurodegenerative diseases, with implications for future neuroimaging research.

Supervisors: Dr. Rajeev Kumar Singh, Dr. Saurabh J. Shigwan
Start Date: 17 August 2020, Completion Date: 26 May 2024
PhD is in an industrial and/or academic context: Abhishek's PhD bridges industry and academia, applying deep learning to Diffusion Tensor Imaging (DTI) for clinical and research applications.

Keywords: Diffusion MRI · Deep learning · Tractography · Neurological disorders

1 Problem Statement

The growing prevalence of mental health disorders necessitates the development of accurate and scalable diagnostic tools for neurodegenerative diseases. Diffusion Tensor Imaging (DTI), a non-invasive method, is instrumental in understanding the brain's white matter structure, which is vital for diagnosing neurological and mental health conditions. However, traditional DTI techniques face limitations, including long scan times, restricted diffusion directions, and noise sensitivity, making large-scale clinical deployment challenging. Addressing these issues through deep learning methodologies could transform mental health diagnosis by providing faster, more accurate, and scalable solutions.

2 State of the Art

Diffusion MRI (dMRI) has emerged as a powerful tool in neuroimaging, offering insights into white matter pathways in the brain(1). However, due to limitations in current techniques, researchers have explored deep learning approaches to improve data quality and reduce scan time(5). Sparse diffusion MRI is one approach to expedite data acquisition, though it presents challenges in terms of noise and incomplete information(4). Traditional DTI parameter estimation methods, such as tensor fitting and model-based approaches, are computationally intensive and often inadequate for sparse data(2).

Recent advances in deep learning have provided promising results for improving the accuracy of DTI parameter estimation from sparse data(3; 7). Specifically, Transformer models, known for their self-attention mechanisms, have shown efficacy in capturing long-range dependencies in data, while CNNs have demonstrated success in processing spatial features in medical imaging(6; 8; 9).

3 Methodology

The methodology employed in thesis focuses on the development and application of advanced deep learning techniques for the estimation and analysis of diffusion tensor imaging (DTI) parameters from sparse diffusion MRI data. The key components of the methodology include:

Innovative DTI Estimation Method (SwinDTI) The thesis introduces *SwinDTI*, a novel approach that utilizes a transformer neural network to efficiently estimate diffusion tensor parameters from sparse diffusion-weighted imaging (DWI) data. This method addresses the limitations of traditional techniques, such as prolonged scan times and restricted generalization across various diffusion directions.

– SwinDTI leverages sophisticated attention mechanisms and patch-based processing, which enhances the speed and accuracy of DTI estimation.

Assessment of Deep Learning Techniques The research evaluates the effectiveness of various deep learning methods in improving the quality of quantitative measures derived from sparse measurements. This assessment is particularly focused on decision-making processes related to chronic and episodic migraines.

– The study investigates the impact of angular resolution in diffusion MRI and challenges multiple teams to utilize deep learning techniques to enhance diffusion metrics.

Framework for Frontotemporal Dementia (FTD) Diagnosis The thesis proposes a deep learning framework that employs the Swin-Transformer architecture to improve the diagnostic accuracy of frontotemporal dementia (FTD). This framework utilizes sparse diffusion measures from neuroimaging data, significantly reducing scanning time while enhancing diagnostic capabilities.

Accelerated Alzheimer's Disease Diagnosis Model An additional contribution includes the introduction of a model aimed at accelerating the diagnosis of Alzheimer's disease, further showcasing the versatility and applicability of the Swin-Transformer in neuroimaging contexts.

4 Contributions

The contributions of thesis can be summarized as follows:

Advancement in Neuroimaging Techniques The introduction of the SwinDTI method represents a significant advancement in neuroimaging, providing a fast and accurate means of estimating DTI parameters from sparse data, which is crucial for clinical applications.

Enhancement of Quantitative Measures The thesis contributes to the understanding of how deep learning can enhance the quality of quantitative measures in medical imaging, particularly in the context of migraine diagnosis, thereby improving decision-making processes in clinical settings.

Improved Diagnostic Framework for FTD The development of a deep learning framework for FTD diagnosis not only enhances diagnostic accuracy but also reduces the time required for scanning, which is a critical factor in clinical practice.

Insights into AI Applications in Medical Imaging The research provides valuable insights into the challenges and potential of applying AI methodologies in medical imaging, emphasizing the importance of maintaining data integrity and accuracy while utilizing advanced techniques.

Contribution to Neurodegenerative Disorder Diagnostics Overall, the thesis significantly enhances the accuracy, efficiency, and reliability of neuroimaging analysis, paving the way for improved diagnostic capabilities and better treatment outcomes in the field of neurodegenerative disorders.

5 Post-PhD Career Plan

After completing the PhD, the goal is to transition into a research-centric role in academia or industry. The focus will be on the continued development of AI-based diagnostic tools in neuroimaging, with applications extending beyond mental health to other neurodegenerative and neurological diseases.
Postdoctoral Research Plan to pursue a postdoctoral position specializing in AI applications in neuroimaging. This will include continued work on improving diagnostic accuracy and exploring applications in novel brain imaging techniques.
Academic Career Long-term, I aim to secure a faculty position to lead independent research groups that focus on AI-driven healthcare innovations, specifically in neurology and psychiatry.
Industry Collaboration Simultaneously, I seek to collaborate with tech and healthcare industries to translate research into commercially viable diagnostic products that can have an immediate clinical impact.

List of Publications

1. **Abhishek Tiwari**, Rajeev Kumar Singh, and Saurabh J. Shigwan, "SwinDTI: Swin transformer based fast estimation of diffusion tensor parameters from sparse data", **Neural Computing and Applications Springer journal Impact Factor = 4.5 Q1 SCI Journal** (2023) [Paper]

2. Santiago Aja-Fernández, **Abhishek Tiwari**, Saurabh J. Shigwan, Rajeev Kumar Singh, Tianshu Zheng, "Validation of Deep Learning techniques for quality augmentation in diffusion MRI for clinical studies", **Elsevier NeuroImage: Clinical journal Impact Factor = 3.4 Q1 SCI Journal** (2023) [Paper]

3. **Abhishek Tiwari**, Ananya Singhal, Saurabh J. Shigwan, Rajeev Kumar Singh, "Deep Learning Framework using Sparse Diffusion MRI for Diagnosis of Frontotemporal Dementia", **IEEE/CVF International Conference on Computer Vision ICCV 2023 (Core Rank A*, h-index 228) workshop on BioImage Computing. [ICCV 2023 Paris, Acceptance rate: 2160 / 8068 = 26.8%]** [Paper]

4. **Abhishek Tiwari**, Ananya Singhal, Saurabh J. Shigwan, Rajeev Kumar Singh,"Early Diagnosis of Alzheimer through Swin-Transformer-Based Deep Learning Framework using Sparse Diffusion Measures",**The 15th Asian Conference on Machine Learning (ACML 2023) in İstanbul, Turkey Acceptance rate = 26.9%, with just four papers selected from India**. [Paper]

Bibliography

[1] Guevara, M., Guevara, P., Román, C., Mangin, J.F.: Superficial white matter: A review on the dmri analysis methods and applications. NeuroImage **212**, 116673 (2020)

[2] Jian, B., Vemuri, B.C.: A unified computational framework for deconvolution to reconstruct multiple fibers from diffusion weighted mri. IEEE transactions on medical imaging **26**(11), 1464–1471 (2007)

[3] Karimi, D., Jaimes, C., Machado-Rivas, F., Vasung, L., Khan, S., Warfield, S.K., Gholipour, A.: Deep learning-based parameter estimation in fetal diffusion-weighted mri. Neuroimage **243**, 118482 (2021)

[4] Lustig, M., Donoho, D., Pauly, J.M.: Sparse mri: The application of compressed sensing for rapid mr imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **58**(6), 1182–1195 (2007)

[5] Shamir, I., Assaf, Y.: Tutorial: a guide to diffusion mri and structural connectomics. Nature Protocols pp. 1–19 (2024)

[6] Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis **88**, 102802 (2023)

[7] Tian, Q., Bilgic, B., Fan, Q., Liao, C., Ngamsombat, C., Hu, Y., Witzel, T., Setsompop, K., Polimeni, J.R., Huang, S.Y.: Deepdti: High-fidelity six-direction diffusion tensor imaging using deep learning. NeuroImage **219**, 117017 (2020)

[8] Tiwari, A., Singh, R.K., Shigwan, S.J.: Swindti: swin transformer-based generalized fast estimation of diffusion tensor parameters from sparse data. Neural Computing and Applications **36**(6), 3179–3196 (2024)

[9] Tiwari, A., Singhal, A., Shigwan, S.J., Singh, R.K.: Deep learning framework using sparse diffusion mri for diagnosis of frontotemporal dementia. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3821–3827 (2023)

Towards Efficient Machine Learning Approach for Stock Price Movement Prediction

Manali Patel

Sardar Vallabhbhai National Institute of Technology, Surat, India
d21co004@coed.svnit.ac.in

1 PhD context

My PhD research is supervised by Dr. Krupa Jariwala (Assistant Professor, Sardar Vallabhbhai National Institute of Technology, Surat, India) and Dr. Chiranjoy Chattopadhyay (Associate Professor, FLAME University, Pune, India). I commenced my PhD studies on 06-08-2021, and the expected completion date is yet to be decided. This PhD is conducted within an academic context, focusing on the development and application of advanced deep learning techniques for stock market prediction.

2 Problem statement

Stock market movement prediction is a challenging task due to considerable volatility and unpredictability induced by numerous factors. The recent approaches consider stocks as individual entities and neglect the relationships between stocks that can explain external events efficiently. To overcome this, we propose three different types of relational approaches, i.e., sequential, knowledge graph-based, and hybrid, that efficiently consider the intra and inter-sector associations and enhance the prediction accuracy. The applicability of the approaches is tested in real life applications such as portfolio optimization or trading strategies. The research problem is formulated as:

Given a tensor $\chi_t = \{X_t^1, X_t^2, \ldots, X_t^N\} \in \mathbb{R}^{N \times p \times d}$ of collected features of N stocks, where p and d represent lag value and feature dimension, respectively. An adjacency matrix $A \in \mathbb{R}^{N \times N}$ defines the relationship strength between stock pairs. Our objective is to predict the future price movement (up (1) and down (-1)) of N stocks at time instance $t+1$ as:

$$\hat{Y}_{t+1} = F(\chi_t, A) \in \mathbb{R}^N \tag{1}$$

where $F(.)$ is a user-defined prediction function.

3 State-of-the-art

Figure 1 depicts the road map of stock market forecasting approaches being applied over a period of time. The statistical approaches [8, 9] have been replaced

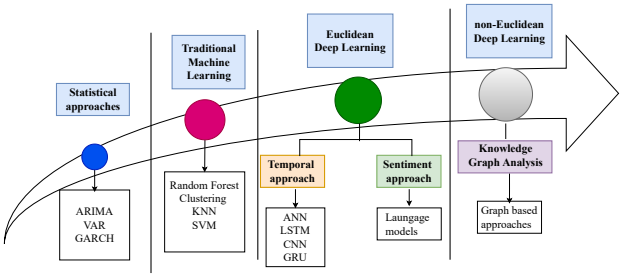


Fig. 1. Evolution of stock market forecasting approaches.

by traditional machine learning (ML) approaches that have the ability to capture non-linear patterns [4,5,7,10]. A subset of ML approaches known as deep learning (DL) techniques have outperformed traditional approaches for stock market prediction due to their self-adaptive and feature extraction abilities. For stock market forecasting, the DL methods are further classified into two categories based on the data representation, i.e., Euclidean and non-Euclidean approaches. The Euclidean temporal DL approaches have been applied to sequential market data or stock images to capture the long range dependency [1–3, 12]. With the rise of online social platforms such as Twitter, online forums, and financial reports, specialized deep learning approaches have been applied to perform sentiment analysis and predict stock prices [6]. With the advent of advanced DL models to process the non-Euclidean geometry space, graph based methods for the stock market have seen a rise in recent years [11, 13]. The research question is modeled as **"How efficiently can a model capture the relational as well as temporal dependency hidden in the data?"**

4 Methodology and Contributions

4.1 Contribution 1: Review paper

Contribution 1.1: A Systematic Review on Graph Neural Network-based Methods for Stock Market Forecasting
Published in: *Patel, M., Jariwala, K., Chattopadhyay, C.: A systematic review on graph neural network-based methods for stock market forecasting. ACM Comput. Surv. 57(2) (Oct 2024).*
1) This article comprehensively reviews and discusses graph-based prediction models for the stock market, categorized by their specific tasks, such as classification, regression, or stock recommendation.
2) This article discusses different data sources, datasets, and evaluation parameters for stock market forecasting.
3) The results from each reviewed article are highlighted.
4) It also discusses open issues and future directions to guide further research in this domain.

4.2 Contribution 2: Introduction of Sequential Relational Features

Contribution 2.1: SM2PNet: A Deep Learning Approach Towards Indian Stock Market Movement Prediction Published in: *Patel, M., Jariwala, K., Chattopadhyay, C.: SM2PNet: A Deep Learning Approach Towards Indian Stock Market Movement Prediction. In: 8th International Conference for Convergence in Technology (I2CT). pp. 1–7 (2023).*
1) We created a model, SM2PNet (Stock Market Movement Prediction Network), which incorporates the characteristics of highly correlated **large capitalization intra-sector stocks** to predict the future trend of the target stock.
2) The top-K companies having a **higher correlation value** with the target company are selected and their closing prices are considered as the relational features and given to the LSTM model.

Contribution 2.2: MSNet: Momentum Spillover Network for Indian Stock Market Movement Prediction
Published in: *Patel, M., Jariwala, K., Chattopadhyay, C.: MSNet: Momentum Spillover Network for Indian Stock Market Movement Prediction. In: 2023 IEEE 20th India Council International Conference (INDICON). pp. 615–620 (2023).*
1) We collected the information about **large, mid, and small capitalization intra-sector stocks** from various sources that was not readily available and proposed a unique data-driven approach, MSNet (Momentum Spillover Network) to incorporate relational dependency.
2) The top K companies selected by the data-driven measure, i.e., **Random Forest Regressor** are used to update the target firm's embedding by concatenating the closing prices of K companies and given to the **Temporal Convolutional Network (TCN)** to predict the future trend.

4.3 Contribution 3: Introduction of Graph Based Relational Feature

Contribution 3.1: Advancing Indian Stock Market Prediction with TRNet: A Graph Neural Network Model
Published in: *Patel, M., Jariwala, K., Chattopadhyay, C.: Advancing Indian Stock Market Prediction with TRNet: A Graph Neural Network Model. In: International Conference on Modeling, Simulation Intelligent Computing (MoSICom). pp. 173–178 (2023).*
1) We proposed the TRNet (Temporal Relational Network) model, which incorporated **industry-sector classification data** to construct a financial knowledge graph for the constituent stocks of the NIFTY-50 index.
2) The Graph Convolution Network (GCN) is utilized to exploit the relational dependency from the built financial graph to predict future price movements.

Contribution 3.2: An Approach Towards Stock Market Prediction and Portfolio Optimization in Indian Financial Sectors
Published in: *Patel, M., Jariwala, K., Chattopadhyay, C.: An Approach Towards Stock Market Prediction and Portfolio Optimization in Indian Financial*

Sectors. IEEE Transactions on Computational Social Systems pp. 1–12 (2024). https://doi.org/10.1109/TCSS.2024.3450291

1) From the review done in 4.1, we identified that the existing graph approaches have considered pre-defined relationships to build the financial knowledge graph. The pre-defined relationships are static in nature, thus failing to capture the dynamic, latent interactions between stock pairs.

2) To overcome this, we have proposed the **D**ynamic **R**elation aware **R**elational **T**emporal **N**etwork (DR2TNet) prediction framework as shown in Figure 2, which does not rely on pre-defined relationships and learns the positive as well as negative **intra and inter-sector associations** in a dynamic manner.

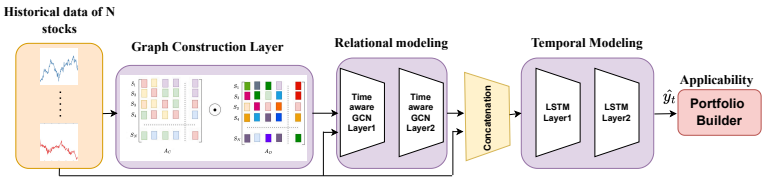


Fig. 2. Proposed DR2TNet model.

3) A time-aware GCN is utilized to capture the dynamic relationships between stock pairs defined by the built financial graph, and a LSTM is applied to learn the temporal patterns.

4) A new loss function is proposed, and the applicability of DR2TNet is also demonstrated in the **portfolio optimization** problem.

4.4 Contribution 4: A Hybrid Relational Approach Towards Stock Price Prediction and Profitability

Published in: *Patel, M., Jariwala, K., Chattopadhyay, C.: A Hybrid Relational Approach Towards Stock Price Prediction and Profitability. IEEE Transactions on Artificial Intelligence pp. 1–10 (2024). https://doi.org/10.1109/TAI.2024.3408129.*

1) We proposed the **hybrid RF2P-TCLM model** as depicted in Figure 3, consisting of a Temporal Convolution Network (TCN) and a linear model (LM) to account for the linear and non-linear components in stock prices.

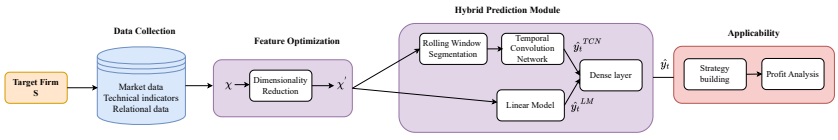


Fig. 3. A schematic representation of the proposed methodology RF2P-TCLM.

2) To prove the generalizability of the RF2P-TCLM model, we experimented with the stocks from the American, Indian, and Korean economies.
3) The real-world validation of the proposed approach is done by developing a **trading strategy** and yielding higher profitability.

5 Future work

1) Considering the volatile nature of the stock data, we will integrate diverse sources of information, such as candlestick charts and sentiment data, along with the relational aspects to enhance the prediction performance.
2) Additionally, we will delve into interpretability aspects and understanding model predictions to enhance user trust.

References

1. Abdelfattah, B.A., Darwish, S.M., Elkaffas, S.M.: Enhancing the prediction of stock market movement using neutrosophic-logic-based sentiment analysis. Journal of Theoretical and Applied Electronic Commerce Research **19**(1), 116–134 (2024)
2. Ali, M., Khan, D.M., Aamir, M., Ali, A., Ahmad, Z.: Predicting the direction movement of financial time series using artificial neural network and support vector machine. Complexity **2021**, 1–13 (2021)
3. Chen, C., Xue, L., Xing, W.: Research on improved gru-based stock price prediction method. Applied Sciences **13**, 8813 (07 2023)
4. Gao, S.: Trend-based k-nearest neighbor algorithm in stock price prediction. In: 3rd International Conference on Digital Economy and Computer Application (DECA 2023). pp. 746–756. Atlantis Press (2023)
5. Illa, P.K., Parvathala, B., Sharma, A.K.: Stock price prediction methodology using random forest algorithm and support vector machine. Materials Today: Proceedings **56**, 1776–1782 (2022)
6. Koukaras, P., Nousi, C., Tjortjis, C.: Stock market prediction using microblogging sentiment analysis and machine learning. Telecom **3**(2), 358–378 (2022)
7. Liagkouras, K., Metaxiotis, K.: Stock Market Forecasting by Using Support Vector Machines, pp. 259–271. Springer International Publishing, Cham (2020)
8. Mohankumari, C., Vishukumar, M., Chillale, N.: Analysis of daily stock trend prediction using arima model. International Journal of Mechanical Engineering and Technology **10**, 1772–1792 (01 2019)
9. Naik, N., Mohan, B., Jha, R.: Garch model identification for stock crises events. Procedia Computer Science **171**, 1742–1749 (01 2020)
10. Sáenz, J.V., Quiroga, F.M., Bariviera, A.F.: Data vs. information: Using clustering techniques to enhance stock returns forecasting. International Review of Financial Analysis **88**, 102657 (2023)
11. Sattiraju, S.A., Chakraborty, A., Shaijumon, C.S., Manoj, B.S.: Corporate linkages and financial performance: A complex network analysis of indian firms. IEEE Transactions on Computational Social Systems **7**(2), 339–351 (2020)
12. Wojarnik, G.: The potential of convolutional neural networks for the analysis of stock charts. Procedia Computer Science **225**, 941–950 (2023)
13. Xu, H., Zhang, Y., Xu, Y.: Promoting financial market development-financial stock classification using graph convolutional neural networks. IEEE Access **11**, 49289–49299 (2023)

GAN and DM Generated Synthetic Image Detection in the Age of Misinformation

Tanusree Ghosh

Department of Information Technology
Indian Institute of Engineering Science and Technology, Shibpur, India
2021itP001.tanusree@students.iiests.ac.in

PhD Context

The research presented in this paper is part of my PhD, which is being conducted under the guidance of Dr. Ruchira Naskar. I began my PhD in February 2022, with an expected completion date of December 2025. The PhD is undertaken in an academic context, focusing on the detection of synthetic media, specifically targeting content generated by Generative Adversarial Networks (GAN) and Diffusion Models (DM) in the context of online social networks. The findings are significant for the industry as well, which could support the development of commercial tools to detect fake content to preserve the integrity of social networks by preventing the propagation of misinformation.

Introduction and Problem Identification

In recent years, the rapid development of Generative Artificial Intelligence technologies, such as Generative Adversarial Networks (GANs) and Diffusion Models (DM), has brought forth a new era of hyper-realistic synthetic images. While these advancements have enriched industries like entertainment and gaming, they have been used to spread misinformation in Online Social Networks (OSNs), which ushered in a formidable social peril[1] [2]. Fake face images are widely used as profile pictures of fake social media profiles that eventually propagate fake news or scam. On the other hand, photorealistic synthetic images are used with fake news to increase their trustworthiness, as shown in Fig. 1.

Hence, the problem at hand is the detection of synthetic images. While it is already a challenging task to detect such images due to their photo-realism and similar statistical properties to camera-generated images, in real-life scenarios, an ideal fake image detector should be able to detect OSN circulated images that go through unknown OSN-specific post-processing changes and should detect any given fake images, as it is impossible to know prior from which generative model a given image possibly could belong. Hence, with correct detection, robustness and generalisation are crucial for synthetic image detectors.

[1] https://www.livemint.com/news/world/aigenerated-image-of-explosion-at-pentagon-goes-viral-creates-chaos-in-stock-market-see-here-11684774645223.html

[2] https://www.cbsnews.com/news/is-that-facebook-account-real-meta-reports-rapid-rise-in-ai-generated-profile-pictures/

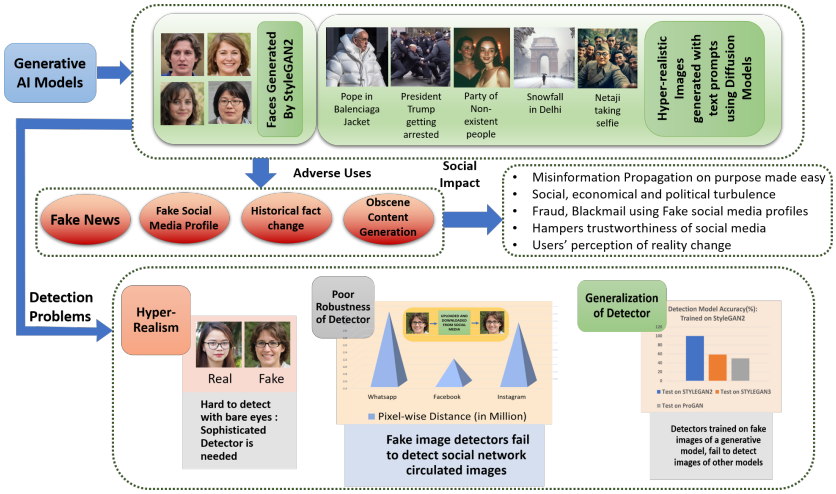


Fig. 1: Current Scenario of Detecting AI-Generated Images: Advanced generative AI technologies like StyleGAN2 and Diffusion Models are creating highly realistic synthetic images with ease. These images are frequently misused by bad actors for spreading misinformation, impersonating others on social media, or creating damaging content aimed at individuals. Such practices are undermining the credibility of social networks, leading to increased incidents of fraud, and causing social and economic unrest. The primary challenge in identifying these images is their hyper-realistic appearance, which often deceives the human eye. While some cutting-edge detection methods boast high accuracy rates, their effectiveness diminishes when applied to images altered by social media platforms. Additionally, these detectors struggle with the 'Generalization Problem', failing to recognize synthetic images produced by previously unseen generative models.

Potential Solution Approaches

Initially, GAN-generated images were identifiable through visible artifacts [5,8,6] like mismatched eye colors, irregular pupil shapes, and inconsistent corneal highlights, allowing trained observers to spot fakes. However, advancements in GANs have erased these artifacts, reducing the observer's ability to discern synthetic images, and increasing susceptibility to misinformation.

Existing detectors fall into two categories: Deep Learning (DL) based, using complex feature sets derived automatically, and those using selected features, based on statistical properties and machine learning classifiers, like hand-crafted features in colour and frequency domains [7,9]. DL methods typically employ models like VGG-Net and modified Xception-Net [2]. However, using these directly in GAN-generator models can render synthetic images undetectable. On the other hand, selected features-based detectors use domain knowledge, such as statistical properties of facial landmarks with classifiers like Support Vector Machines, co-occurrence matrices, cross co-occurrence matrices [9,1,10], etc. Re-

cently, combining statistical properties with Deep Neural Network classifiers has proven effective in detecting synthetic images, especially in the context of OSN, an approach we adopt in our work.

Detection of fake images can broadly be viewed as efficient coordination of two sub-modules, identification of differentiating features and classifier, as shown in Fig. 2. Here it is evident that in certain feature spaces, they have visible differences. Here we have shown, a pixel-wise average of 500 gradient magnitude and direction and luminosity components for both real and fake images.

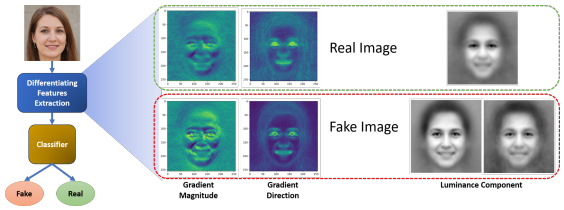


Fig. 2: Broad overview of working of fake image detector.

Proposed Solutions

Our synthetic face detection solutions involve a feature extractor and a classifier. The proposed solutions are:

- **Relative Chrominance-Based Detection** [3]: RCD-Net uses Relative Chrominance Distance in various color spaces (LAB, YUV, HLS) with High-Frequency Residual (HFR) and grayscale features for enhanced detection. Achieves 99.20% accuracy for StyleGAN2 images with high efficiency.
- **Gradient-Based Detection** [4]: Uses gradient magnitude and direction matrices (Sobel, Scharr) for detection. Combines GM, GD, GF-Net, and DCG-Net for feature extraction, followed by a CNN classifier.
- **STN-Net** [5]: In this work, we propose Sine Transformed Noise (STN) for feature extraction using Gaussian and Laplacian filters, enhancing detection through Sine Transformations followed by a custom CNN classifier.

[3] Tanusree Ghosh, and Ruchira Naskar. "Less is more: A minimalist approach to robust GAN-generated face detection." Pattern Recognition Letters 179 (2024): 185-191.

[4] Tanusree Ghosh and Ruchira Naskar. "Leveraging Image Gradients for Robust GAN-Generated Image Detection in OSN context." 2023 IEEE International Conference on Visual Communications and Image Processing (VCIP). IEEE, 2023.

[5] Tanusree Ghosh, and Ruchira Naskar. "STN-Net: A Robust GAN-Generated Face Detector." International Conference on Information Systems Security. Cham: Springer Nature Switzerland, 2023.

Table 1: Detection Performance Comparison on Test Set. Metric: Accuracy (%)

Metric	STN-Net (Best Model)	GF-Net (Best Model)	Qiao[11]	Nowroozi [10]	Nataraj [9]	Chen [3]	Frank [4]	Yu [12]	Li [7]
Accuracy (%)	99.53	99.80	99.95	99.33	96.11	97.70	98.58	97.70	99.50

Table 2: Robustness testing with various post-processing operations.

Operations	Parameters	Gradient based Variants								Qiao [11]	Nataraj [9]	Nowroozi [10]	STN-Net Variants		
		GM-Net		GD-Net		GF-Net		DCG-Net					Baseline (STN-Net)	STN-Net	Dual layer STN-Net
		Sobel	Scharr	Sobel	Scharr	Sobel	Scharr	Sobel	Scharr						
Median Filter	3 x 3	98.20	98.36	97.54	97.57	98.56	98.45	97.54	98.54	99.35	81.48	85.13	91.62	96.92	98.78
	5 x 5	88.88	91.96	80.53	85.62	88.12	91.89	85.57	92.53	93.80	75.98	83.65	91.12	63.77	78.92
Gaussian Noise	1.0	99.04	98.96	99.13	98.74	99.18	98.89	98.54	98.88	94.43	76.35	93.68	91.82	99.21	99.38
	2.0	99.01	98.93	99.21	98.55	99.02	98.83	98.61	98.98	74.25	76.73	96.80	91.84	99.28	99.33
CLAHE	3 x 3	98.74	98.86	99.08	98.86	98.86	98.81	98.66	98.81	94.70	51.43	50.32	85.54	97.10	98.99
Average Blurring	3 x 3	76.54	88.57	96.35	90.80	98.09	97.88	97.47	96.88	97.30	93.68	86.90	91.37	86.11	94.84
	5 x 5	50.02	50.19	70.14	85.57	75.15	82.99	81.85	78.89	82.68	88.23	76.63	90.40	50.12	53.77
Gamma Correction	0.8	98.74	98.93	99.28	98.71	99.12	98.99	98.51	98.99	95.08	82.28	86.90	90.10	99.28	99.26
	0.9	98.96	98.91	99.13	98.59	99.11	99.03	98.56	99.01	98.00	87.23	90.98	91.69	99.23	99.45
	1.2	98.88	98.93	99.08	98.66	98.96	98.99	98.66	98.91	96.90	87.20	85.53	89.56	99.40	99.36
Resizing	0.5	63.44	79.69	95.66	90.50	97.95	97.27	96.92	95.24	79.80	57.93	92.47	91.42	72.40	86.31
Average	-	88.22	91.11	94.10	94.74	95.63	**96.54**	95.53	95.96	91.48	78.04	84.11	90.58	87.53	**91.67**

Table 3: Robustness testing with various JPEG compression factors.

Quality Factor	STN-Net Variants			Qiao [11]	Nowroozi [10]	Nataraj [9]	Gradient-based Variants							
	Baseline (STN Net)	STN-Net	Dual-Layer STN-Net				GM-Net		GD-Net		GF-Net		DCG-Net	
							Sobel	Scharr	Sobel	Scharr	Sobel	Scharr	Sobel	Scharr
90	91.38	99.33	99.33	97.53	94.50	95.58	98.71	98.99	99.40	98.56	98.96	98.89	98.67	98.99
80	91.37	98.71	99.08	97.44	88.66	94.93	98.49	98.76	99.06	98.34	98.83	98.81	98.39	98.60
70	91.29	97.99	98.69	97.23	83.50	94.03	98.41	98.54	98.77	98.29	98.59	98.55	98.19	98.44
60	91.34	96.90	98.26	96.83	94.00	94.65	98.39	97.94	98.26	97.42	98.39	97.99	97.77	98.16
50	91.22	96.88	**97.89**	96.51	80.05	96.66	98.19	97.30	97.69	96.60	**98.31**	97.79	95.12	97.72

– **LPQ-Net** [6]: Uses Local Phase Quantization (LPQ) with a lightweight CNN classifier, achieving 99.98% accuracy for StyleGAN2 images, and robust performance on OSN images.
– **Transfer learning-based solution for DM image detection** [7]: In this work, we use the ResNet-50 backbone, fine-tuned with a custom classification head to achieve over 96% detection accuracy and 93% source attribution accuracy for Diffusion Model images.

Comparison of our few proposed methods with other state-of-the-art solutions for real vs. StyleGAN2-generated synthetic face image classification are shown in Tab. 1. Robustness comparison of our proposed solutions, that mimic OSN scenario are present in Tab. 2), and Tab. 3.

Conclusion and Future Work Direction

Our existing techniques for identifying GAN-generated facial images demonstrate remarkable effectiveness when tested on synthetic images from the same

[6] Srijit Kundu, Tanusree Ghosh, and Ruchira Naskar. "Using Local Phase Quantization to Identify Fake Faces in Online Social Networks", IEEE Region 10 Conference 2024 (Tencon 2024). (*To appear*)
[7] Sanandita Das, Dibyarup Dutta, Tanusree Ghosh, and Ruchira Naskar. "Universal Detection and Source Attribution of Diffusion Model Generated Images with High Generalization and Robustness." In International Conference on Pattern Recognition and Machine Intelligence, pp. 441-448. Cham: Springer Nature Switzerland, 2023.

dataset. These methods focus on identifying features that withstand common transformations in online social networks (OSNs).

Moving forward, our research aims to enhance both the generalization and robustness of these detection methods. With the proliferation of diverse generative models, the creation of a universal detector capable of recognizing synthetic images from any model is increasingly crucial.

Additionally, there is an unexplored potential in examining artifacts produced by Diffusion Model (DM) generated images across various domains, such as spatial and frequency, which warrants further investigation.

References

1. Barni, M., Kallas, K., Nowroozi, E., Tondi, B.: Cnn detection of gan-generated face images based on cross-band co-occurrences analysis. In: 2020 IEEE international workshop on information forensics and security (WIFS). pp. 1–6. IEEE (2020)
2. Chen, B., Ju, X., Xiao, B., Ding, W., Zheng, Y., de Albuquerque, V.H.C.: Locally gan-generated face detection based on an improved xception. Information Sciences **572**, 16–28 (2021)
3. Chen, B., Liu, X., Zheng, Y., Zhao, G., Shi, Y.Q.: A robust gan-generated face detection method based on dual-color spaces and an improved xception. IEEE Transactions on Circuits and Systems for Video Technology **32**(6), 3527–3538 (2021)
4. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning. pp. 3247–3258. PMLR (2020)
5. Guo, H., Hu, S., Wang, X., Chang, M.C., Lyu, S.: Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2904–2908. IEEE (2022)
6. Hu, S., Li, Y., Lyu, S.: Exposing gan-generated faces using inconsistent corneal specular highlights. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2500–2504. IEEE (2021)
7. Li, H., Li, B., Tan, S., Huang, J.: Identification of deep network generated images using disparities in color components. Signal Processing **174**, 107616 (2020)
8. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). pp. 83–92. IEEE (2019)
9. Nataraj, L., Mohammed, T.M., Chandrasekaran, S., Flenner, A., Bappy, J.H., Roy-Chowdhury, A.K., Manjunath, B.: Detecting gan generated fake images using co-occurrence matrices. arXiv preprint arXiv:1903.06836 (2019)
10. Nowroozi, E., Mekdad, Y.: Detecting high-quality gan-generated face images using neural networks. Big Data Analytics and Intelligent Systems for Cyber Threat Intelligence pp. 235–252 (2023)
11. Qiao, T., Chen, Y., Zhou, X., Shi, R., Shao, H., Shen, K., Luo, X.: Csc-net: Cross-color spatial co-occurrence matrix network for detecting synthesized fake images. IEEE Transactions on Cognitive and Developmental Systems (2023)
12. Yu, Y., Ni, R., Zhao, Y.: Mining generalized features for detecting ai-manipulated fake faces. arXiv preprint arXiv:2010.14129 (2020)

Identifying Deepfakes based on Human Physiological Signals

Rajat Chakraborty[0000−0002−1653−2699]

Department of Information Technology
Indian Institute of Engineering Science and Technology, Shibpur, India
rajat.rs2023@it.iiests.ac.in

1 PhD Context

I enrolled in the PhD program in August 2023 under the supervision of Dr. Ruchira Naskar, with an expected completion date of June 2027. My PhD research focuses on recent advancements in robust deepfake detection systems that utilize remotely extracted human physiological signals from facial videos. I am exploring the application of Time Frequency (TF) algorithms to these signals to enhance detection resilience and am working toward developing a novel technique for remote heart rate estimation from facial videos, contributing a new dimension to deepfake detection methodologies. Given the significant impact of deepfakes on the legal, media, and broadcasting industries, developing robust and accurate detection methods is need of the hour.

2 Problem Statement

"Deepfakes" are hyper-realistic audio, video, or multimedia content generated by deep learning models. Since their emergence around 2017, deepfake technologies have quickly gained traction in media, film, and e-commerce.

Current Challenges in Deepfake Detection: As deepfake technology advances, creating hyper-realistic content through methods like face swapping [9], facial reenactment [5], attribute editing [8], and face synthesis [4] is rapidly outpacing traditional detection methods. Figure 1 shows outputs from these tools. Face swapping maps a target face onto a source for seamless expression integration, while reenactment transfers one person's expressions to another. Attribute editing, often achieved through GANs, modifies features like hair color and age, and face synthesis creates entirely new, realistic identities. While some detectors perform well on simpler manipulations, they struggle against sophisticated deepfakes that minimize visual artifacts, especially across demographic and cultural variations. This calls for adaptive detection methods to keep pace with evolving synthetic media.

Need for Physiological Signal based Deepfake Detection: Physiological signals, such as subtle changes in heart rate, eye movements, and auditory or facial muscle features, offer promising indicators for deepfake detection, as

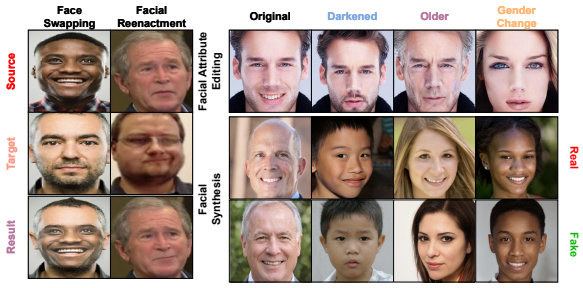


Fig. 1. Face Swapping Method, Facial Reenactment Method, Facial Attribute Editing Technique and Facial Synthesis Techniques.

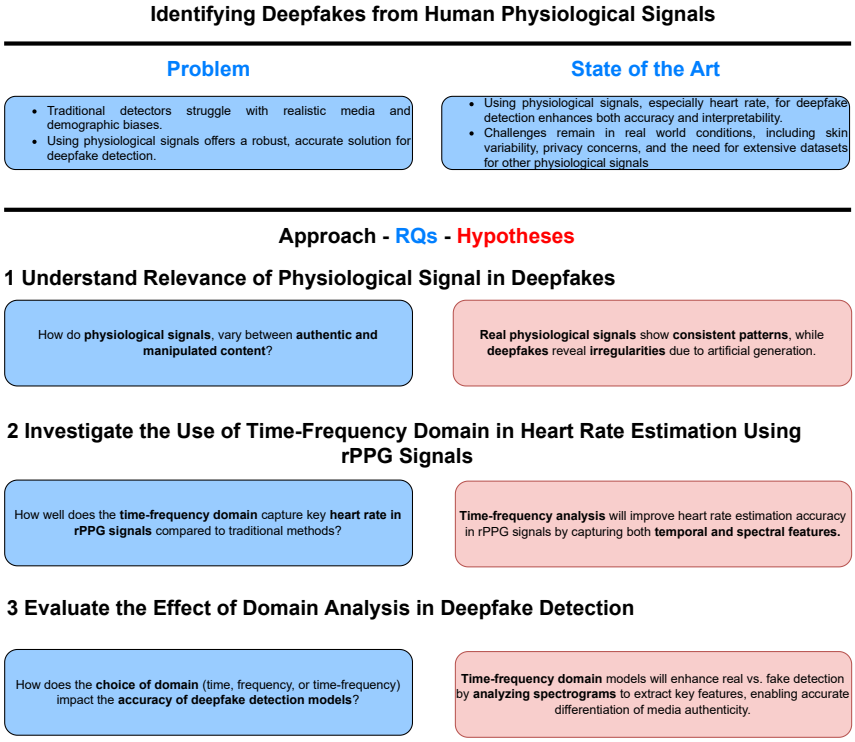


Fig. 2. Overview of the Three Approach Framework for Deepfake Detection.

they are inherently difficult for AI to replicate. This approach can improve accuracy, robustness, reduce demographic biases, and address key limitations in traditional detection methods.

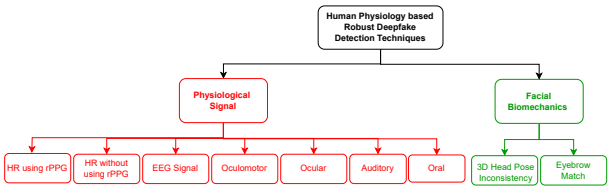


Fig. 3. Illustration of Physiological Signals and Facial Biomechanics for Deepfake Detection.

Research Objective: This study aims to develop a physiological signal based deepfake detection method that remains resilient against adversarial attempts, accurately detects deepfakes across varied media, and mitigates social and demographic biases. An overview of the approach, along with research questions and their respective hypotheses, is provided in Figure 2.

3 Deepfake Detection: Current Research

Leveraging heart rate features for deepfake detection provides enhanced interpretability, accuracy, and valuable insights, especially in high resolution videos. This approach effectively identifies sophisticated deepfakes through continuous heart rate monitoring. However, challenges include the impact of real world skin conditions, scene variations, and head movements on reliable heart rate extraction with remote photoplethysmography (rPPG) methods. Ethical and privacy concerns also arise with camera based monitoring, necessitating algorithms that modify physiological signals to protect subjects' privacy [2].

Beyond heart rate (HR), other physiological cues are also investigated for deepfake detection. Electroencephalography (EEG) captures brain responses to visual stimuli, though large scale dataset collection and privacy issues present challenges [10]. Oculomotor cues, like blink rates and pupil shape, offer a promising direction due to the difficulty of replicating involuntary eye movements in deepfakes [7]. Oral physiology analyzes mouth movements to differentiate real from fake videos, although datasets are limited [3]. Finally, facial biomechanics focus on subtle, authentic facial expressions, providing robust resistance to advanced deepfake techniques [6]. Figure 3 visualizes these physiological signals and biomechanics.

Given these advancements and challenges in leveraging physiological signals, especially heart rate, for deepfake detection, this study introduces a comprehensive model that addresses these limitations and enhances detection robustness through an innovative combination of TF analysis and ML techniques.

4 Our Solution Strategy and Research Contribution

We propose an innovative model using Time Frequency (TF) algorithms and machine learning (ML) for heart rate extraction and deepfake detection. The

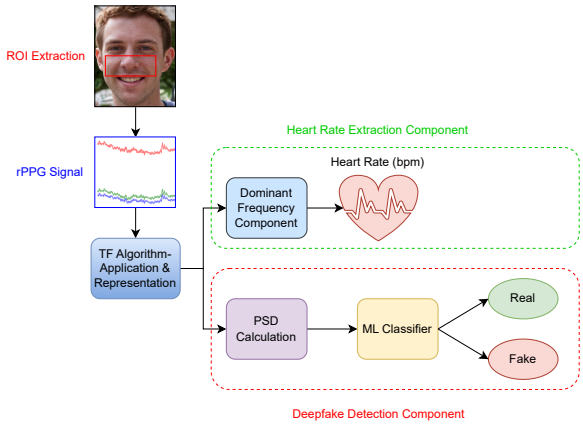


Fig. 4. Proposed model for heart rate extraction and deepfake detection

framework consists of three phases: (1) extracting a physiological signal from facial regions of interest (ROIs) and performing initial preprocessing, (2) analyzing this signal using TF algorithms to derive heart rate metrics from its frequency components, and (3) employing statistical measures obtained from the power spectral density (PSD) of the TF analysis to facilitate deepfake detection through ML classification methods.

Signal Extraction & Preprocessing: The process begins by isolating specific ROIs from video frames and extracting relevant green signal information, followed by techniques for noise reduction and trend removal to improve data quality for further analysis.

HR Calculation via TF Algorithm: Next, TF analysis tools are utilized to generate a representation of the processed signal, enabling heart rate calculation based on the peak frequency identified through this representation.

ML Based Deepfake Detection: For accurate deepfake detection, some enhancements are applied to both original and altered videos before re-extracting the TF model data from each frame. This data is then represented as PSD, from which key metrics are derived and used as inputs for ML algorithms to classify videos as authentic or fake. Figure 4 illustrates the entire model described in this section.

In this research, we emphasize our published survey on deepfake detection [1], which focuses on integrating physiological signals and facial biomechanics. This survey provides a foundation on current advancements and challenges, enriching discussions on effective detection methods.

While our proposed heart rate based methodology offers a structured approach to deepfake detection, rapid advancements in deepfake technology reveal promising areas for future research.

5 Plan of Action

The following steps outline the key phases of our research:

- Published a comprehensive survey on the use of physiological signals and facial biomechanics in deepfake detection.
- To develop a novel technique for calculating HR from rPPG signals using TF domain analysis.
- To design an unique and robust method for detecting visual deepfakes based on rPPG signal characteristics, leveraging TF algorithms for enhanced detection accuracy.
- To investigate the correlation between rPPG signals and changes in affective states, contributing to potential new insights in the field of affective computing.

References

1. Chakraborty, R., Naskar, R.: Role of human physiology and facial biomechanics towards building robust deepfake detectors: A comprehensive survey and analysis. Computer Science Review **54**, 100677 (2024)
2. Chen, M., Liao, X., Wu, M.: Pulseedit: Editing physiological signals in facial videos for privacy protection. IEEE Transactions on Information Forensics and Security **17**, 457–471 (2022)
3. Elhassan, A., Al-Fawa'reh, M., Jafar, M.T., Ababneh, M., Jafar, S.T.: Dft-mf: Enhanced deepfake detection using mouth movement and transfer learning. SoftwareX **19**, 101115 (2022). https://doi.org/https://doi.org/10.1016/j.softx.2022.101115, https://www.sciencedirect.com/science/article/pii/S2352711022000759
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
5. Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., Liu, Y.: Countering malicious deepfakes: Survey, battleground, and horizon. International journal of computer vision **130**(7), 1678–1734 (2022)
6. Liao, X., Wang, Y., Wang, T., Hu, J., Wu, X.: Famm: facial muscle motions for detecting compressed deepfake videos over social networks. IEEE Transactions on Circuits and Systems for Video Technology **33**(12), 7236–7251 (2023)
7. Nguyen, H.M., Derakhshani, R.: Eyebrow recognition for identifying deepfake videos. In: 2020 international conference of the biometrics special interest group (BIOSIG). pp. 1–5. IEEE (2020)
8. Ning, X., Xu, S., Nan, F., Zeng, Q., Wang, C., Cai, W., Li, W., Jiang, Y.: Face editing based on facial recognition features. IEEE Transactions on Cognitive and Developmental Systems **15**(2), 774–783 (2022)
9. Nirkin, Y., Masi, I., Tuan, A.T., Hassner, T., Medioni, G.: On face segmentation, face swapping, and face perception. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 98–105. IEEE (2018)
10. Tarchi, P., Lanini, M.C., Frassineti, L., Lanatà, A.: Real and deepfake face recognition: An eeg study on cognitive and emotive implications. Brain Sciences **13**(9), 1233 (2023)

Audit & Mitigation of Gender Biases in Human-AI Platforms

Siddharth D Jaiswal

Indian Institute of Technology, Kharagpur, India
siddsjaiswal@kgpian.iitkgp.ac.in

1 PhD Context

I am 4th year PhD student under Prof. Animesh Mukherjee at the Dept. of CSE, IIT Kharagpur. I started my PhD in January 2021, and my expected timeline for completion is by June 2025. I am studying biases in Human-AI platforms developed in the industry and academia from a rigorous academic lens.

2 Research Direction

Human-AI platforms are AI-based systems that humans directly interface with, for addressing some personal, social or societal need. Examples range from media recommendation platforms (Netflix) and e-commerce websites (Amazon) to face (AWS Rekognition) and speech (OpenAI Whisper) recognition platforms. These platforms have now become ubiquitous. While the benefits of using such systems are well known, a wide variety of biases have also been reported against different stakeholders. For example- there have been reports of discrimination against dark-skinned people by Face Recognition Systems (FRSs) [1], biases against non-binary shoppers [2] and users of social media platforms [3]. These biases can and have had far-reaching consequences on society. It is essential to identify and mitigate these biases to ensure fair treatment for all individuals involved. To identify these biases, researchers perform (third-party) audit studies that probe the platforms with various inputs (standard & adversarial) and analyze the outputs along different demographic dimensions. To mitigate the biases, interventions are introduced in one or more of the following stages of model development– pre-processing (changes to the training dataset), in-processing (changes to the objective function) or post-processing (regularization/smoothing) steps. Such studies form the bedrock of Responsible AI.

Problem Statement: As part of my PhD studies, I am focusing on the broad problem of *gender related biases in Human-AI systems*, making contributions to every part of the Responsible AI pipeline– (i) Adversarial *Audits* in different human-AI platforms to identify biases against binary & non-binary gender groups, (ii) Novel gender-inclusive *datasets* and (iii) Low-resource *bias mitigation* algorithms. The Human-AI systems that I am studying as part of my PhD thesis are– (a) Face Recognition Systems (FRSs), (b) e-commerce platforms, (c) text-based gender analyzers and (d) Vision Language Models (VLMs). My thesis is

divided into three parts– (i) Adversarial audit of FRSs [2, 5, 4] for binary genders, (ii) Data-centric [4] bias mitigation in FRSs for binary genders using real and synthetic datasets, and (iii) Audit of non-binary gender bias in other Human-AI systems– visual-search enabled e-commerce [2], text-based gender analyzers [3] and VLMs (current work).

Existing studies have primarily focused on standard audits of these platforms, using benchmark datasets, which often lack real-world representations in terms of gender, race, socio-political demographics or other realistic features. To audit FRSs, I have applied simple but adversarial filters like RGB, Blur, and face masks to create realistic variants of these images that have been used to audit both commercial [1] and open-source FRSs [5]. This has led to two interesting observations– (a) FRSs are not robust to such adversarial inputs, and (b) Existing biases against dark-skinned individuals are exacerbated under adversarial scenarios. In [5], I have also observed that humans are similarly biased and hence vanilla humans-in-the-loop solutions are not a good enough bias mitigation strategy. In [4], where we study data-centric bias mitigation, we contribute a new face dataset with a majority contribution from the Global South countries (with realistic adversarial variants), which is then used to mitigate biases in FRSs. In our audit of visual search enabled e-commerce platforms, I studied biases against non-binary shoppers [2] and observed that the platforms focus more on the perceived gender (from the face) rather than the item of clothing the person is searching for. Similarly, our audit on text-based gender analyzers [3] indicated that neither the task-specific models nor LLMs are designed to identify non-binary authors and often classify them as female, perpetuating the stereotype that non-binary individuals are effeminate. In many of our studies, I have performed extensive human surveys to support our findings.

3 Research Questions and Contributions

I am addressing the following research questions as part of my PhD thesis–

RQ1. Adversarial audits for FRSs: Audit studies are performed to evaluate AI systems for their performance in real-world deployment scenarios. Multiple audit studies in literature [6] have exposed deeply ingrained biases in many AI platforms like Face Recognition Systems, LLMs and Speech-to-Text systems, especially against historically disadvantaged or marginalized members of society. These audits are performed using benchmark datasets, most of which are either sourced from the internet or from volunteers and for either face matching or facial attribute analysis.

Gaps identified: Existing research has significant gaps in two major areas– (a) Real-world inputs are often poisoned by environmental factors like natural rain, dust etc, or societal factors like face masks and (b) Existing audits are mostly unidimensional and do not attempt to understand the impact of realistic adversarial inputs on biases.

Research Contributions: I perform a new, more realistic audit, namely "adversarial" audit. The inputs are adversarial in that they simulate the presence of

realistic environmental and societal factors like rain and face masks, etc and thus are *not* out-of-distribution. This addresses Gap (1). Next, the adversarial inputs also allow for the study of model robustness towards bias. I study this bias against different demographic intersectional groups [1]. Currently, I have completed two audit studies– (a) [1]– We audit three commercial FRSs for various facial attribute analysis tasks and discover that biases against Black females increase significantly. (b) [5]– We audit thirteen FRSs (commercial and open-source) for face matching with masked faces and discover large-scale biases against Black females; a survey with over 80 participants returns similar results indicating that such biases equally perpetuate amongst humans.

RQ2. Developing data-centric bias mitigation solutions: Bias mitigation in AI platforms can be performed at one of the three intervention stages– pre-processing, in-processing or post-processing. In my PhD, I am focusing on the pre- and in-processing stages using data-centric bias mitigation solutions. All AI-based platforms need well-sampled and representative data to function in an unbiased manner. Thus, the development and availability of diverse, fair datasets are of utmost importance not just for model development but also for audits. Due to privacy concerns and lack of representation from the Global South, synthetic datasets are urgently needed to augment existing face image datasets.

Gaps identified: I have identified the following major gaps in existing literature– (1) Existing datasets are highly imbalanced in terms of racial/ethnic and gender representations and the models trained on these datasets carry forward the biases, (2) A majority of existing datasets violate the privacy of regular citizens as their photos are collected from their social media profiles without their permission, (3) There is a lack of sufficient datasets from the Global South countries. As these countries are the hotbed of deployment for AI software, it is important to ensure that their citizens are fairly represented in the data, and (4) There are very few adversarial datasets for training and auditing FRS models.

Research Contributions: We have developed a benchmark dataset [4] that has representation from various ethnicities and races (addresses Gap (1)) and is composed of publicly famous individuals (addresses Gap (2)). Our dataset has more than 50% individuals belonging to the Global South countries (addresses Gap (3)) and has five adversarial variants (addresses Gap (4)). This large-scale dataset (more than 40k images, including the adversarial variants) will allow us to train, test and audit both commercial and open-source FRSs under standard and adversarial conditions. In my current work, I am developing synthetic face datasets using conditioned and unconditioned diffusion models that are representative of images from the Global South.

RQ3. Auditing human-AI systems for biases against non-binary individuals: Human-AI platforms like visual search enabled e-commerce, text-based gender analyzers and VLMs are used for various personal and professional purposes, directly interfacing with humans. Developers must design them with due acknowledgement to the minority groups who may use these platforms. Non-

[1] Any group formed by combining race and gender labels– white male, black female, etc.

binary individuals are one such group who often face discrimination in society, and if this discrimination is baked into the AI models deployed at scale around the world, these individuals may face unprecedented biases. Thus it is important to audit and fix existing platforms for such biases using sufficiently diverse datasets.

Gaps identified: We have identified the following major gaps in existing literature– (1) Existing audit studies rarely look at biases against non-binary individuals, even in domains where such datasets may be available and, (2) Most existing models are not designed with gender fluidity in mind, either due to developer oversight or deliberate choice.

Research Contributions: I have curated two novel datasets where more than 50% of the data points belong to the non-binary gender group– an image dataset of fashion wear for male, female and non-binary clothing [2], and a text-dataset of Reddit and Tumblr comments/posts from self-declared non-binary individuals [3] that can be used to audit visual search enabled e-commerce and text-based gender analyzers (addresses Gap (1)). I am currently working on developing a similar multi-modal dataset for auditing VLMs. I have also developed a gender-inclusive text-based gender analyzer that predicts for the non-binary gender group [3] along with males and females.

RQ2b (developing synthetic face datasets) and **RQ3c** (auditing VLMs) are the current focus of my work, and I plan to make further developments to both these topics.

3.1 List of Publications

1. **Jaiswal, S.**, Ganai, A., Dash, A., Ghosh, S., & Mukherjee, A. (2024, October). Breaking the Global North Stereotype: A Global South-centric Benchmark Dataset for Auditing and Mitigating Biases in Facial Recognition Systems. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Vol. 7, pp. 634-646).
2. **Jaiswal, S.**, Verma, A. K., & Mukherjee, A. (2023, November). Auditing gender analyzers on text data. In IEEE/ACM ASONAM (pp. 108-115).
3. **Jaiswal, S.**, Duggirala, K., Dash, A., & Mukherjee, A. (2022, May). Twoface: Adversarial audit of commercial face recognition systems. In AAAI ICWSM (Vol. 16, pp. 381-392).
4. **Jaiswal, S.**, & Mukherjee, A. (2022, April). Marching with the pink parade: Evaluating visual search recommendations for non-binary clothing items. In ACM CHI EA (pp. 1-8).
5. Rai, A. K., **Jaiswal, S. D.**, Prakash, S., Sree, B.P. & Mukherjee, A. (2024). DENOASR: Debiasing ASRs through Selective Denoising. In IEEE ICKG. *To Appear.*
6. Rai, A. K., **Jaiswal, S. D.**, & Mukherjee, A. (2024, May). A Deep Dive into the Disparity of Word Error Rates across Thousands of NPTEL MOOC Videos. In AAAI ICWSM (Vol. 18, pp. 1302-1314).
7. **Jaiswal, S.**, & Mukherjee, A. (2023, April). A History of Diversity in The Web (Conference). ACM TheWebConference (Companion) (pp. 625-632).

8. Das, M., Dash, A., **Jaiswal, S.**, Mathew, B., Saha, P., & Mukerjee, A. (2022). Platform governance: Past, present and future. ACM GetMobile: Mobile Computing and Communications, 26(1), 14-20.

4 Future Directions

In the immediate future, before submitting my thesis, I plan to complete the following tasks.
(A) I have developed an adversarial audit strategy that has been used for the tasks of face matching and facial attribute analysis. Next, I plan to make this audit strategy more explainable, especially for non-domain experts that will allow users to understand the model's preference towards the different facial regions while performing the face matching or attribute analysis task. This will help in addressing **RQ1** and allow researchers to design better bias mitigation algorithms.
(B) I have developed an initial data-centric bias mitigation solution using a new face dataset [4] which has the majority of images from the Global South. In the current and future work, I plan to augment this dataset using synthetically generated face images that are representative of the Global South demographic. I am using unconditioned and text-to-image diffusion models for this purpose, fine-tuned on more representative Global South face images. This will address the challenges highlighted in **RQ2**.

References

1. Jaiswal, S., Duggirala, K., Dash, A., Mukherjee, A.: Two-face: Adversarial audit of commercial face recognition systems. In: AAAI ICWSM (2022)
2. Jaiswal, S., Mukherjee, A.: Marching with the pink parade: Evaluating visual search recommendations for non-binary clothing items. In: ACM CHI Extended Abstracts (2022)
3. Jaiswal, S., Verma, A.K., Mukherjee, A.: Auditing gender analyzers on text data. In: IEEE/ACM ASONAM (2023)
4. Jaiswal, S.D., Ganai, A., Dash, A., Ghosh, S., Mukherjee, A.: Breaking the global north stereotype: A global south-centric benchmark dataset for auditing and mitigating biases in facial recognition systems. In: AAAI/ACM AIES (2024), to appear
5. Jaiswal, S.D., Verma, A.K., Mukherjee, A.: Mask-up: Investigating biases in face re-identification for masked faces. arXiv preprint arXiv:2402.13771 (2024)
6. Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C.: Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry **22**(2014), 4349–4357 (2014)

Development of a forest mapping system using UAVs and AI models

Francisco Raverta Capua[1,2][0009−0004−3337−7741]

[1] Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Computación. Buenos Aires, Argentina.
[2] CONICET-Universidad de Buenos Aires. Instituto de Ciencias de la Computación (ICC). Buenos Aires, Argentina.
fraverta@icc.fcen.uba.ar

1 PhD context

I am currently pursuing a PhD in Computer Science under the supervision of PhD. Pablo De Cristóforis. My doctoral studies began on April 1, 2021, and are expected to conclude on May 31, 2026. This research is conducted in a fully academic context, supported by CONICET (National Council for Scientific and Technical Research, Consejo Nacional de Investigaciones Científicas y Técnicas) through a call for doctoral proposals on strategic topics. Among the strategic topics available, this work is framed in the use of robotics and AI tools aimed at environmental development.

2 Main Topic

My PhD research focuses on developing an innovative system tailored for Argentinian institutions related to forest management and conservation. This system will provide a wide range of functionalities, including a tridimensional computational model of the study area, the estimation of structural parameters (such as crown height, density of individuals, cover percentage and plant stratification), detection of unauthorized logging, automatic classification of species and calculation of vegetation indices.

The country has faced an intense deforestation process over the last decades, and the remaining forests are mostly degraded. My PhD aims to develop a monitoring system mature enough for the agencies in charge of forest management to adopt and apply it directly on several national parks along the country. This project is being developed in collaboration with groups of ecologists and different national parks authorities with whom we maintain collaboration.

3 State of the Art

Forests play a vital role in the health of the planet. They are fundamental for biodiversity, climate regulation and climate change mitigation, oxygen production, water cycle, soil conservation, the economy and human well-being. Their

preservation and management are essential to ensure a sustainable future for our planet [1]. Consequently, the monitoring and management of forest resources have strategic importance for countries, not only from a conservation standpoint but also for economic-productive development.

Recently, there has been a great interest in using unmanned aerial vehicles (UAVs) as remote sensors for ecology, and for forest monitoring in particular [2]. UAVs offer several advantages over satellite imagery: arbitrary revisit time, high spatial resolution, independence from cloud cover, low cost and operational risk. Field work, on the other hand, is useful for obtaining high-detailed information, but it is less efficient in terms economic cost-analyzed area. This is reflected in the proliferation of organizations at an international level that use UAVs for environmental monitoring. The use of remote sensing in this field has grown significantly in recent years, thanks to the development of Terrestrial Laser Scanning (TLS), Aerial Laser Scanning (ALS), and Aerial Photogrammetry, techniques widely used in precision forestry [3]. Both LiDAR and camera sensors made it possible to easily acquire three-dimensional data of the studied environment, accurately representing it with high precision level point clouds. Both have been widely used in forest environments for health monitoring, species classification, tree parameter estimation, and even illegal logging detection amidst other applications [4]. In Argentina, although there are some ongoing projects to develop and/or apply this technologies, their practice is still in early stages, making it essential to invest in its growth.

Nowadays, several software tools are available for processing aerial images captured with UAVs, bot commercial (e.g. Agisoft Metashape) and open-source (e.g. OpenDroneMap). These systems use a computer vision technique called Structure from Motion (SfM) to generate the 3D reconstruction of the survayed area. This 3D models, as well as those obtained with LiDARs, are used to generate the digital terrain model (DTM) of the forest. Finding an accurate DTM is helpful with the posterior estimation of most of the structural parameters of the forest, as it allows to level the forest floor before estimating volumes, heights and widths of the trees, and it makes it easier to detect individual canopies. Although there are several geometrical ways to obtain an initial DTM, we proposed using AI models, particularly based on transformer technology, to improve it and to address the problem of tree species classification simultaneously via segmentation of the point clouds. Similar works can be seen in [5–8].

Very few point cloud datasets of forest environments are publicly available. Therefore, if training a deep learning architecture for forested environment is needed, a new dataset for this purpose has to be generated. For example, [5] developed a dataset from the regions of the Southern Sierra Nevada Mountains, USA, [6] from Australia and New Zealand, and [7] from Evo, Finland. All these works were conducted for forest segmentation. Of the three, only the latter dataset is publicly available, limiting the repeatability of the experiments and the comparison between the cited works.

Regarding species detection from forest imagery, recent advances have closely followed developments in the field of artificial intelligence. For instance, stud-

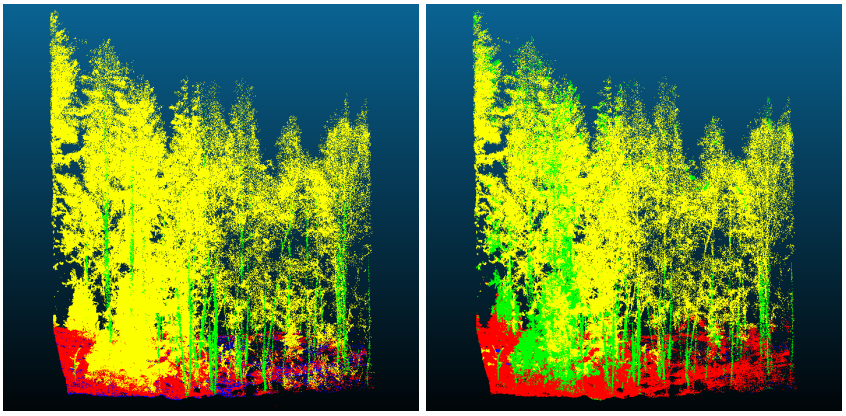
ies [9–11] have utilized multispectral imagery, while [12–14] have relied solely on RGB imagery.

4 Methodology and Contributions

As part of this project, an UAV capable of carrying out survey missions autonomously and carrying out a detailed photographic survey is being built. This vehicle is based on a Skywalker fixed-wing fuselage with 1900mm wingspan, equipped with a Pixhawk autopilot running ArduPilot navigation software and with several sensors, including an inertial unit, barometer, digital compass, wind sensor, GPS, radiocontrol and telemetry link. A 24-megapixel Sony A6100 controlled by the autopilot is being installed, together with a RTK module for the GPS, which will decrease geo-reference errors from meters to centimeters.

I am currently working towards the detection of digital terrain models (DTM) on 3D forest models, generated either by LiDAR or by applying Structure from Motion techniques to photos taken from the environment, via point segmentation with deep learning models. In the paper accepted for the ICPR2024, we have selected four networks of the state-of-the-art for segmenting point clouds, named PointNeXt, PointBERT, PointMAP, and PointGPT, and trained them with the synthetic points obtained via the developed forest simulator to differentiate between four categories: ground, trunk, canopy and understorey. We then test this models in real forest data, and evaluate whether training with synthetic data was good enough for this segmentation task in real forests. A sample result can be seen in Fig. 1. We additionally evaluate the accuracy gain after doing a fine-tuning over those networks with a very few portion of the real-forest dataset, concluding that it is enough to retrieve only a small data quantity for the forest to analyze to obtain good results. I am also analyzing the uncertainty associated with the label of each of the points of the point cloud, for all the networks used. The objective is to identify which are the zones of the point cloud that has labeling errors with high and low uncertainty, and whether it could be resolved by doing changes on the developed simulator or by using more information of those zones in training time in order to obtain a higher accuracy when segmenting and labeling real forest data.

The developed simulator also offers the feature of simulating a flight survey over the forest scene, capturing RGB images of the view from above, as can be seen in Fig. 2. This images allows to use SfM algorithms to obtain a 3D representation of the scene that relates to the ones obtained with a similar method in real forests. It also has the possibility to extract the ground-truth segmentation of the images, indicating pixel-by-pixel to which category they belong to. Nowadays, most SfM algorithms does not have the feature of using semantic information to reduce errors. We aim to use this information to generate a more 3D representation of the scenes, and then translate this to real-world forest representations.



(a) Ground Truth (b) PointMAE

Fig. 1: Fraction of the Evo Dataset classified with PointMAE, one of the selected networks: terrain (blue), trunks (green), canopy (yellow) and understorey (red).

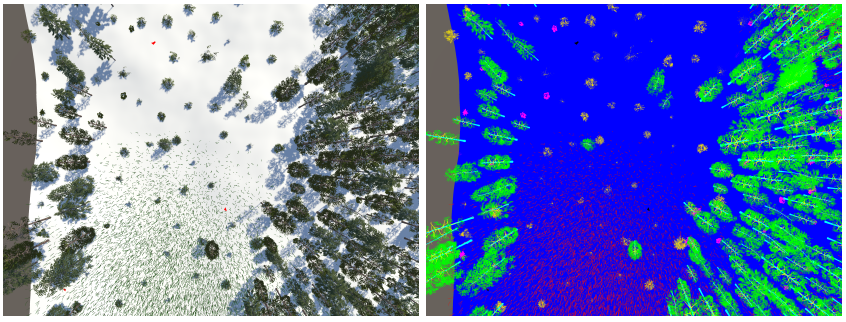


Fig. 2: Generated forest scene. Left: RGB view. Right: Segmented view.

5 Planned Actions Before PhD Completion

Based on the work already done, and pursuing the proposed objectives of my doctoral thesis, I have the following planned actions:

1. Extend the work presented in the ICPR2024 with other real forest datasets, and repeat the experiments with simulated flight surveys, using SfM to extract point clouds from images and ensure a more realistic data acquisition. Analyze the uncertainty of the results of segmenting the point clouds.
2. Finish integrating the hardware on the UAV. Make surveys on forest areas of interest. Generate our own real forest datasets to use them to validate our research approach.
3. Use deep learning techniques over images obtained with the UAV for tree species identification in forest areas of interest, which have multiple and varied species cohabiting.

References

1. IPBES. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES secretariat, Bonn, Germany. 56 pp. (2019)
2. Krause, S., Sanders, T. G., Mund, J. P., Greve, K. UAV-based photogrammetric tree height measurement for intensive forest monitoring. Remote sensing, 11(7), 758 (2019).
3. Murtiyoso, A., Holm, S., Riihimäki, H., Krucher, A., Griess, H., Griess, V. C., Schweier, J: Virtual forests: a review on emerging questions in the use and application of 3D data in forestry. International Journal of Forest Engineering **35**(1), 29–42 (2024)
4. Guimarães, N., Pádua, L., Marques, P., Silva, N., Peres, E., Sousa, J. J.: Forestry remote sensing from unmanned aerial vehicles: A review focusing on the data, processing and potentialities. Remote Sensing **12**(6), 1046 (2020)
5. Jin, S., Su, Y., Zhao, X., Hu, T., Guo, Q.: A point-based fully convolutional neural network for airborne lidar ground point filtering in forested environments. IEEE journal of selected topics in applied earth observations and remote sensing **13**, 3958–3974 (2020)
6. Krisanski, S., Taskhiri, M. S., Gonzalez Aracil, S., Herries, D., Turner, P.: Sensor agnostic semantic segmentation of structurally diverse and complex forest point clouds using deep learning. Remote Sensing **13**(8), 1413 (2021)
7. Kaijaluoto, R., Kukko, A., El Issaoui, A., Hyyppä, J., Kaartinen, H.: Semantic segmentation of point cloud data using raw laser scanner measurements and deep neural networks. ISPRS Open Journal of Photogrammetry and Remote Sensing **3**, 100011 (2022)
8. Li, B., Lu, H., Wang, H., Qi, J., Yang, G., Pang, Y., Dong, H., Lian, Y.: TerrainNet: A Highly-Efficient, Parameter-Free, and Easy-to-Use Deep Neural Network for Ground Filtering of UAV LiDAR Data in Forested Environments. Remote Sensing **14**(22), 5798 (2022)
9. Nezami, S., Khoramshahi, E., Nevalainen, O., Pölönen, I., Honkavaara, E. Tree species classification of drone hyperspectral and RGB imagery with deep learning convolutional neural networks. Remote Sensing, 12(7), 1070 (2020).
10. Miyoshi, G. T., Arruda, M. D. S., Osco, L. P., Marcato Junior, J., Gonçalves, D. N., Imai, N. N., ... Gonçalves, W. N. . A novel deep learning method to identify single tree species in UAV-based hyperspectral images. Remote Sensing, 12(8), 1294 (2020).
11. Abdollahnejad, A., Panagiotidis, D. Tree species classification and health status assessment for a mixed broadleaf-conifer forest with UAS multispectral imaging. Remote Sensing, 12(22), 3722 (2020).
12. Ferreira, M. P., de Almeida, D. R. A., de Almeida Papa, D., Minervino, J. B. S., Veras, H. F. P., Formighieri, A., Ferreira, E. J. L. Individual tree detection and species classification of Amazonian palms using UAV images and deep learning. Forest Ecology and Management, 475, 118397 (2020).
13. Beloiu, M., Heinzmann, L., Rehush, N., Gessler, A., Griess, V. C. Individual tree-crown detection and species identification in heterogeneous forests using aerial RGB imagery and deep learning. Remote Sensing, 15(5), 1463 (2023).
14. Zhang, C., Zhou, J., Wang, H., Tan, T., Cui, M., Huang, Z., Zhang, L. Multi-species individual tree segmentation and identification based on improved mask R-CNN and UAV imagery in mixed forests. Remote Sensing, 14(4), 874 (2022).

Audio Forensics: Leveraging Deep Learning and Adaptive Learning for Audio Deepfake Detection

Taiba Majid Wani[1][0000−0002−2335−0420]

Sapienza University of Rome, Italy
https://www.uniroma1.it/
@diag.uniroma1.it

1 PhD Context

My PhD journey began on 1st November 2022 under the supervision of Prof. Irene Amerini at the Department of Computer, Control, and Management Engineering, Sapienza University of Rome. The PhD is primarily conducted within an academic context and focuses on the field of Audio Forensics, with a specialized interest in *deepfake detection* using advanced methodologies. The expected completion date for my PhD is October 2025. Through this research, I aim to contribute significantly to the challenges posed by synthetic and manipulated audio, which has far-reaching implications for security and trust in digital media

2 Main Topic and Problem Statement

Audio deepfakes are synthetic audio clips generated using artificial intelligence techniques like Text-to-Speech (TTS) synthesis and Voice Conversion (VC). These methods allow for the creation of highly realistic imitations of human voices, making it sound as though a person is saying something they never actually said. Audio deepfakes have been used for various purposes, including entertainment, virtual assistants, and content creation. However, they also pose serious risks when exploited for malicious activities. The growing complexity of audio deepfakes presents a significant threat to digital security and trust. Deepfakes are often used in identity theft, fraud, disinformation campaigns, and blackmail, where an individual's voice can be convincingly manipulated to deceive listeners. The seamless nature of these forgeries makes it difficult for traditional detection systems to differentiate between real and fake audio, leading to widespread potential misuse.

Given the increasing threat of audio deepfakes, the relative lack of research compared to video deepfakes, and the scarcity of comprehensive datasets, there is an urgent need for advanced and adaptive detection systems. Existing methods struggle to keep pace with the evolving nature of audio deepfakes, leaving a void in audio forensics. Developing a robust detection framework is crucial for accurately distinguishing between real and fake audio, maintaining trust in digital communications, and preventing the misuse of audio deepfakes in key sectors such as media, politics, and security.

3 State-of-Art

Several works have been proposed in the field of audio deepfake detection, utilizing deep learning techniques and datasets such as ASVspoof and FoR. Wang et al. [1] introduced DeepSonar, using neuron behavior analysis to detect fake and real speeches with an accuracy of 98.1% on the FoR dataset. Camacho et al.[2] developed a two-stage model using CNNs for raw data transformation, achieving 88% accuracy on the FoR-original dataset. Luo et al. [3] proposed a Capsule Network architecture with a modified dynamic routing algorithm, improving generalization and achieving an EER of 3.19% on ASVspoof. Ma et al. [4] designed a ConvNeXt-based model incorporating Res2Net and efficient channel attention, reaching an EER of 0.64% on ASVspoof. Hamza et al. [5] explored various machine learning models, with SVM achieving 98.83% accuracy on FoR-rerec and VGG16 reaching 93% on FoR-original. Mittal et al. [6] combined static and dynamic CQCC features, obtaining an EER of 0.009 on ASVspoof. Continual learning approaches, such as Detecting Fake Without Forgetting (DFWF) by Ma et al. [7] and Radian Weight Modification (RWM) by Zhang et al., [8] have been introduced to address the challenge of catastrophic forgetting and improve adaptability in detecting new deepfake attacks. From the literature survey, we drew inspiration from various approaches and incorporated key elements into our research. Building on these insights, we developed several novel methodologies, which are detailed in the following section.

4 Methodology and Contributions

Convolutional Neural Networks (CNNs) have been extensively used for feature extraction from Mel spectrograms in audio deepfake detection. We proposed a novel custom convolutional neural network (cCNN) to detect audio deepfakes using mel spectrograms [9]. The cCNN architecture consists of four convolutional layers and two fully connected layers, with smaller filter sizes and dropout layers to prevent overfitting. The model was trained on the FoR dataset, which was divided into sub-datasets (for-norm, for-2-s, and for-rerecording). Mel spectrograms, a frequency-based representation of audio, were used as input to the network to capture essential frequency patterns. Additionally, two pretrained models, VGG16 and MobileNet, were employed for comparison. Data augmentation was performed to improve generalization.

However, CNNs face limitations in handling complex transformations, often leading to misclassifications in challenging scenarios. To overcome the limitations of CNNs, we introduced Capsule Networks (CapsNet) into the detection pipeline. In [10] a novel ABC-CapsNet architecture was introduced to improve audio deepfake detection. The proposed model employs VGG18 for feature extraction, followed by an attention mechanism that enhances critical feature regions. Two Capsule Networks (CapsNet) are arranged in a cascaded architecture to preserve spatial hierarchies and improve generalization. Capsule Network 1 extracts complex audio features, while Capsule Network 2 refines them, producing final activity vectors for classification.

Recognizing the need for enhanced temporal analysis, we further explored feature concatenation techniques and hybrid models. In [11], we proposed a hybrid architecture that integrates Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) to detect audio deepfakes. The novelty lies in the concatenation of four distinct acoustic features: Mel Frequency Cepstral Coefficients (MFCC), Mel spectrograms, Constant-Q Cepstral Coefficients (CQCC), and Constant-Q Transform (CQT). The CNN extracts spatial features from the input, and these features are concatenated into two multi-dimensional feature sets. The BiLSTM network processes these feature sets to capture temporal dynamics and contextual dependencies in the audio data.

Additionally, we addressed the evolving nature of audio deepfake attacks through continual learning framework. In [12], we introduced a continual learning approach for audio deepfake detection, addressing catastrophic forgetting and incremental learning. The methodology integrates the feature extraction power of SincNet and the computational efficiency of LightCNN. Knowledge is transferred from SincNet to LightCNN using Feature Distillation, while Dynamic Class Rebalancing (DCR) adjusts the learning strategy based on the similarity of class features across tasks. This model was evaluated on the ASVspoof 2015, ASVspoof 2019, and FoR datasets, showing significant improvements in detecting deepfakes while maintaining performance on older tasks.

Through this work, we present a robust progression from CNN-based methods to capsule networks, hybrid architectures, and continual learning techniques, offering a comprehensive and scalable solution to the growing challenge of audio deepfake detection. Our multi-faceted approach not only improves detection accuracy across a variety of deepfake techniques but also enhances the model's ability to generalise by continuously learning and adapting to evolving threats.

4.1 Publications

- Wani, T. M., Amerini, I. (2023, September). Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks. In International Conference on Image Analysis and Processing (pp. 156-167). Cham: Springer Nature Switzerland.
- Wani, T. M., Gulzar, R., Amerini, I. (2024). ABC-CapsNet: Attention based Cascaded Capsule Network for Audio Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2464-2472).
- Wani, Taiba Majid, Syed Asif Ahmad Qadri, Danilo Comminiello, and Irene Amerini. "Detecting Audio Deepfakes: Integrating CNN and BiLSTM with Multi-Feature Concatenation." In Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security, pp. 271-276. 2024.
- Taiba M and Irene A. Icpr 2024: Audio Deepfake Detection: A Continual Approach with Feature Distillation and Dynamic Class Rebalancing. In 2024 27th International Conference on Pattern Recognition (ICPR) 2024 Dec 02. IEEE.

– Taiba M. and Irene A. 2024. Comprehensive Framework for Audio Deepfake Detection: From Capsule Networks to Hybrid and Continual Learning Models. Accepted in Women in Machine Learning workshop in NeurIPS 2024, Canada.

Under Review

– Wani, T. M., Amerini, I. GCAD: Gender-Conditioned Audio Deepfake Detection with Latent Diffusion Models. submitted to ICCASP 2025, India.
– Wani, T. M., Ahmad. F, Amerini, I. STC-CapsNet: Detecting Audio Deepfakes with Spatio-Temporal Convolutions and Capsule Networks. submitted to IEEE SCCI. 2025, Norway.
– Wani, T. M.,Madleen U, Amerini, I.HCN-TA: Hierarchical Capsule Network with Temporal Attention for a Generalizable Approach to Audio Deepfake Detection. In 40th ACM/SIGAPP Symposium On Applied Computing, (2025), Sicily, Italy
– Wani, T. M., Syed,A. Farooq, A. Amerini, I.Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons. (2024) submitted to Foundations and Trends in Privacy And Security in NOW PUBLISHERS.

5 Planned Actions and Future Research Directions

Before completing my PhD, I plan to refine my methodology by expanding my research to generate novel deepfake audio datasets and explore emotion embeddings to enhance the detection of deepfakes by analyzing emotional inconsistencies in manipulated speech. This approach aims to address the shortage of comprehensive datasets and improve detection by incorporating emotional cues that could indicate falsified audio. Looking ahead, I intend to investigate hybrid models combining GANs with Capsule Networks to improve detection accuracy through advanced feature extraction. Additionally, I will develop a hybrid ensemble model with regression-based anomaly detection to capture subtle manipulations in speech that might be overlooked by conventional methods.

Post-PhD Career Plan Upon completing my PhD, I plan to pursue a postdoctoral position in multimedia forensics, focusing on both audio and video deepfake detection. My long-term objective is to advance digital media authentication through the development of robust, real-world security measures. By collaborating across disciplines, I aim to contribute innovative techniques applicable in cybersecurity, legal forensics, and media integrity, helping to safeguard against synthetic media misuse and uphold trust in digital communications.

References

1. R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu, "Deepsonar: Towards effective and robust detection of AI-synthesized fake voices," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1207–1216, 2020.

2. S. Camacho, D. M. Ballesteros, and D. Renza, "Fake speech recognition using deep learning," in *Applied Computer Sciences in Engineering: 8th Workshop on Engineering Applications, WEA 2021, Medellín, Colombia, October 6–8, 2021, Proceedings 8*, pp. 38–48, 2021, Springer.
3. A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6359–6363, 2021, IEEE.
4. Q. Ma, J. Zhong, Y. Yang, W. Liu, Y. Gao, and W. W. Y. Ng, "Convnext based neural network for audio anti-spoofing," *arXiv preprint arXiv:2209.06434*, 2022.
5. A. Hamza, A. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via MFCC features using machine learning," *IEEE Access*, vol. 10, pp. 134018–134028, 2022.
6. A. Mittal and M. Dua, "Static–dynamic features and hybrid deep learning models based spoof detection system for ASV," *Complex & Intelligent Systems*, vol. 8, no. 2, pp. 1153–1166, 2022, Springer.
7. H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, "Continual learning for fake audio detection," *arXiv preprint arXiv:2104.07286*, 2021.
8. X. Zhang, J. Yi, C. Wang, C. Y. Zhang, S. Zeng, and J. Tao, "What to remember: Self-adaptive continual learning for audio deepfake detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, pp. 19569–19577, 2024.
9. T. M. Wani, R. Gulzar, and I. Amerini, "ABC-CapsNet: Attention based Cascaded Capsule Network for Audio Deepfake Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2464–2472, 2024.
10. T. M. Wani, S. A. A. Qadri, D. Comminiello, and I. Amerini, "Detecting Audio Deepfakes: Integrating CNN and BiLSTM with Multi-Feature Concatenation," in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, pp. 271–276, 2024.
11. T. M. Wani, S. A. A. Qadri, D. Comminiello, and I. Amerini, "Detecting Audio Deepfakes: Integrating CNN and BiLSTM with Multi-Feature Concatenation," in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, pp. 271–276, 2024.
12. T. M. Wani, and I. Amerini, "Audio Deepfake Detection: A Continual Approach with Feature Distillation and Dynamic Class Rebalancing," in *International Conference on Pattern Recognition (ICPR)*, 1-5 December, 2024.

Handwritten Text Recognition and Generation for Historical Documents - A Norwegian Perspective

Aniket Gurav[1] - PhD Student

Department of Computer Science and Communication, Østfold University College, Halden, Norway
aniketag@hiof.no

Keywords: Handwritten Text Recognition · Historic Documents · Synthetic Data Generation · Diffusion Models · Zero-Shot Learning

1 PhD Context

1.1 Supervisors and Timeline

- **Primary Supervisor:** Dr. Sukalpa Chanda,
  Østfold University College, Norway

- **Co-Supervisors:** Dr. Marius Pedersen, NTNU, Norway;
  Dr. Narayanan C. Krishnan, IIT Palakkad, India

PhD Timeline: Start Date: April 2022, Expected Completion Date: April 2025.

This PhD research aims to advance Handwritten Text Recognition (HTR) for historical documents, focusing particularly on Norwegian archives, which encounter significant challenges due to limited labeled data and wide handwriting variability. The research is conducted in academic contexts under the HUGIN-MUNIN project, funded by the Norwegian Research Council. Collaborating institutions include Østfold University College, NTNU, IIT Palakkad, and the National Library of Norway (Nationalbiblioteket), along with industry partners Teklia (France), Anahit, and TidVis (Norway). This interdisciplinary team works to create scalable, adaptable HTR systems for the large-scale digitization of historical collections across libraries, archives, and museums throughout Norway.

2 Main Topic and Problem Statement

This PhD research focuses on advancing HTR through two primary strategies: enhancing recognition accuracy and generating synthetic data. The initial phase addresses the development of a zero-shot learning framework (ResPho(SC)Net [1]) to improve HTR accuracy for Norwegian historical texts, facilitating recognition across various writer styles, including unseen writers and languages like English. The second phase focuses on the use of diffusion models for synthetic data generation, targeting both word [2] and line-level [3] handwriting. This synthetic data, rich in style variations, serves to improve model robustness and generalization, mitigating the challenges posed by limited annotated data in historical document analysis in the context of Norwegian historical data figure 1.a. Together, these approaches aim to create a versatile HTR system capable of recognizing diverse and complex handwritten text styles in historical archives.

3 Brief State-of-the-Art

HTR models is essential for digitizing historical documents, allowing scanned manuscripts to become machine-readable. Traditional HTR approaches rely on supervised deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which typically require extensive labeled datasets that are scarce for historical archives. Digitizing Historical archives faces significant challenges due to the scarcity of labeled data and the high variability in handwriting styles [4,5].

Recent advancements, such as ResPho(SC)Net & Pho(SC)Net [1,6,7], address these limitations through zero-shot learning. These methods generalizes across unseen handwriting styles, which is critical for historical documents with limited labeled data.

GANs, Transformer and Diffusion Models, including HiGAN+,HWT, WordStylist, and HTLDIFF [8,9,10,4,3], support synthetic data generation. GAN-based models, such as GANwriting [9] and HiGAN+ [10], have been successful at the word level; however, they face challenges with line-level synthesis and maintaining realistic inter-word spacing.

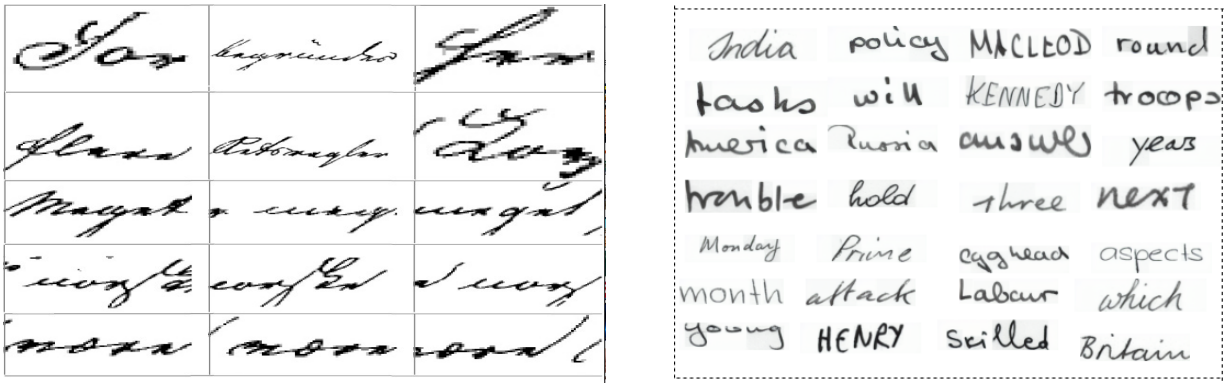
Diffusion Models [11] offer a promising alternative, allowing for better control over style transfer and variability in synthetic handwriting. Diffusion models like WordStylist [4] have shown improvements at the word level figure 1.b, though line-level and Out of Vocabulary (OOV) synthesis remains challenging. Our work HTLDIFF [3], demonstrates the ability to produce coherent, stylized text lines that maintain realistic spatial relationships, addressing limitations found in previous based models [4,9,10] also our work Word-Diffusion [2] proposes faster OOV word generation method. This research builds on these frameworks to improve both recognition accuracy and data generation quality for HTR in Norwegian historical documents.

4 Methodology and Contributions

In this section, we present our novel methodologies and contributions, focusing on improving HTR through zero-shot learning and synthetic data generation.

4.1 Recognition Enhancement through Zero-Shot Learning

This phase focuses on developing the ResPho(SC)Net [1] framework, a zero-shot learning model that uses a ResNet-18 backbone combined with PHO(SC) [6] representation. This enables the model to generalize across unseen words and handwriting styles with minimal labeled data, effectively enhancing HTR for diverse historical documents [5]. ResPho(SC)Net surpasses existing state-of-the-art methods such as PHO(SC)Net [6] in both generalized zero-shot learning (GZSL) and traditional zero-shot learning (ZSL) tasks, especially on challenging datasets like IAM and Norwegian historical archives. Compared to PHO(SC)Net [6], which has 134.44 million parameters, ResPho(SC)Net [1] achieves higher accuracy for both seen and unseen classes and is significantly lighter with only 48.12 million parameters—nearly three times fewer. This substantial reduction in model size enhances ResPho(SC)Net's [1] efficiency and generalization ability, making it well-suited for zero-shot learning tasks in handwritten text recognition without compromising performance.



(a) Norwegian Handwritten Images (b) DiffWord generated images

Fig. 1: Handwritten Data

4.2 Synthetic Data Generation with Diffusion Models

Diffusion Models are used to generate synthetic handwriting at word and line levels:

– **Word-Diffusion Model [2]:** This model generates high-quality handwritten word images, including OOV words, which are essential for expanding the diversity of training data using minimal labeled samples. To improve efficiency, the model implements an early sampling technique, reducing the typical diffusion steps from 600-1000 to just 120 for the IAM [5] dataset and 200 for CVL [12]. This approach accelerates data generation without sacrificing quality, allowing for the rapid creation of diverse and stylistically consistent handwritten word images. By incorporating both OOV capabilities and this efficient sampling strategy, the model enhances training diversity while maintaining

high fidelity to realistic handwriting styles.

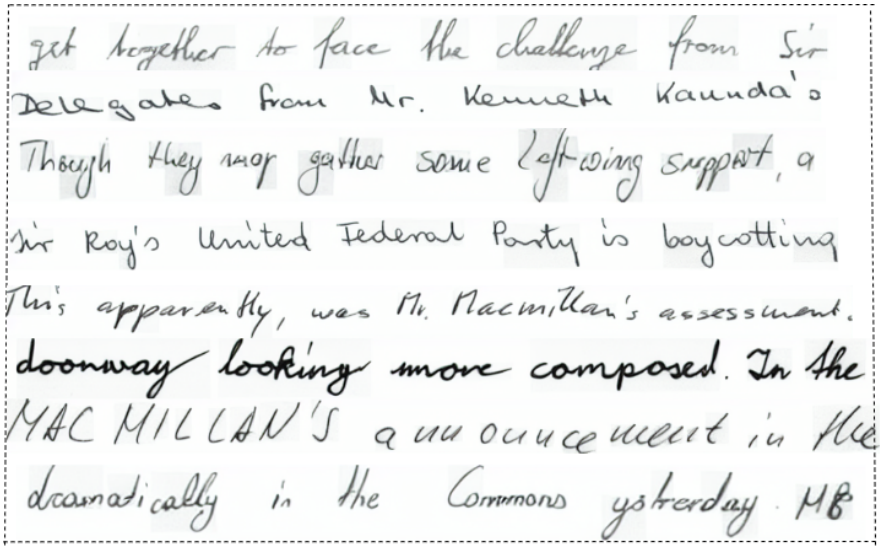


Fig. 2: HTLDIFF generated lines

– **HTLDIFF Model[3]:** This is the first method to use diffusion models for line-level generation of handwritten text. HTLDIFF [3] produces coherent, stylized text lines with realistic spacing and character placement, addressing the limitations of prior GAN-based and diffusion methods [4,8,9,10]. By leveraging modified unet, HTLDIFF [3] ensures that synthetic line data maintains spatial coherence and stylistic consistency in figure 2, making it particularly effective for historical documents where accurate line layout is essential [3].

4.3 Contributions

Our key contributions include the development of the ResPho(SC)Net [6] framework for zero-shot learning. Additionally, this research introduces the Word-Diffusion [2] and HTLDIFF [3] Models for synthetic data generation, providing controlled stylistic variations at both word and line levels to enhance HTR robustness. These combined approaches offer scalable solutions for HTR applications in historical and cultural archives.

– **ResPho(SC)Net Framework:** A zero-shot learning model supporting HTR by addressing data scarcity, generalized for unseen handwriting styles [1].

– **Word-Diffusion and HTLDIFF Models:** Synthetic data generation models enhancing HTR with controlled stylistic variations. Word-Diffusion [2] is selected for ICPR 2024, and HTLDIFF [3] is presented at IPTA 2024.

5 Planned Actions Before Completing the PhD

Future work includes:

– **Training free methods for diverse data generation** Developing methods to prevent diffusion models from generating data too similar to the training data, thereby increasing diversity.
– **Generating Out-of-Vocabulary (OOV) Data:** Enhancing HTR with realistic OOV data generation, increasing robustness to unseen words [6].
– **Architectural Modifications for Style Transfer and End-to-End Word Spotting:** Modifying U-Net components to improve style transfer and facilitate word spotting directly.
– **Using Diffusion Models as Zero-Shot Classifiers for Word-Level Identification:** Applying diffusion models for zero-shot word recognition [13].

References

1. A. Gurav, J. Jensen, N. C. Krishnan, and S. Chanda, "Respho(sc)net: A zero-shot learning framework for norwegian handwritten word image recognition," in *Pattern Recognition and Image Analysis: 11th Iberian Conference, IbPRIA 2023, Alicante, Spain, June 27–30, 2023, Proceedings*, 2023, pp. 182–196.
2. A. Gurav, N. C. Krishnan, and S. Chanda, "Word-diffusion: Diffusion-based handwritten text word image generation," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2024.
3. A. Gurav, S. Chanda, and N. C. Krishnan, "Htldiff: Handwritten text line generation - a diffusion model perspective," in *Proceedings of the International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2024.
4. K. Nikolaidou, G. Retsinas, V. Christlein, M. Seuret, G. Sfikas, E. B. Smith, H. Mokayed, and M. Liwicki, "Wordstylist: Styled verbatim handwritten text generation with latent diffusion models," in *Document Analysis and Recognition - ICDAR 2023*, G. A. Fink, R. Jain, K. Kise, and R. Zanibbi, Eds. Cham: Springer Nature Switzerland, 2023, pp. 384–401.
5. U.-V. Marti and H. Bunke, "The iam-database: an english sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, pp. 39–46, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:29622813
6. A. Rai, N. C. Krishnan, and S. Chanda, "Pho(sc)net: An approach towards zero-shot word image recognition in historical documents," in *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, ser. Lecture Notes in Computer Science, J. Lladós, D. Lopresti, and S. Uchida, Eds., vol. 12821. Springer, 2021, pp. 19–33.
7. R. K. Bhatt, A. Rai, N. C. Krishnan, and S. Chanda, "Pho(sc)-ctc—a hybrid approach towards zero-shot word image recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 26, pp. 51–63, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:247597101
8. J. Gan, W. Wang, J. Leng, and X. Gao, "Higan+: Handwriting imitation gan with disentangled representations," *ACM Trans. Graph.*, vol. 42, no. 1, 2022. [Online]. Available: https://doi.org/10.1145/3550070
9. L. Kang, P. Riba, Y. Wang, M. Rusiñol, A. Forn'es, and M. Villegas, "Ganwriting: Content-conditioned generation of styled handwritten word images," *ArXiv*, vol. abs/2003.02567, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:212414769
10. J. Gan and W. Wang, "Higan: Handwriting imitation conditioned on arbitrary-length texts and disentangled styles," in *AAAI Conference on Artificial Intelligence*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235306524
11. P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *ArXiv*, vol. abs/2105.05233, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:234357997
12. F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, "Cvl-database: An off-line database for writer retrieval, writer identification and word spotting," in *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 560–564.
13. A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," 2023. [Online]. Available: https://arxiv.org/abs/2303.16203

Unsupervised Video Representation Learning: developments and applications

Elena Bueno-Benito

ebueno@iri.upc.edu

1 PhD context

Dr Mariella Dimiccoli supervises my PhD research at the Institut de Robòtica i Informàtica Industrial, CSIC-UPC. I began my doctoral studies in November 2021, with an expected completion in November 2025. The research is part of the Doctoral Degree in Automatic Control, Robotics, and Computer Vision (ARV), situated in an academic context. This work explores key challenges in robotics, machine learning, and video understanding, bridging theory with practical applications.

2 Introduction

Videos recording real-world scenarios often capture dynamic interactions between objects and individuals and constitute a rich source of information for several applications, ranging from video surveillance systems [10] to sports analytics [6]. However, processing long-untrimmed videos is challenging and despite major progress in video understanding in the last ten years, many current methods remain task-specific and depend heavily on large, manually annotated datasets. The reliance on labelled data presents a substantial challenge, particularly in tasks such as action segmentation, localization, and anticipation, where annotating data is both labour-intensive and time-consuming.

In light of these limitations, there is an increasing need to develop weakly-supervised and unsupervised methods that can effectively learn video representations for these tasks without the burden of extensive manual labelling, while maintaining adaptability to real-world applications. One of the first steps for understanding untrimmed videos is action segmentation, which involves the classification of each frame of an untrimmed video depicting a complex activity into distinct actions. This task plays a pivotal role in a range of applications, including video surveillance, sports analysis, and robotics [6, 15]. The weakly-supervised paradigm learns to partition videos into action segments using only transcript annotations for each video, typically in the form of action transcripts (ordered lists of actions) or action sets (unique actions derived from narrations, captions or meta-tags) [8, 12, 14]. This approach contrasts with unsupervised methods [5, 7, 11, 13], which can be broadly categorized into three types based on the matching objective: video-level, activity-level, and global-level [4]. In the absence of explicit action labels during training, these methods often employ the Hungarian Matching algorithm for evaluation.

The planned thesis aims at contrasting the dominance of supervised methods by proposing efficient solutions for action segmentation through innovative weakly/unsupervised techniques. It seeks to establish a new learning paradigm

that reduces dependence on extensive data, aligns more closely with human learning processes, and addresses domain adaptation issues, thus advancing both the methodology and practical applications in video understanding.

3 State-of-Art

Traditional methods for action segmentation relied heavily on hand-crafted features combined with classifiers [1, 2, 9]. However, with the advent of deep learning, there has been a shift towards end-to-end models that learn video representations directly from data. Despite the impressive performance of these supervised models, their reliance on large labelled datasets limits their scalability and generalization to new domains.

To mitigate the need for large annotated datasets, weakly-supervised techniques have been introduced [8, 12, 14]. These methods utilize only transcript annotations for each video, typically in the form of action transcripts (ordered lists of actions) or action sets (unique actions derived from narrations, captions or meta-tags). POC [8] introduces a loss function to ensure the output order of any two actions aligns with the transcript. Conversely, ATBA [12] propose an approach that incorporates alignment by directly localizing action transitions for efficient pseudo-segmentation generation during training, eliminating the need for time-consuming frame-by-frame alignment.

Unsupervised approaches aim to eliminate the need for manual annotations. Video-level matching [7] matches the cluster concerning the ground truth actions of a single video. Activity-level matching associates clusters to labels within each complex activity [11, 13]. As we can see here, the activity level of grouping leads to the assignment changes denoted by the coloured arrows.

Finally, recent work has addressed global action segmentation across entire datasets [5]. CAD [5] directly tackled global-level segmentation by relying on complex activity labels to discover constituent actions. However, it does not explicitly model the alignment of actions across videos of the same activity using high-level activity labels. Consequently, [5] can be considered a weakly-supervised method on a global scale, occupying a unique position in the spectrum of action segmentation approaches.

4 Methodology and Contributions

In this section, we outline our key contributions so far in two main areas: **Learning without data annotations (fully unsupervised)** and **Learning with pseudo-labels (weakly-supervised)**.

[3] E.Bueno-Benito, B.Tura and M.Dimiccoli (2023). "Leveraging triplet loss for unsupervised action segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*

⋆ E.Bueno-Benito and M.Dimiccoli (2024). "2by2: Weakly-Supervised Learning for Global Action Segmentation". In: *To appear in the International Conference on Pattern Recognition (ICPR).*

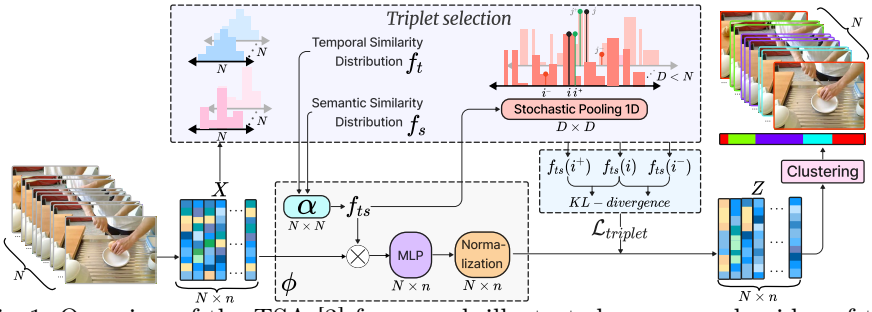


Fig. 1: Overview of the TSA [3] framework illustrated on a sample video of the Breakfast Dataset: network architecture transforming the initial features X into the learned features Z through a shallow network with a novel triplet selection strategy and a triplet loss based on similarity distributions.

4.1 Learning without data annotations (fully unsupervised)

Our primary contribution in this area, as introduced in [3], is a self-supervised training methodology that leverages both temporal and semantic relatedness in video data. This approach is based on two key observations: (i) actions in videos are composed of temporally continuous sequences of images, meaning that adjacent frames are likely to belong to the same action; (ii) frames depicting the same action, though not necessarily adjacent in time, should share similar representations, encoding the underlying common semantic.

To address these observations, we propose Temporal-Semantic Aware (TSA) representations, which are specifically designed for video action segmentation tasks. We hypothesize that atomic actions can be effectively modelled as clusters in a representational space. Our framework maps the initial feature space of a video into a new space where temporal-semantic clusters corresponding to atomic actions are revealed.

The technical innovation of our approach lies in its action representation learning, which employs a shallow neural network and a novel triplet loss function. This loss operates on similarity distributions, and our unique triplet selection strategy is based on a downsampled temporal-semantic similarity weighting matrix (as depicted in Fig. 1). Our TSA framework demonstrates superior performance compared to the state-of-the-art in action segmentation on the *Breakfast* (BF) and *Youtube INRIA Instructional* (YTI) benchmark datasets. For details on the method, ablation study, and results, see [3].

4.2 Learning with pseudo-labels (weakly-supervised)

As part of our second contribution, we present a framework that use activity labels to predict action labels at a global level, given a set of videos including different activities. This work, termed 2by2, proposes a triadic action learning approach(see Fig. 2), aiming at modeling:

(i) *Intra-video action discrimination* (video level): Video frames sharing the same action with their nearest neighbours exhibit temporal consistency.

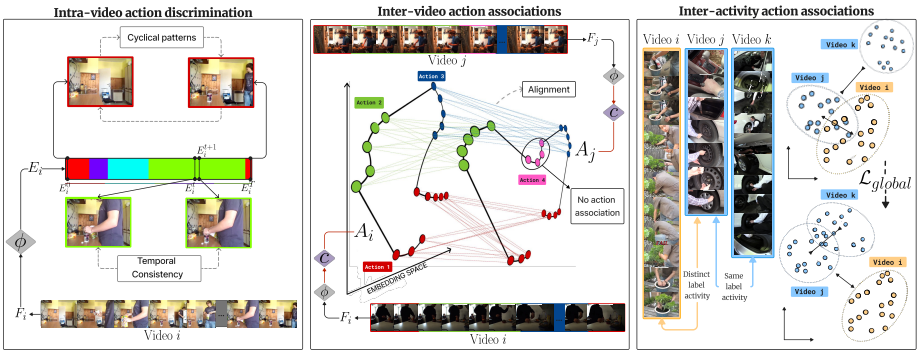


Fig. 2: Overview of the 2by2. The figure illustrates our triadic learning approach: intra-video action discrimination, which enhances cross-temporal consistency within a single video (first box); inter-video action associations, which align action frames among similar videos (second box); and inter-activity action associations, which establish global correspondence between different videos (third box). The red arrows indicate steps specific to the training phase.

Moreover, actions typically do not occur at the beginning or end of videos. Thus, a video can be interpreted as a cyclic temporal sequence.

(ii) *Inter-video action associations* (activity level): For videos categorized under the same activity, segments within these videos exhibit similarity, facilitating the alignment of actions across them.

(iii) *Inter-activity action associations* (global level): Videos representing different activities that share common actions should be closer in the representational space compared to those that do not share actions.

To learn these aspects, we construct a Siamese transformer-based network that takes input pairs of videos and determines if they belong to the same activity. If they do, the videos are also temporally aligned. A key innovation of our approach is the direct action alignment between videos, which is crucial for accurately matching corresponding segments. This is enabled by the Siamese two-stage architecture that ensures robust initialization for temporal alignment. The multi-level learning strategy in 2by2 significantly outperforms state-of-the-art approaches on the challenging BF and YTI datasets.

5 Actions planned before finishing the PhD

In the upcoming months, I plan to develop a novel transformer-based approach for unsupervised action segmentation that will integrate ideas from clustering techniques to jointly learn video segments and their labels. Additionally, I plan to extend my research to the related task of Long Term Anticipation from untrimmed video, since action segmentation is required to understand the observation interval.

References

1. Bahrami, E., Francesca, G., Gall, J.: How much temporal long-term context is needed for action segmentation? In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2023)
2. Behrmann, N., Golestaneh, S.A., Kolter, Z., Gall, J., Noroozi, M.: Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
3. Bueno-Benito, E., Tura, B., Dimiccoli, M.: Leveraging triplet loss for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
4. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: An analysis of modern techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
5. Ding, G., Yao, A.: Temporal action segmentation with high-level complex activity labels. IEEE Transactions on Multimedia (2023)
6. He, Y., Yuan, Z., Wu, Y., Cheng, L., Deng, D., Wu, Y.: Vistec: Video modeling for sports technique recognition and tactical analysis. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2024)
7. Li, Y., Xue, Z., Xu, H.: Otas: Unsupervised boundary detection for object-centric temporal action segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)
8. Lu, Z., Elhamifar, E.: Set-supervised action learning in procedural task videos via pairwise order consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19871–19881 (2022)
9. Lu, Z., Elhamifar, E.: Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
10. Sharma, R., Sungheetha, A., et al.: An efficient dimension reduction based fusion of cnn and svm model for detection of abnormal incident in video surveillance. In: Journal of Soft Computing Paradigm (JSCP). vol. 3, pp. 55–69 (2021)
11. Tran, Q.H., Mehmood, A., Ahmed, M., Naufil, M., Konin, A., Zia, M.Z.: Permutation-aware activity segmentation via unsupervised frame-to-segment alignment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)
12. Xu, A., Zheng, W.S.: Efficient and effective weakly-supervised action segmentation via action-transition-aware boundary alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
13. Xu, M., Gould, S.: Temporally consistent unbalanced optimal transport for unsupervised action segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
14. Zhang, R., Wang, S., Duan, Y., Tang, Y., Zhang, Y., Tan, Y.P.: Hoi-aware adaptive network for weakly-supervised action segmentation. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2023)
15. Zuckerman, I., Werner, N., Kouchly, J., Huston, E., DiMarco, S., DiMusto, P., Laufer, S.: Depth over rgb: automatic evaluation of open surgery skills using depth camera. International Journal of Computer Assisted Radiology and Surgery pp. 1–9 (2024)