Vol. 8, December 2023

Published on : 20-12-2023

GYANVESHAN

In Pursuit of Knowledge

The philosophy behind the name...

You may notice that "AN" is smaller than other letters in the title. It signifies something important. GYANVESHAN is a name built by combining two words – GYAN and ANVESHAN (with the common AN being merged into one) with GYAN meaning knowledge, and ANVESHAN meaning search or pursuit. That brings us to the tagline "In pursuit of knowledge".

Muthoot
Institute of Technology & Science

From Editor's Desk

The editorial board is proud to bring you the Eighth edition of our Technical Magazine Gyanveshan !!. In accordance with the name , this edition also contains a variety of research and thought stimulating articles that align with the motto " In Pursuit of Knowledge".

Fourteen technical articles in various branches of Engineering education by faculty members who are actively involved in research are included in this edition. Like previous editions, this edition of 'Gyanveshan' also exhibits our relentless passion for excellence.

The editorial board would like to take this opportunity to thank our Executive Director, Principal ,Vice-Principal and Dean, Academics for their valuable comments and extensive support for making this possible.

With great pleasure we are submitting this edition for your reading. Thanks to all who supported this edition through authoring, editing & publishing . Wishing all a pleasant reading!!

Dr. Sumam Mary Idicula

Faculty Chairperson, R&C Cell

GYANVESHAN, The Book of Articles - Vol 8, December 2023

"The common facts of today are the products of yesterday's research"

- Duncan MacDonald

Table of Contents

A comparative study on Fractional Order PID controller

Ambili Mohan & Breeza Paulose
Department of Electrical and Electronics Engineering

The PID (Proportional-Integral-Derivative) controllers are widely used in process control, manufacturing industries, power systems, robotics etc. even with the development of sophisticated control laws. PID controllers, which combine proportional, integral, and derivative elements, hold a fundamental position in control engineering as it is simple to incorporate in a system and each controllers work independent of each other. But they have limitations related to sensitivity towards parameter changes and nonlinearities. This has spurred exploration into alternative control approaches.

Fractional PID controllers introduce a fractional order into the standard PID structure, presenting a more adaptable way of handling system dynamics still keeping the advantage of simplicity as in traditional PID Controller. The examination of fractional PID controllers and conventional PID controllers within the control systems realm has attracted considerable interest. This study seeks to comprehend their performance in a specific category of systems, offering insights into their respective strengths and limitations.

This fractional aspect allows a more comprehensive representation of intricate systems, capturing behaviours that integer-order models cannot. This study aims to reveal how these fractional components affect controller performance, especially in conditions where traditional PID controllers might fail. Focusing on a specific set of systems adds a unique angle to this investigation. By honing in on particular systems, the goal is to pinpoint nuanced requirements and obstacles that might determine the superiority of one controller over the other. Whether it's a complex industrial process or a dynamic mechanical system, comprehending how fractional and traditional PID controllers interact with and impacts these systems holds vital importance in optimizing control strategies.

Fractional PID Controller Methodology

A Fractional PID controller, often known as an "FOPID" controller, operates within a control system framework utilizing fractional calculus to execute the PID control algorithm. These controllers are utilized in situations where system characteristics cannot be precisely managed using standard integer-order PID controllers. Essentially an evolved form of the traditional PID controller, the FOPID controller incorporates concepts from fractional calculus. Differing from integer-order PID controllers, it integrates fractional-order derivatives and integrals, offering greater adaptability in capturing intricate system behaviours. The controller's parameters include the conventional proportional, integral, and derivative gains, alongside fractional orders. This innovative approach enhances the controller's ability to handle complex processes, resulting in enhanced performance, reduced steady-state errors, and robustness across a wide array of applications spanning from control systems to robotics.

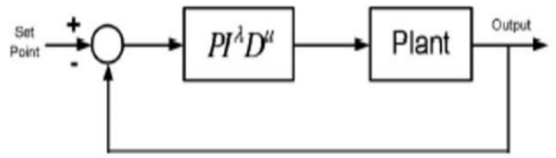


Fig 1 General block diagram of FPID

The transfer function of FPID controller is,

$$G_c(s) = \frac{U(s)}{E(s)} = K_P + K_I s^{-\lambda} + K_D s^{\mu}, \quad (\lambda, \mu > 0)$$

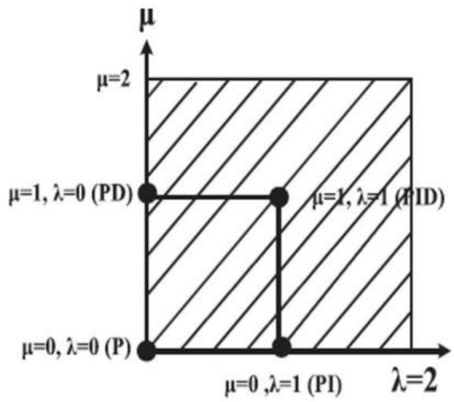


Fig 2. Plane of FOPID controller

Generally , the range of order for λ and μ is in between 0 - 2.

Here,

- When λ is 0 and μ is 0 , then it is called classical P controller.
- When λ is 1 and μ is 1 , then it is called classical PID controller.
- When λ is 0 and μ is 1 , then it is called classical PD controller.
- When λ is 1 and μ is 0 , then it is called classical PI controller

Tuning FOPID controllers poses a challenge compared to traditional PID controller due to the increased number of controller parameters. Traditional methods are often insufficient, leading to the use of optimization techniques. Gradient-based algorithms, evolutionary algorithms, metaheuristics, and hybrid approaches are commonly used. The choice depends on problem complexity, computational resources, and desired performance. Optimization requires a suitable objective function, appropriate constraints, and good initial parameters. These techniques have proven successful in applications like AVR systems, motion control, and process control. Another method is

Chaotic Atom Search Optimization (ChASO) for Fractional Order PID (FOPID) Controllers.

ChASO is a novel optimization technique specifically designed to tune FOPID controllers. It combines the strengths of the Atom Search Optimization (ASO) algorithm with the exploration capabilities of chaotic sequences. ASO mimics the atomic motion model in nature, where atoms attract and repel each other based on their potential energies. It iteratively updates the controller parameters based on these interactions, seeking the global minimum of the objective function.Chaotic sequences, like the logistic map, introduce randomness into the search process, helping the algorithm escape from local minima and explore a wider search space. This significantly improves the convergence speed and accuracy compared to the original ASO algorithm.

The steps to obtain the optimal values are

- Initialization: Generate a set of random solutions representing the FOPID controller parameters.
- Evaluation: Evaluate the objective function for each solution. This function typically reflects desired performance criteria like settling time, overshoot, and steady-state error.
- Attraction and Repulsion: Calculate the attraction and repulsion forces between each pair of solutions based on their objective function values.
- Update Positions: Update the position of each solution based on the combined forces and a chaotic sequence. This ensures both exploitation and exploration during the search process.
- Iteration: Repeat steps 2-4 until a stopping criterion is met, such as a maximum number of iterations or a desired convergence tolerance.

FOPID controller is implemented for the control of DC motor and the simulation depicted the outcomes in figures below, indicating that the FOPID controller delivers an improved response. The findings suggest that FOPID yields reduced error compared to traditional PID. This indicates an enhanced steady-state performance of FPID over traditional PID. Moreover, FOPID showcases swifter response and improved transient response than classical PID. Despite FPID's advantages over PID, it remains considerably complex due to tuning challenges.

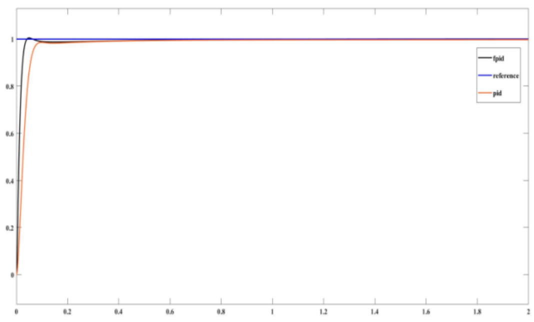


Fig 3 Step response of FPID AND PID controller

Unlike a traditional PID controller, a FOPID controller can meet five distinct specifications, presenting a level of versatility that is unattainable with the conventional counterpart. Specifically, higher-order systems and systems with extended time delay demonstrate improved performance with a FOPID controller while exhibiting comparatively impaired performance with a classical PID controller.. Additionally, in terms of robustness and stability, FOPID controllers are superior to classical PID controllers, especially in non-minimum phase systems. Furthermore, in scenarios involving non-linear systems that are linearized at different operating points with corresponding controllers, a FOPID controller typically proves sufficient for the task at hand.

Simulation results further affirm the dominance of FOPID over PID, displaying lower error rates and quicker responses. Despite of these advantages, FOPID's main drawback lies in the complexity of tuning controller parameters.

References:

1. Kumar, V.; Rana, K. “Comparative study on fractional order PID and PID controllers on noise suppression for manipulator trajectory control”,Fractional Order Control and Synchronization of Chaotic Systems; Springer: New York, NY, USA, 2017
2. Bettayeb, M., and R. Mansouri. “*Fractional IMC-PIDfilter controllers design for non-integer order systems*”. Journal of Process Control, 24.4: 261-271.
3. Sagar S, Kaur S, Sondhi S. “ *Fractional Order IMCPID Controller Design for the Pressurized Heavy Water Reactor*”. International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences (EEECOS), 2016.

Advancements in Olfactory Assistance through Artificial Intelligence

Divya .K.S
Department of Artificial Intelligence & Data Science

Abstract

The human olfactory system plays a crucial role in daily life, influencing perception, memory, and overall well-being. However, replicating this intricate sense in artificial intelligence (AI) systems presents a significant challenge due to the complexity and subjective nature of smell. This article explores the advancements and potential applications of AI in aiding olfactory functions. Applications of AI in olfactory assistance span across diverse sectors. In the food and beverage industry, AI-powered e-noses facilitate quality control by detecting odors indicative of spoilage or contamination. Environmental monitoring benefits from AI's ability to identify specific odorous compounds, aiding in pollution control and assessment. Additionally, AI-driven olfaction holds promise in healthcare, enabling the detection of diseases through breath analysis and assisting individuals with olfactory impairments.

Index Terms - Olfactory , odor classification, AI nose

Introduction

Recent research has focused on leveraging AI to enhance olfactory capabilities, primarily through the development of electronic noses (e-noses) equipped with sensor arrays. These devices detect and analyze volatile compounds, generating complex data patterns that AI algorithms interpret to identify and categorize odors. Machine learning models, including neural networks and pattern recognition algorithms, have been employed to recognize and differentiate between various smells based on acquired datasets..

Challenges persist in replicating the nuanced human olfactory system accurately. Subjectivity in odor perception, the vast array of possible odors, and the need for extensive and diverse datasets for AI training are significant hurdles. Achieving robust and reliable odor identification akin to human capabilities remains an ongoing pursuit in AI research. An artificial-intelligence system can describe how compounds smell simply by analysing their molecular structures — and its descriptions are often similar to those of trained human sniffers.[1]

The human olfactory system is a remarkable sensor, capable of detecting and distinguishing thousands of distinct odors. Yet, replicating this intricate sense in machines has proven to be a challenging endeavor. The ability to identify, classify, and interpret odors is a multifaceted process that involves the detection of volatile compounds and the intricate processing of sensory information in the brain.

AI approaches to olfactory identification

Odor identification using AI involves teaching machines to recognize and differentiate between various smells or odors. While humans have a remarkable ability to discern and categorize odors, replicating this capability in machines presents unique challenges due to the complexity and subjective nature of smell perception.

Electronic Nose (E-Nose):

AI can be integrated into electronic nose devices that contain an array of sensors to detect and analyze odors. These sensors generate data that AI algorithms can interpret to identify and classify different smells. Machine learning models can be trained on these data sets to recognize patterns and associate them with specific odors.

Machine Learning and Pattern Recognition:

AI algorithms, particularly machine learning models like neural networks, can be trained on large datasets of odor samples to learn patterns and associations. This allows them to identify and classify smells based on the input data they receive.

Gas Chromatography-Mass Spectrometry (GC-MS):

AI can be used to analyze data from GC-MS, a technique used to separate and analyze compounds within a sample. AI algorithms can interpret these complex data sets to identify and classify specific odor compounds.[4]

AI Nose:

An AI nose, often referred to as an electronic nose or e-nose, is a technology designed to replicate and mimic human olfactory capabilities using artificial intelligence and sensor arrays. While it doesn't possess the biological complexity of a human nose, an AI nose aims to detect, analyze, and identify odors by employing various sensors and AI algorithms.

AI nose and its principal components

Sensor Array: A collection of different types of sensors capable of detecting volatile compounds present in odors. These sensors generate data based on the chemical composition of the detected compounds. The sensors used can be multichannel sensor, Fiagro 822 gas sensors etc.

Data Acquisition System: This gathers and processes the data from the sensor array, translating the chemical information into digital signals that can be interpreted by AI algorithms.

Artificial Intelligence (AI) Algorithms: Machine learning models, such as neural networks or pattern recognition algorithms, are employed to interpret the sensor data. These algorithms learn from training datasets to recognize and classify different odors based on the patterns in the data.eg PCA, Regression models ANN etc[4]

Applications of AI aided olfactory mechanism

Quality Control: In industries like food and beverage, an AI nose can detect odors associated with spoilage or contamination, aiding in quality control processes.eg : In the field coffee tasting and quality assurance. Aromas and flavors are among the most important sensory attributes that consumers consider when assessing coffee quality and liking[2,3,5]

Environmental Monitoring: It can be used for detecting and identifying pollutants or harmful gases in the environment, assisting in monitoring air quality or identifying hazardous substances.eg: The presence of Carbon Monoxide in the testing lab and closed air conditioned spaces.

Healthcare: AI noses hold potential in healthcare for disease detection through breath analysis. They could assist in identifying specific volatile organic compounds (VOCs) associated with certain illnesses.eg: Predicting the presence of a disease based on the odor of the patient.

Security and Safety: AI noses might be used in security systems to detect explosives, narcotics, or other illegal substances based on their distinct odors.eg : Identifying the drug and illicit trafficking in airports.

Conclusion

As research progresses, continued efforts in refining sensor technologies, advancing machine learning algorithms, and expanding datasets will likely propel AI-driven olfactory assistance further. Strides in these areas hold the potential to not only refine existing applications but also unlock novel uses for AI noses, transcending the limitations of human olfaction and offering innovative solutions to real-world challenges.

References:

1. Lee, B. K. et al. "AI predicts chemicals' smells from their structures". Science 381, 999–1006 (2023). doi:https://doi.org/10.1038/d41586-023-02714-2
2. Claudia et.al "Integrating a Low-Cost Electronic Nose and Machine Learning Modelling to Assess Coffee Aroma Profile and Intensity". https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7998415 doi: 10.3390/s21062016
3. Bona E., Silva R.S.D.S.F.D., Borsato D., Bassoli D.G. Optimized Neural Network for Instant Coffee Classification through an Electronic Nose. *Int. J. Food Eng.* 2011;**7**:7. doi: 10.2202/1556-3758.2002.
4. Michishita et.al "Gas Chromatography/ Olfactometry and Electronic Nose Analyses of Retronasal Aroma of Espresso and Correlation with Sensory Evaluation by an Artificial Neural Network". *J. Food Sci.* 2010;75:S477–S489. doi: 10.1111/j.1750-3841.2010.01828
5. Gianmarco et.al ."Accelerated estimation of coffee sensory profiles using an AI-assisted electronic tongue".(2022) https://www.sciencedirect.com/science/article/pii/S1466856422002909 https://doi.org/10.1016/j.ifset.2022.103205

Augmented Reality in Manufacturing: A Comprehensive Review

Dr. Bibin Jose
Department of Mechanical Engineering

Abstract

Augmented reality (AR) serves as a unique interface, seamlessly merging virtual computer-generated information with real-world applications. Recently, its versatile potential has been recognized across diverse fields, spanning military training, surgery, entertainment, maintenance, assembly, product design, and other production processes. Crucially, AR plays a pivotal role in addressing Industry 4.0 challenges, facilitating technology integration. In the manufacturing sector, AR fosters a shift from mass production to mass customization, with a myriad of industrial applications leveraging AR technologies. Its popularity stems from easy implementation in production and broad accessibility through devices like smartphones and tablets. The article provides a comprehensive assessment of the latest AR applications and advancements in manufacturing, shedding light on challenges and potential future solutions. The widespread adoption of AR systems is anticipated, marking a transformative trajectory for the manufacturing industry.

Index Terms - Augmented reality, Manufacturing, Product design, Industry 4.0

Introduction

In the dynamic landscape of modern manufacturing, the integration of cutting-edge technologies is reshaping traditional paradigms and propelling the industry towards unprecedented efficiency and innovation [1-3]. One such transformative technology that has captured the attention of manufacturing professionals worldwide is AR. AR represents a paradigm shift, bridging the gap between the physical and digital realms, and offering a myriad of possibilities to enhance various facets of manufacturing processes. At its core, AR involves the overlay of digital information onto the real-world environment, thereby augmenting the user's perception and interaction with their surroundings. In the context of manufacturing, this means that workers can access relevant data, instructions, and insights in real-time, directly within their field of view. This paper delves into the realm of AR in manufacturing, aiming to provide a comprehensive understanding of its evolution, core technologies, applications, benefits, challenges, and future prospects.

The manufacturing sector has always been at the forefront of adopting innovative technologies to drive efficiency and productivity [4]. From the advent of the assembly line to the integration of automation and robotics, each era has seen transformative changes. AR is the latest addition to this technological continuum, promising to revolutionize the way manufacturing processes are executed, managed, and optimized.

The journey of AR in manufacturing is intertwined with the evolution of hardware and software capabilities. Over the years, advancements in computer vision, tracking systems, and display technologies have enabled the seamless integration of digital information into the physical workspace. The convergence of AR with other transformative technologies like the Internet of Things (IoT) and Artificial Intelligence (AI) further amplifies its potential impact on manufacturing operations [5]. As manufacturing processes become more complex and data-driven, the need for intuitive and real-time information becomes paramount. AR addresses this need by providing a contextual layer of information directly in the user's field of vision. Whether it's assisting assembly line workers in intricate tasks, providing maintenance technicians with step-by-step guidance, or offering real-time quality control insights, AR empowers manufacturing professionals with unprecedented capabilities.

This paper aims to unravel the multifaceted dimensions of AR in manufacturing. Through a systematic exploration of its historical evolution, core technologies, applications, benefits, challenges, and future trends, we seek to provide a comprehensive and insightful overview. By doing so, we aspire to contribute to the broader understanding of Augmented Reality's role in shaping the future of manufacturing, fostering a landscape where human expertise converges seamlessly with the power of digital augmentation.

Core Technologies in Augmented Reality

The core technologies powering AR include computer vision, tracking systems, display technologies, and sensory input devices [4]. Computer vision enables AR to interpret and understand the user's physical environment, while tracking systems ensure accurate alignment of virtual content with the real world. Display technologies project digital information seamlessly into the user's field of view, enhancing their perception. Sensory input devices, such as cameras and sensors, facilitate interaction and feedback. Together, these technologies create a dynamic interface, allowing users to seamlessly interact with digital information overlaid on the physical world, shaping the immersive and transformative experiences characteristic of augmented reality applications.

Augmented Reality Applications in Manufacturing

Augmented Reality applications in manufacturing have become increasingly prevalent, offering innovative solutions to streamline processes and enhance efficiency. Some notable applications include:

Assembly Guidance: AR assists workers in assembling complex products by providing step-by-step visual instructions overlaid onto the physical components. This reduces errors, accelerates assembly times, and aids in training new personnel.

Maintenance and Repairs: Maintenance technicians benefit from AR by accessing real-time information and instructions during equipment repairs. AR overlays relevant data, schematics, and guidelines, improving accuracy and minimizing downtime.
Training Simulations: AR-based training simulations allow workers to practice tasks in a virtual environment before executing them in real-world scenarios. This enhances skill development, particularly for intricate or hazardous processes.
Quality Control: AR is employed for quality assurance by overlaying digital indicators and specifications onto physical products. This ensures consistency and precision in the inspection process, reducing defects and improving overall product quality.
Product Design and Prototyping: AR facilitates collaborative product design by projecting virtual prototypes into the physical workspace. This enables real-time visualization and adjustments, fostering efficient communication among design teams.
Logistics and Warehousing: AR aids in optimizing logistics operations by providing visual cues for inventory management, order picking, and packing. This results in faster and more accurate fulfilment processes within warehouses.
Safety Training: AR is utilized for immersive safety training scenarios, allowing workers to experience and respond to potential hazards in a controlled virtual environment. This enhances safety awareness and preparedness.
Customization in Mass Production: AR supports the transition from mass production to mass customization by providing real-time customization options on the production line. This allows for flexibility in meeting diverse customer requirements.
Data Visualization: AR visualizes complex data sets and analytics in real-time, providing managers and operators with insights into manufacturing processes. This data visualization aids in decision-making and process optimization.

These applications collectively showcase the versatility of augmented reality in manufacturing, offering solutions that range from enhancing precision and productivity to improving safety and training methodologies. As technology continues to advance, the scope for innovative AR applications in the manufacturing sector is likely to expand further.

Challenges and Limitations

Despite the promising applications of Augmented Reality (AR) in manufacturing, several challenges and limitations impede its seamless integration [6]. One significant challenge lies in the upfront costs associated with implementing AR systems, including hardware, software, and training expenses. Additionally, the technology's maturity poses hurdles, with certain industries still navigating the learning curve and awaiting standardized practices. Data security concerns also loom large, as the reliance on interconnected devices increases vulnerability to cyber threats. Furthermore, achieving optimal user experiences requires addressing issues such as field of view limitations and the potential for visual clutter. As the technology advances, interoperability between AR systems and existing infrastructure remains a challenge, hindering widespread adoption. Overcoming these obstacles demands concerted efforts from industry stakeholders, technological innovators, and regulatory bodies to ensure that the benefits of AR in manufacturing can be fully realized.

Future Trends and Prospects

The future of Augmented Reality in manufacturing holds exciting prospects with several emerging trends shaping the landscape [7]. As technology continues to evolve, the integration of 5G connectivity is set to enhance AR applications by providing faster and more reliable communication [8-10]. This enables real-time data transfer, improving responsiveness in manufacturing processes. Wearable devices, such as AR-enabled smart glasses, are anticipated to become more prevalent, offering hands-free interaction and a seamless user experience. The convergence of AR with Artificial Intelligence (AI) is another promising trend. AI algorithms can enhance AR applications by providing intelligent insights, predictive analytics, and dynamic adaptability. This synergy enables more sophisticated and context-aware interactions within the manufacturing environment.

The evolution of AR software is expected to focus on creating more intuitive and user-friendly interfaces, reducing the learning curve for operators and technicians. Additionally, advancements in spatial computing technologies will contribute to more accurate and immersive AR experiences.

As manufacturing embraces the concept of smart factories, AR is poised to play a central role in creating interconnected, data-driven environments. The integration of AR into broader Industry 4.0 initiatives will further optimize production processes, enabling greater efficiency, customization, and flexibility.

The future trends and prospects of AR in manufacturing paint a picture of a technologically advanced and interconnected industry. The combination of 5G, wearables, AI integration, and improved software interfaces will likely propel AR into becoming a ubiquitous tool, transforming how manufacturing processes are executed and managed. As these trends unfold, AR is poised to contribute significantly to the ongoing evolution of the manufacturing sector.

Conclusion

Augmented Reality (AR) has emerged as a transformative force within the manufacturing sector, offering a versatile set of applications that redefine traditional processes. The comprehensive review of AR in manufacturing has illuminated its historical evolution, core technologies, diverse applications, benefits, challenges, and future prospects. AR's impact on manufacturing is undeniable, evident in its ability to enhance assembly processes, streamline maintenance tasks, revolutionize training methods, and improve overall quality control. The technology's role in facilitating the transition from mass production to mass customization underscores its significance in an era where flexibility and adaptability are paramount. However, challenges such as upfront costs, technological maturity, and data security concerns pose barriers to widespread adoption. Overcoming these challenges requires concerted efforts from industry stakeholders and ongoing technological advancements. Looking ahead, the future of AR in manufacturing holds promise, with trends such as 5G connectivity, wearable devices, and the integration of AI poised to elevate its capabilities. The envisioned smart factories of tomorrow will likely rely heavily on AR to create interconnected, data-driven environments that optimize production processes. As AR continues to evolve, addressing current challenges and embracing emerging trends will be pivotal in unlocking its full potential within the manufacturing domain. This review aims to contribute to the collective understanding of AR's current state while providing insights that can guide future research, development, and implementation efforts in the dynamic intersection of augmented reality and manufacturing.

References:

1. Chryssolouris G, Mavrikios D, Papakostas N, et al. Digital manufacturing: History, perspectives, and outlook. Proc Inst Mech Eng Part B J Eng Manuf 2009; 223: 451–462.
2. Schmalstieg D, Höllerer T. Augmented reality: Principles and practice. In: 2017 IEEE Virtual Reality (VR). 2017, pp. 425–426.
3. Wu S-K, Hu SJ, Wu SM. Optimal door fitting with systematic fixture adjustment. Int J Flex Manuf Syst 1994; 6: 99–121.
4. Santi GM, Ceruti A, Liverani A, et al. Augmented Reality in Industry 4.0 and Future Innovation Programs. Technologies; 9. Epub ahead of print 2021. DOI: 10.3390/technologies9020033.
5. Ong SK, Yuan ML, Nee AYC. Augmented reality applications in manufacturing: a survey. Int J Prod Res 2008; 46: 2707–2742.
6. Caruso G, Re GM. Interactive augmented reality system for product design review. In: Proc.SPIE, p. 75250H.
7. Baroroh DK, Chu C-H, Wang L. Systematic literature review on augmented reality in smart manufacturing: Collaboration between human and computational intelligence. J Manuf Syst 2021; 61: 696–711.
8. Eswaran M, Bahubalendruni MVAR. Challenges and opportunities on AR/VR technologies for manufacturing systems in the context of industry 4.0: A state of the art review. J Manuf Syst 2022; 65: 260–278.
9. Park J. Augmented Reality Based Re-formable Mock-Up for Design Evaluation. In: 2008 International Symposium on Ubiquitous Virtual Reality. 2008, pp. 17–20.
10. Chen CJ, Ong SK, Nee AYC, et al. Haptic-based interactive path planning for a virtual robot arm. Int J Interact Des Manuf 2010; 4: 113–123

Best Subset Regression

Emil Baby
Basic Science and Humanities Department

Abstract

Regression analysis is concerned with the study of the relationship between one variable called the explained or dependent variable and one or more other variables called independent variable or explanatory variables. Regression is a cause and effect relationship between dependent and independent variables. If the relationship between a dependent variable *Y* and an independent variable *X* are linear, it is known as simple linear regression of *Y* on *X*. The functional relationship of a variable *Y* with other two or more independent variable is termed as multiple regression. In this case A large number of independent variables are available for the purpose of prediction. But all of these available variables are not necessary for prediction. So we want to select a subset of these available variables for the accuracy of prediction.

Index Terms - Coefficient of determination R^2, Residual Mean Square S^2 , Mallows's C_P Statistic

Introduction

The aim of regression analysis is to discover the relationships, if any, between the response Y and the explanatory variables $X_1, X_2, ..., X_k$. Ultimately, we may wish to use these relationships to make predictions about *Y* based on observing $X_1, X_2, ..., X_k$. A large number of independent variables are available for the purpose of prediction. But all of these available variables are not necessary for prediction. So we want to select a subset of these available variables for the accuracy of prediction.

Building a regression model that includes only a subset of the available regressors involves two conflicting objectives.

We would like the model to include as many regressors as possible so that the information content in these factors can influence the predicted value of Y. We want the model to include as few regressors as possible because the variance of the prediction $\hat{Y}$ increases as the number of regeressors increases. Also the more regressors there are in a model, the greater the costs of data collection and model maintenance. The process of finding a model that is a compromise between these two objectives is called selecting the "best" regression equation.

The literature on model selection and prediction sometimes makes the assumption that the true model is one of the forms.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \quad (1)$$

In this case a key issue is identifying the variables that are not related to the response, that is identifying the β's that are zero. This is a model selection in its pure form.

If we want to select a subset of variables, there are two main approaches. In the first, known as Best Subset Regression, we define criteria of model goodness, evaluate the criteria for each possible subset of variables and then choose the subset that optimizes the criteria.

The second approach is the Stepwise Regression method, in which a sequence of hypothesis tests is applied to the problem in order to identify the non zero β 's in (3.1). These techniques, of which forward selection, backward elimination and stepwise regression are the best known examples, obviously make the assumptions that (3.1) is the true model.

Best Subset Regression Method

The Best Subset Regression procedure requires that the analyst fits all the regression equations involving one candidate regressor, two candidate regressors and so on. These equations are evaluated according to some criterion and the "best" regression model is selected. If we assume that the intercept term β_0 is included in all equations, then if there are *k* candidate regressors, there are 2^k total equations to best estimated and examined.

The best subsets regression procedure can be used to select a group of likely models for further analysis. The general method is to select the smallest subset that fulfills certain statistical criteria. The reason that we would use a subset of variables rather than a full set is because the subset model may actually estimate the regression coefficients and predict future responses with smaller variance than the full model using all predictors.

Criteria for evaluating best subset regression models

Coefficient of determination R^2

Residual Mean Square S^2

Mallows's C_P Statistic

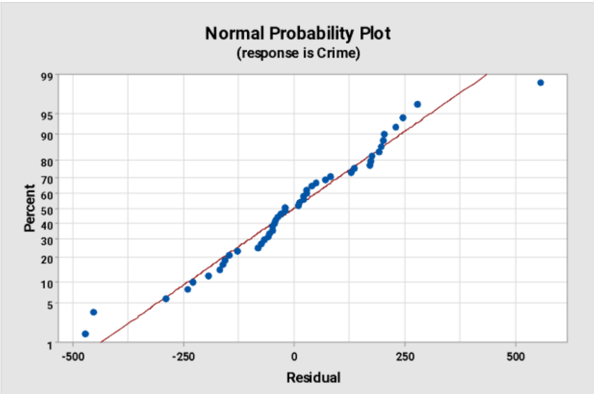
In order to find a best regression model for modelling the response variable 'crime rate' using the other variables using aggregate data on 47 states of the USA for 1960 is collected from online website www.statsci.org. we apply the best subset regression procedure and the output is summarized in the following table

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time
1	47.3	46.1	39.3	40.0	283.93				X											
1	44.5	43.2	35.9	44.5	291.44					X										
1	19.5	17.7	11.8	83.8	350.90												X			
2	58.0	56.1	48.6	25.1	256.19				X									X		
2	56.2	54.3	48.0	27.9	261.59	X			X											
2	55.5	53.5	45.0	29.1	263.81					X								X		
3	66.6	64.2	57.5	13.6	231.31			X	X									X		
3	64.6	62.2	55.6	16.7	237.88				X			X						X		
3	63.8	61.3	53.5	18.0	240.75			X		X								X		
4	70.0	67.2	60.9	10.2	221.54	X		X	X									X		
4	70.0	67.1	59.8	10.3	221.78			X	X									X	X	
4	68.8	65.8	58.8	12.1	226.14				X			X						X	X	
5	73.8	70.6	64.1	6.3	209.72	X		X	X									X	X	
5	73.0	69.7	63.7	7.6	213.01	X		X	X							X		X		
5	72.4	69.1	60.7	8.4	215.16	X		X	X								X	X		
6	76.6	73.1	66.6	3.9	200.69	X		X	X							X		X	X	
6	75.0	71.3	63.0	6.3	207.17	X		X	X								X	X	X	
6	75.0	71.3	63.8	6.3	207.29	X		X	X			X						X	X	
7	77.5	73.4	64.9	4.5	199.43	X		X	X							X	X	X	X	
7	77.4	73.3	66.2	4.6	199.76	X		X	X						X	X		X	X	
7	77.3	73.3	66.0	4.7	199.98	X		X	X				X			X		X	X	
8	78.9	74.4	66.8	4.2	195.55	X		X	X			X			X	X		X	X	
8	78.3	73.8	65.9	5.1	198.02	X		X	X				X		X	X		X	X	
8	78.3	73.7	64.6	5.2	198.27	X		X	X				X			X	X	X	X	
9	79.3	74.2	65.0	5.6	196.36	X		X	X			X			X	X	X	X	X	
9	79.1	74.0	65.1	5.9	197.03	X		X	X			X	X		X	X		X	X	
9	79.1	74.0	65.4	5.9	197.20	X		X	X	X		X			X	X		X	X	
10	79.6	73.9	63.7	7.1	197.50	X		X	X			X	X		X	X	X	X	X	
10	79.5	73.8	63.6	7.3	198.14	X		X	X	X		X			X	X	X	X	X	
10	79.4	73.7	63.0	7.4	198.21	X		X	X	X			X		X	X	X	X	X	
11	79.8	73.5	62.1	8.7	199.11	X		X	X	X		X	X		X	X	X	X	X	
11	79.7	73.3	61.5	9.0	199.80	X		X	X			X	X	X	X	X	X	X	X	
11	79.6	73.3	61.7	9.0	200.03	X	X	X	X			X	X		X	X	X	X	X	
12	80.0	72.9	60.0	10.5	201.16	X		X	X	X		X	X	X	X	X	X	X	X	
12	79.9	72.9	60.1	10.6	201.49	X	X	X	X	X		X	X		X	X	X	X	X	
12	79.9	72.9	59.6	10.6	201.52	X		X	X	X	X	X	X		X	X	X	X	X	
13	80.2	72.3	57.2	12.2	203.40	X		X	X	X	X	X	X	X	X	X	X	X	X	
13	80.2	72.3	55.7	12.2	203.44	X		X	X	X		X	X	X	X	X	X	X	X	X
13	80.1	72.3	54.7	12.3	203.69	X		X	X	X	X	X		X	X	X	X	X	X	X
14	80.3	71.7	52.4	14.0	205.77	X		X	X	X	X	X	X	X	X	X	X	X	X	X
14	80.2	71.5	52.8	14.2	206.45	X	X	X	X	X		X	X	X	X	X	X	X	X	X
14	80.2	71.5	53.6	14.2	206.55	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
15	80.3	70.8	48.6	16.0	209.06	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

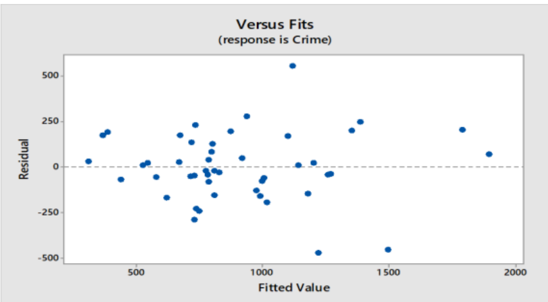
Model Adequacy Checking

In a regression model, the error terms, ε_t, is assumed to follow certain assumptions. The error terms should be independent normal random variables with mean zero and a constant variance. In practice we can't observe the error terms, but we do have estimates of them; we have the residuals, e_t, calculated from the estimated model. If the fitted regression model is a good model for the data, the residuals should satisfy the assumptions of random error terms.

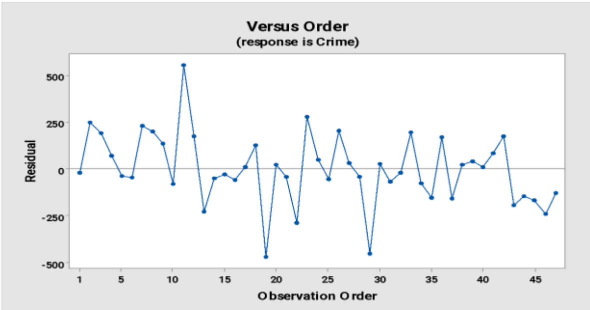
In order to check the normality assumption, Normal probability plot of the residuals are drawn and are shown in the following figure.



The normal probability plot of residuals shows that the normality assumptions of the error terms is satisfied.The assumption i) the error term have mean zero and ii) the error terms have constant variance can be checked using the using the residual plot shown in the following figure.



Here the plot for residual versus fitted value shows that mean zero and constant variance assumption of error terms is true since the plot does not have a pattern and the scattering is equal on both sides.



Here the plot for residuals versus order does not show any strong pattern in the ordered residuals. So the assumption of independence of error terms is also satisfied.

Thus the diagnostic checking reveals that the fitted multiple regression model is statistically adequate and we may use this model to predict the crime rate using the variables.

Conclusion

Best Subset Regression emerges as a robust method for variable selection, offering a systematic exploration of predictor combinations. Its ability to balance simplicity with predictive accuracy, handle multicollinearity, and provide clear model comparisons makes it a valuable tool in regression analysis. Despite its computational demands, advancements in technology make its implementation increasingly feasible, solidifying its relevance in the evolving landscape of data analysis.

References:

1.Douglas C. Montgomery: Introduction To Linear Regression analysis, Third edition, John Wiley and sons, New York.

2.Mallows, C.L (1995). More comments on Cp. Technometrics.

3.Anderson, T.W (June 1984), 'An Introduction to Multivariate Statistical Analysis', John Wiley and sons, New York.

4.Damodar Gujarati (1999), 'Essentials of Econometrics', second edition, McGraw Hill Company

5. Hald, A. (1952). Statistical Theory with Engineering Applications. New York : Wiley

ClinicalBERT- A TOOL FOR PROCESSING CLINICAL TEXTS

Dr. Jina Varghese
Department of Artificial Intelligence & Data Science

Abstract

Modern digital era is worn out with vast amount of versatile data. Scientists are exploring the data to make human life relaxed. A major contributor of this complex data is healthcare industry, which generates data in various forms. Analysis of health care data is more vital since it directly affects the humankind and it is a complex task as the data remains unstructured. And the operational and computational cost of data analysis are also increasing exponentially. Also, it is a tedious task to deduce a summary from medical documents. An efficient process to analyse the clinical data is highly appreciated. Clinical information has several forms such as texts, images, videos and audio clips. Clinical texts are textual descriptions such as information about the medicines, personal information, history of treatments, treatment scenarios, special notes etc. To aid medical practices, AI is expanding to medical sciences. Artificial Intelligence techniques (AI) can reduce the cost by automating the data analysis. Encompassing NLP (Natural Language Processing- a subfield of AI) to medical analysis will be a solution to the current medical field problems since it can reduce the operational and computational costs.

Different sources are generating clinical data of patients such as Electronic Health Records (EHR), information from various medical departments like radiology and clinical variables such as lab values and medications etc. Integration of all these notes is essential for the proper diagnosis. Most of the reports are written in English language. Natural Language Processing has showcased its ability to process the texts written a natural language, particularly English. NLP employs several preprocessing steps to extract the useful information from these data and it can estimate the relationship or association between these variables or documents. Clinical narratives are other forms of clinical texts which delivers more detailed and personalized history and assessments. Scientists employ this narrative for building the model. Since the clinical data are high dimensional, and sparse, a more sophisticated AI-NLP model is on demand.

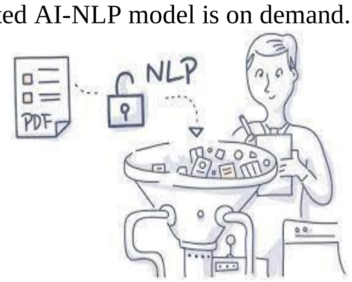


Figure 1: Natural Language Processing in Medical Field

How to process clinical texts

In order to make the system understand the natural language, documents should be translated to machine readable form through pre-processing. Basic pre-processing operations for inputs in natural languages are tokenization (split the sentence into words), stemming (convert each word to its base form, for eg: 'walking' is converted to 'walk'), removal of stop words (stop words are words in a language without any specific meaning, for eg: 'of', 'the', etc) etc. Encoding and embedding mechanisms are used for converting the text in natural language to numerical vector. Encoding, specifically one-hot encoding creates a vector (of size equal to the size of vocabulary), for each word, where all the elements in the vector are 0s except for the respective word for which the value is 1.

For eg: If the corpus = ['This is the first document.',
... 'This document is the second document.',
... 'And this is the third one.',
... 'Is this the first document?',
...].
The vocabulary includes the words array(['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third','this']) and the encoded vector is shown below.
array([[0, 1, 1, 1, 0, 0, 1, 0, 1],
[0, 2, 0, 1, 0, 1, 1, 0, 1],
[1, 0, 0, 1, 1, 0, 1, 1, 1],
[0, 1, 1, 1, 0, 0, 1, 0, 1]])

As written in the encoded vector, it contains the frequency of each words of the vocabulary in the corpus. Encoding schemes fails to express the semantic association between words in a text. Hence, we convert encoded vectors to embedded vector where relationships between words in the corpus will be expressed as numerical vectors. When we say, processing texts, the model operates on embedded numerical vector, which is machine understandable.

1 Basic architectures to process natural language

Generally, texts are sequential in form. Sequential data preserves the order of the input and takes previous data into account while processing a word. Generally used architectures for processing sequential data are *Recurrent Neural Network (RNN), LSTM, Transformers and BERT and its variants.*

RNN [1], **LSTM** [2]: RNNs (see figure 2) are generally used to process the sequential information. Each word in the input language is given sequentially. As shown in figure 2, input in each timestep uses the processed information from its previous timestep. Word at time t is processed along with the essence generated from the previous word at time t-1 (through hidden layer), and the output is passed to next timestep where it process the word at time t+1. RNNs often have short memories and start to forget previous inputs, in long sentences. But as the length of the input clinical text increases, it leads to Vanishing and Exploding gradients issue. LSTM and GRU are better solutions for this problem. They incorporate memory cell to remember these essential information from long sequences. Different gates are used in LSTM and GRU to forget the unnecessary information from past words and to add new useful information from the current word, to the memory. They perform better than normal RNN.

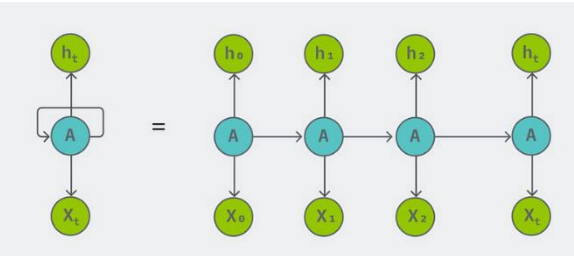


Figure 2: RNN-Recurrent Neural Network

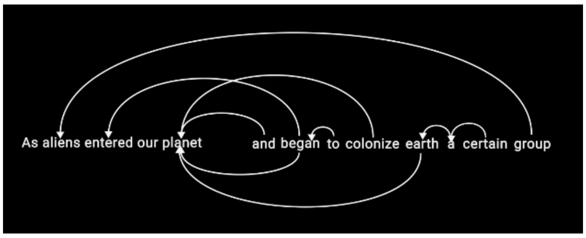


Figure 3: Attention mechanism focusing on different words(tokens) while processing each work

Transformers [3]: It is a better architecture with faster training time compared to LSTMs, as they allow for better parallelization during training. Transformers can handle larger input sequences because of their self-attention layers and ability to analyze words in parallel. Transformers are able to generate long-term relationships between the data elements. *Transformers* employs *self-attention* mechanism as shown in figure 3, to identify and weigh the importance of different parts of the input sentence by attending to itself. It learns and selects the appropriate past words for further processing. Architecture of transformers is given in figure 6, a bit complex structure. It follows Encoder-Decoder structure (shown in figure 5, wherein encoder takes the input(clinical text) and generates the essence of the input text in the form of numerical vectors(Encoded vector), and pass to decoder to generate the possible output numerical value corresponds to an output text/word. In the architecture of transformer, (see figure 6), the left part is the encoder part and the Decoder is on the right. Both Encoder and Decoder are composed of various modules that can be stacked on top of each other multiple times. The modules consist mainly of Multi-Head Attention (see figures 3 and 4) and Feed Forward layers. Now, relationships between words are expressed in terms of attention scores.

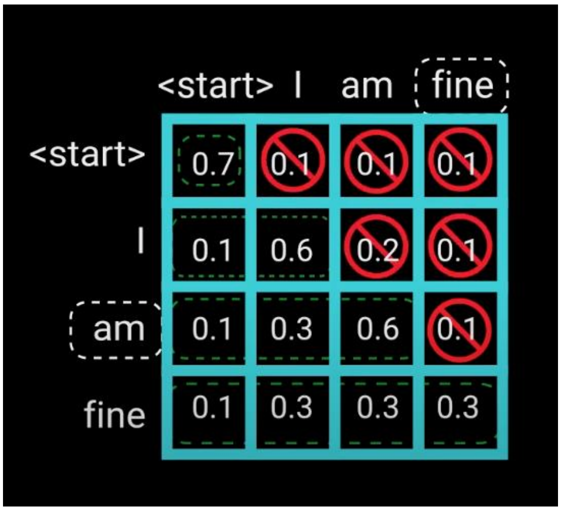


Figure 4: MultiHeadAttention Scores: To prevent computing attention scores for future words, some values are masked (values in red circle).

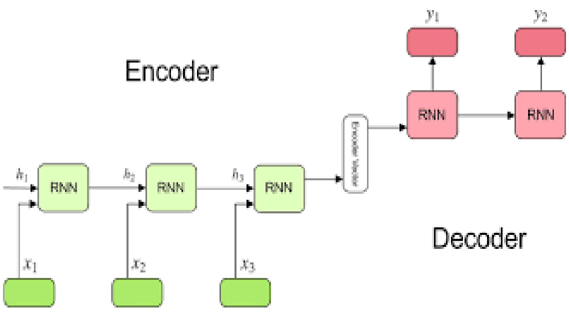


Figure 5: Encoder-Decoder Architecture

BERT [4], is a new language representation model (Bidirectional Encoder Representations from Transformers). Bidirectional architecture are used when a specific word in an English sentence is influenced by the words in both directions. It is a pre-train architecture for un labelled textual data which operates in both directions of input text to extract the context of the sentence. This pre-trained model can be fine-tuned with just one additional output layer for wide range of tasks. The architecture is of BERT is shown in figure 7.

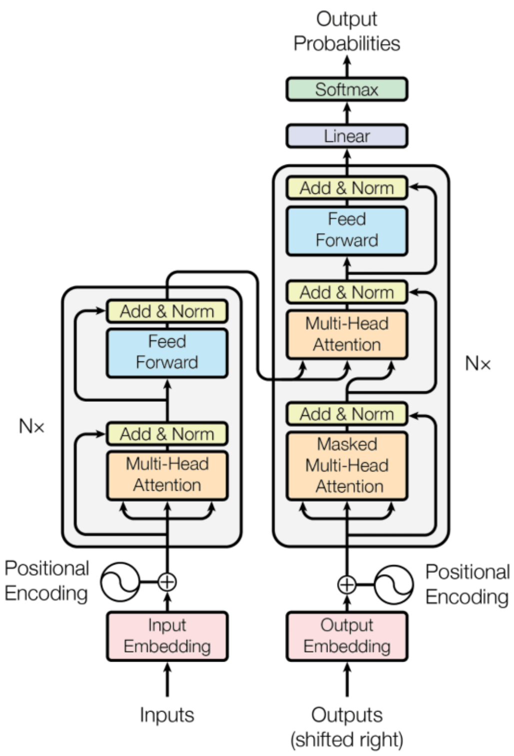


Figure 6: Transformers

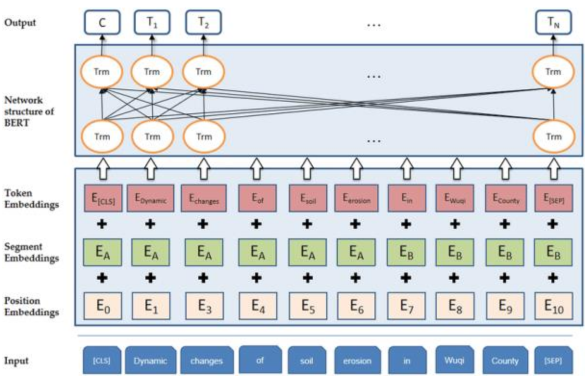


Figure 7: BERT

Several variants of BERT have been developed, each with its unique characteristics and applications. And they are built upon the basic BERT model to increase the performance, reduce computational complexity and to improve the efficiency. Some of the most notable BERT variants include:

- ALBERT [5]: (A Lite BERT) - used to reduce the number of parameters and computational complexity without sacrificing performance.
- RoBERTa: (Robustly Optimized BERT) [6] - used in larger dataset and for a longer period using more advanced training techniques.
- ELECTRA: (Efficiently Learning an Encoder that Classifies Tokens Accurately) [7] - a BERT model used to generate high-quality text representations.
- DistilBERT [8]: A BERT model that has been distilled (or simplified) to be used when computational resources are limited, such as on mobile devices.
- TinyBERT [9]: TinyBERT to be be used in even more resource-constrained situations.
- clinicalBERT [10]: It is constructed on basic BERT model specifically for clinical data analysis. ClinicalBERT is a modified BERT model: Specifically, the representations are learned using medical notes and further processed for downstream clinical tasks.

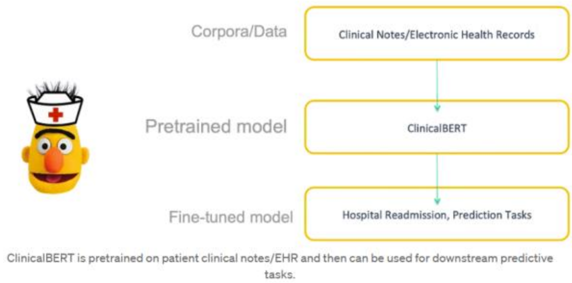


Figure 8: Clinical BERT

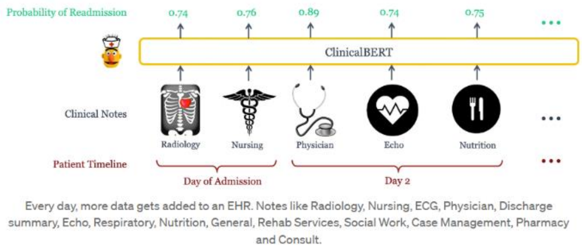


Figure 9: Clinical BERT

ClinicalBERT

In Clinical BERT, the inputs are unstructured, high-dimensional and sparse clinical notes (see the architecture Figure 9).This BERT model can be even used to predict 30-day hospital readmission probability (see figure 8) at various timepoints of admission. General BERT models use Wikipedia and BOOK Corpus as corpus, which is not suited for clinical BERT. Hence, specialized preprocessing is required. The trained parameters of Clinical BERT can also be used to interpret predictions. Empirically, Clinical BERT is an accurate language model and captures physician-assessed semantic relationships in clinical text.

ClinicalBERT [10] used for identifying the association between clinical terms are shown in figure 10. The quality of the input data will be reflected highly on the identified associations and has great influence on the performance of the model. Clinical notes have jargon, abbreviations, with a different syntax and grammar than common language in books or encyclopaedias. Also, clinical notes may contain not-readily accessible information about patients. Most of the information are generated dis-jointly, making the system more complex. And it affects the performance capabilities of each individual solution. Even then, Clinical BERT proves its efficiency to manage the clinical data and it ensures efficient information extraction, enhanced decision-making, seamless data integration, advanced data assessment, automated data extraction, improved data quality and accelerated research.

Figure 10: ClinicalBERT identifies qualitative relationships between clinical terms

Conclusion

ClinicalBERT is a sophistical model, which uses techniques such as bidirectional transformers. ClinicalBERT can capture the high-quality relationships between medical concepts. And it can be used for predicting the readmission chances of a patient within 30 days by analyzing the discharge summaries and the ICU (Intensive Care Unit) data. Statistical analysis of different types of notes creates complexity as different data follows different distributions, which may lead to performance degradation. Hence careful pre-preprocessing is to be designed for specific applications.

References

1. J.T. Connor, R.D. Martin, and L.E. Atlas. Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, 5(2):240–254, 1994.
2. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
4. Sulaiman Aftan and Habib Shah. A survey on bert and its applications. In *2023 20th Learning and Technology Conference (LT)*, pages 161–166, 2023.
5. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.
6. Lan You, Fanyu Han, Jiaheng Peng, Hong Jin, and Christophe Claramunt. Ask-roberta: A pretraining model for aspect-based sentiment classification via sentiment knowledge mining. *Knowledge-Based Systems*, 253:109511, 2022.
7. Galangkangin Gotera, Radityo Eko Prasojo, and Yugo Kartono Isal. Developing a singlish neural language model using electra. In *2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 235–240, 2022.
8. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.
9. Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding, 2020.
10. Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342, 2019

Explainable AI in Clinical Decision Support Systems: An Analysis of SHAP Values

Sujithra Sankar
Department of Computer Applications

Abstract

In the era of advanced artificial intelligence techniques, the development of accurate predictive models in medical domain for complex disease conditions has shown remarkable progress. However, the widespread adoption of these models in practice demands not only predictive performance but also interpretability and transparency. This paper explores the explainability in prediction of machine learning algorithms, focusing on the application of SHAP (SHapley Additive exPlanations) values as a powerful tool for explaining these models. The paper highlights the potential of SHAP values in bridging the gap between the machine learning predictions and the interpretability of this prediction in clinical decision support systems (CDSS). SHAP has the potential to provide valuable insights into the decision-making processes of popular machine learning algorithms.

Index Terms - Explainable AI; SHAP values; Clinical Decision Support System; Prediction; Machine Learning models;

Introduction

Machine learning models have become indispensable tools in a wide range of applications, from healthcare to autonomous vehicles and natural language processing. The integration of advanced machine learning techniques has paved the way for enhanced Clinical Decision Support Systems (CDSS), playing a pivotal role in diagnosis, prognosis, and treatment planning. The challenge of model explainability is an important concern when machine learning systems are deployed, especially in medical domain where transparency, accountability, and trust are most valued. Researchers have been actively developing techniques to enhance the interpretability of machine learning models, enabling users of CDSS ie., clinicians and patients, to understand how and why these models arrive at specific predictions or decisions.

SHAP values are based on game theory and assign an importance value to each feature in a model [1]. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is.

Related Studies

In 2016, Marco Tulio Ribeiro, Sameer Singh, et al. [2] introduced LIME, an innovative method for explaining predictions generated by any classifier in a manner that is both understandable and faithful. This is achieved by constructing an interpretable model locally around the prediction.

In 2022, researchers Du, Y., Rafferty, A.R., McAuliffe, F.M., et al. [3] developed an interpretable clinical decision support system (CDSS) using explainable machine learning to identify women at risk who require targeted pregnancy interventions. SHAP explanations were employed to elucidate the machine learning models, thereby enhancing the credibility and acceptability of the system. Various models were created for diverse use cases and implemented as a publicly accessible web server for academic purposes.

In a separate study conducted in 2023, Mohammad Naiseh et al. [4] conducted an empirical study evaluating four classes of eXplainable Artificial Intelligence (XAI) for their impact on trust calibration, with clinical decision support systems as a case study. The findings are presented as guidelines for designing XAI interfaces.

In 2022, Panigutti et al. [5] demonstrated the applicability of XAI to explain a clinical Decision Support System (DSS) and designed an initial prototype of an explanation user interface. Testing the prototype with healthcare providers and collecting their feedback, the study obtained evidence that explanations increase users' trust in the XAI system, providing valuable insights into perceived deficiencies in healthcare professionals' interaction with the system.

In a study in 2022, Sumayh S. Aljameel [6] created an explainable artificial neural network (EANN) model to distinguish between benign and malignant thyroid nodules while uncovering predictive factors associated with malignancy.

Another study in 2022 by Arjaria, S. K., et al., employed the widely adopted SHAP method grounded in coalition game theory [7] to interpret outcomes. This study sheds light on the system's behavior at both local and global scales, illustrating how machine learning can discern the causality of diseases and assist healthcare professionals in recommending optimal treatments. The research not only presents machine learning algorithm outcomes but also offers insights into feature importance and model explanations.

In a different study in 2022, A. K. M. S. Hosain and M. Golam Rabiul Alam et al. [8] utilized eXplainable Artificial Intelligence (XAI) architectures LIME and SHAP to comprehensively interpret the model's decisions, addressing the 'Black Box' nature of machine learning prediction process.

SHAP Plots

A brief overview of various SHAP plots commonly used for interpretation are listed here.

Summary Plot: Provides a global view of feature importances across all instances in the dataset. It displays each feature's mean SHAP value, indicating its overall impact on model predictions.

Force Plot: Illustrates the SHAP values for a single prediction, showcasing how each feature contributes to the final prediction. It helps in understanding the direction and magnitude of each feature's impact on the prediction.

Dependency Plot: Shows the relationship between the value of a specific feature and the corresponding SHAP values. It is useful for understanding how a feature's value influences the model's prediction.

Waterfall Plot: Represents the breakdown of a prediction's output into individual feature contributions. It is useful for visualising the sequential addition of SHAP values, creating a waterfall-like structure.

Decision Plot: Maps cumulative SHAP values for each prediction. It shows how strongly individual features contribute to a single prediction, aiding in understanding feature values influencing the prediction.

These SHAP plots collectively offer interpretation of machine learning models, giving valuable insights into model behaviour, feature importance, and the decision-making process.

Making sense of SHAP Values

Sign and magnitude of the SHAP values for each feature indicates the impact of the presence or absence of that feature on the model's prediction. The sign of the SHAP values provide information about the direction of the impact (positive or negative), and the magnitude indicates the strength of that impact.

Conclusion

SHAP plots are a powerful tool for understanding feature relationships and their impact on model predictions. The integration of SHAP values into CDSS augments predictive accuracy, and also ensures a level of transparency necessary for the adoption of these systems in clinical practice. By bridging the gap between the complex nature of machine learning models and the interpretability required in healthcare decision-making, SHAP values contribute significantly to the evolution of CDSS, making them more accessible, reliable, and, more effective in improving patient care.

References

1. SHAP Documentations. Available online: https://shap.readthedocs.io/en/latest/index.html.
2. Marco Tulio Ribeiro, Sameer Singh et al. Why Should I Trust You? Explaining the Predictions of Any Classifier, arXiv:1602.04938v1 [cs.LG] 16 Feb 2016
3. Du, Y., Rafferty, A.R., McAuliffe, F.M. et al. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. Sci Rep 12, 1170 (2022). https://doi.org/10.1038/s41598-022-05112-2
4. J. Mohammad Naiseh, Dena Al-Thani, Nan Jiang, Raian Ali, How the different explanation classes impact trust calibration: The case of clinical decision support systems,International Journal of Human-Computer Studies, Volume 169, 2023, 102941, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2022.102941
5. Panigutti, Cecilia et al., Co-Design of Human-Centered, Explainable AI for Clinical Decision Support, Association for Computing Machinery 2023, issn :2160-6455, https://doi.org/10.1145/3587271{10.1145/3587271}.
6. Aljameel S.S., A Proactive Explainable Artificial Neural Network Model for the Early Diagnosis of Thyroid Cancer. Computation; 10(10):183. https://doi.org/10.3390/computation10100183, 2022.
7. Arjaria, S. K., Rathore, A. S., & Chaubey, G. Developing an Explainable Machine Learning-Based Thyroid Disease Prediction Model. International Journal of Business Analytics (IJBAN), 9(3), 1-18. http://doi.org/10.4018/IJBAN.292058. 2022
8. A. K. M. S. Hosain and M. Golam Rabiul Alam, Demystifying Hypothyroidism Detection with Extreme Gradient Boosting and Explainable AI. 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 260-265, doi: 10.1109/ICCIT57492.2022.10055791. 2022.

Fragility Analysis in Earthquake Engineering: OpenSees as a tool for Comprehensive Structural Vulnerability Assessment

Daniel A V
Department of Civil Engineering

Abstract

Fragility analysis plays a pivotal role in earthquake engineering by quantifying the vulnerability of structures to seismic events. This article delves into the concept of fragility analysis and explores the capabilities of OpenSees, a finite element analysis software package, in conducting comprehensive and accurate structural assessments. The application of opensees in conducting the seismic vulnerability assessment for various type of structures highlights the versatility of the tool.

Index Terms - Fragility analysis, seismic vulnerability, OpenSees, structural assessment, earthquake engineering.

Introduction

In earthquake-prone regions, seismic vulnerability of proposed and existing structures is crucial in planning of urban development, retrofitting of existing structures and also in disaster management at the occurrence of seismic events. Fragility analysis serves as a quantitative approach to assess the likelihood of structural damage or failure under varying seismic intensities. The need of the hour is to have methodologies to assess the risk parameters and tools that can give reliable results at maximum possible reduced computational cost.

Fragility Analysis: Theoretical Foundation

Fragility analysis involves the development of fragility curves, which depict the probability of exceeding a specific damage state as a function of seismic intensity. Fragility curves represent log normal distribution between the conditional probability of exceeding predefined damage states (DS) for given levels of EDP (engineering demand parameters) and are used to estimate the amount of damage for a particular level of shaking defined with Intensity Measures (IM). The key EDPs commonly used in fragility analysis are

- Displacement-based EDPs: Roof displacements, Inter-storey drift
- Force-based EDPs: Base Shear, storey shear
- Energy-based EDPs: Hysteretic Energy

Widely used IMs for quantifying the severity of ground shaking during a seismic event are

- Peak Ground Acceleration (PGA): PGA is widely used as an intensity measure, especially in regions where ground acceleration is a dominant factor in structural response.
- Spectral Acceleration (Sa): Sa is valuable for capturing the frequency content of ground motion, aiding in assessing the resonance characteristics of structures.
- Peak Ground Velocity (PGV)
- Peak Ground Displacement (PGD)
- Duration

Most generalized mathematical expression of a fragility curve is as follows

$$Pf(\alpha) = P(EDM > DS)|\, IM = \alpha$$

Lognormal distribution function is the most suitable for representing fragility curve. The structural capacity is assumed to be lognormally distributed with median A_m and lognormal standard distribution (log-std) β. It allows the possibility of formulating the fragility curves as Cumulative Distribution Function (CDF) [1].

$$Pf(\alpha) = \phi(\frac{ln\alpha - lnA_m}{\beta})$$

Where $\phi(u)$ designs the CDF of a normalized Gaussian random variable u.

There are several methods for creating fragility curve of a structure, the one in focus of this study is Incremental Dynamic Analysis (IDA)

Incremental dynamic analysis (IDA)

This is based on numerical simulation by applying a set of accelerograms scaled to have same IM. For a given set of accelerogram, N non-linear time history analysis are performed for each intensity level. The set of N accelerograms are scaled to increasing intensity levels, until failure is reached. Using this method, each accelerogram has a single capacity value associated with onset of collapse. Collapse is linked to either the load intensity where response becomes a flat line or when the time history analysis does not converge and dynamic instability of the structure occurs [1]. That intensity level is considered as the collapse intensity level.To conduct N number of non–linear time history analysis for different intensity level of seismic excitation, an efficient tool is required for the purpose.

OpenSees: A Tool for Fragility Analysis

The Open System for Earthquake Engineering Simulation (OpenSees) has a history rooted in the need for a flexible and open-source platform to facilitate earthquake engineering research and analysis. Creating fragility curves using OpenSees involves simulating the response of structures to a range of ground motion intensities and then analyzing the resulting structural damage. The procedure to follow can be summarised as follows

- Define the Structural Model: Start by defining the structural model in OpenSees. This includes specifying the geometry, material properties, and element types that represent the structure.

- Select Ground Motion Records: Choose a set of ground motion records that represent the seismic hazard at the site.
- Establish Damage States: Define the damage states that you want to assess in your fragility analysis.
- Iterative Seismic analysis:

  a) Apply the ground motion to the structural model using OpenSees dynamic analysis commands.
  b) Extract the engineering demand parameters (EDPs) from the analysis results.
  c) Compare the extracted EDPs to predefined thresholds associated with each damage state
  d) Determine the damage state the structure experiences for the given ground motion.

- Analyze Results and Calculate Probabilities: After simulating the response for all ground motions, calculate the probability of exceeding each damage state at different ground motion intensities.
- Plot Fragility Curves: Using the probabilities and ground motion intensities, plot fragility curves.

Applications of Fragility Analysis with OpenSees

Several studies are available in literature where fragility curves are plotted and assessed using the tool Opensees. A study conducted by (Davide Forcellini, 2021), fragility curve of shallow founded structures under the influence of Soil-Structure Interaction (SSI) was done making using of the Opensees [2]. The non-linearity involved in the problem and the advanced analysis procedures forming the capability of the software interested the researcher to choose it. The study concluded the adverse effect of neglecting the SSI which leads to prediction of lower probability of failures than realistic magnitudes.

Studies on large varieties of structures [3] other than buildings had been performed using the help of finite element analysis software package Opensees. Fragility analysis of bridges [4] forms the majority.

Conclusion

The iterative seismic analysis methodology, guided by OpenSees, ensures a systematic approach to assess structural vulnerabilities under diverse seismic intensities. Applications in literature, such as studies on soil-structure interaction and bridges, showcase OpenSees' adaptability. As a result, the synergy between fragility analysis and OpenSees emerges as a cornerstone in advancing seismic risk assessment and resilient structural design practices.

References:

1. Irmela Zentner, Max Gündel, Nicolas Bonfils, Fragility analysis methods: Review of existing approaches and application, Nuclear Engineering and Design, Volume 323, 2017, Pages 245-258, ISSN 0029-5493
2. Davide Forcellini, Analytical fragility curves of shallow-founded structures subjected to Soil-Structure Interaction (SSI) effects, Soil Dynamics and Earthquake Engineering, Volume 141, 2021, 106487, ISSN 0267-7261
3. Liu, C. & Li, H. & Li, D. & Li, L.. (2017). Seismic fragility of SRC structures based on OpenSEES and IDA. World Information on Earthquake Engineering. 33. 112-121.
4. Ngoc-Long Tran, Trong-Ha Nguyen, Van-Tien Phan, Duy-Duan Nguyen, Seismic fragility analysis of reinforced concrete piers of steel box girder bridges: A parametric study, Materials Today: Proceedings, Volume 38, Part 5, 2021, Pages 2310-2315, ISSN 2214-785

Harmony Search Optimization of Sliding Mode Control for DC Servo

Jim George
Department of Electrical and Electronics Engineering

Abstract

Harmony Search (HS) is a metaheuristic optimization algorithm inspired by the musical improvisation process of searching for a perfect state of harmony. It was introduced by Zong Woo Geem in 2001 and has been applied to various optimization problems. The algorithm is particularly useful for solving complex optimization problems where the search space is large and the objective function is non-linear, non-continuous, or lacks derivatives. In this article position control of dc servo with control parameter optimization via Harmony Search optimization is discussed. The results with a perspective sliding mode control and phase trajectory optimization is presented.

Index Terms — DC Servo, Sliding Mode Control, Harmony Search optimization

I. Introduction

DC Motor position control of is vital in applications for accuracy control framework. Electric rotating actuators are utilized in the electric power industry, high-power changing gears and bundling applications. Position control framework is a closed loop control framework whose result is the ideal precise position of the DC servo. The servo whose position is to be controlled is associated in a closed loop framework.

Optimization is one of the pivotal elements in control design. Optimization of control parameters is performed to match the performance index. Conventional performance indices considered are integral square error (ISE), integral absolute error (IAE), integral time absolute error (ITAE). This paper introduces "optimal course by length" using Subterranean Insect State Streamlining clubbed with Travelling Salesman Problem (TSP) which deal in common with the aforementioned performance indices. That is "optimum course by length" can be considered as a performance index in lieu of ISE, ITAE or ISE.

Sliding Mode Control is a control technique that creates a surface toward which a system asymptotically moves to zero. Filipov and Utkin pioneered SMC in the 1960s, also known as variable structure control due to its switching nature. SMC modifications, such as Twisting SMC, Super Twisting SMC, and Terminal SMC, enhance the performance and robustness by requiring trajectories to slide around a limited ρ-c subspace with infinite gain and can handle unknown disturbances using excitation-response frameworks, discontinuous models, and probabilistic frameworks. The paper focuses on applying ρ-c Control as a nonlinear state feedback control method to control the position and angular velocity of the system. Inherent consideration is given to the influence of both.

The organisation of this paper is as follows. Section 2 provides a brief schematic of servo system. Section 3 give specifications of dc servomotor. Section 4 accounts for ρ-c Control formulation. Section 5 outlines Hardware in Loop implementation of SMC to the servo system. In section 6, determination of optimal periodic phase trajectory is performed with intelligent Subterranean Insect State Streamlining technique. Section 7 highlights an outline of results and its interpretation. Section 8

II. General control structure of DC Servo

As unveiled in Figure 1, Position control of a servo system requires sensing of angular velocity and angular position. The sensors are encoders and tachometer. To apply control part, a data acquisition card (DAQ card) is embedded in the servo mechanism. The DC motor used foe servo control is having a rated voltage of 18V and drive used is direct- drive. The dc servo is powered from PWM amplifier with current sensor integrated. Overall schematic of real time servo system is shown in Figure 1. The data acquisition card working in the mode of controlling servomotor has encoder input and analog signal output as power command. Current sensing command is fed as analog input to data card. Hene velocity and position control is performed by current control method. The control command to the DAQ device is performed by the track of USB. The control algorithms for position and velocity control is directed by MATLAB. The MATLAB writes required logic to the DAQ card which in turn takes action for desired control. The aim of this research is to implement sliding mode controller and to determine optimal control parameters applying optimization method having intelligence factor.

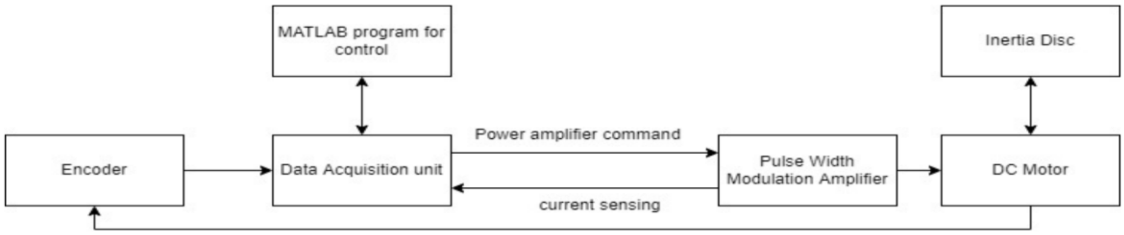


Fig. 1. General control structure

A. Sliding mode Control

SMC is a type of indirect method which is performed in two steps. In the pilot step, that is reaching phase, the system is guided on a path to follow a certain plane or control surface or sliding surface. Once the system glides near to this surface, then by a switching (sign) function, the system is guided to follow the sliding surface. This in turn is fosters the stability of the system as in the transient part, a rigorous control is there and the steady state part a surface control is there. That is in a bird's eye perspective, system is getting locked to the sliding surface even though control effort diminishes. The main draw of SMCis it defies with zero error, but make the system stable with minimal error. SMCis applicable only when a bounded or limited disturbance applied to the system, In the event of high disturbance, a parallel track needs to be provided to stabilize the system during the transient state. Mathematical expressions foe design of a SMC is highlighted in the next section. The key advantage of SMC is by varying two parameters, control of the system can be performed.

B. Mathematics of SMC

ρ-c control law is highlighted in [16]. For dc servo Degree of Freedom are considered to be angular position θ and angular velocity ω.
Considering limited disturbance χ

$$|f(\theta, \omega, t)| \leq \chi > 0 \quad (1)$$

Considering state space model of the system

$$\frac{d\theta}{dt} = \omega \quad (2)$$

$$\frac{d\omega}{dt} = u + f(\theta, \omega, t) \quad (3)$$

For zero convergence, control rule is

$$u = K_1\theta + K_2\omega \text{ ; } K_1 < 0, K_2 < 0 \quad (4)$$

Dynamics with required compensation is given by

$$\frac{d\theta}{dt} + c\theta = 0; c > 0 \quad (5)$$

For the convergence in asymptotic mode,

$$\theta(t) = \theta(0)e^{-ct}$$

$$\omega(t) = -c\theta(0)e^{-ct} \quad (6)$$

$$\zeta = \zeta(\theta, \omega) = \omega + c\theta; c > 0 \quad (7)$$

The input of control u is given as

$$\frac{d\zeta}{dt} = c\omega + f(\theta, \omega, t) + u$$

$$\zeta(0) = \zeta_0 \quad (8)$$

Lyapunov function is selected as

$$V = \frac{1}{2}\zeta^2 \quad (9)$$

The singular point is ζ=0

Stability constraints follows.

$$V = \infty$$

$$\frac{dV}{dt} < 0; \zeta \neq 0 \quad (10)$$

Function V should be negative definite is an essential constraint for stability.
Lyapunov function derivative is highlighted as,

$$\frac{dV}{dt} = \zeta\frac{d\zeta}{dt} = \zeta(c\omega + f(\theta, \omega, t) + u) \quad (11)$$

$$u = -c\omega + v$$

$$v = -\rho * sgn(\zeta) \quad (12)$$

Control law for SMC is

$$u = -c\omega - \rho sgn(\zeta) \quad (13)$$

C. Formulation of sliding surface

Mathematical expression of sliding surface is given in equation (7) by considering equilibrium point ζ=0. c is the factor which controls slope of the sliding surface. Chattering effect is based on the value of c. Figure 2 shows sliding surfaces for variation of c.

III. Hardware in loop implementation of sliding mode controller

Sliding Mode Controller is implemented in dc motor with encoder applying Hardware in Loop Model. The block diagram for implementation is shown in Figure 3 and Figure 4. Since the control is applied to hardware model, non-linearity arises. This is due to friction, backlash effect, torsion and presence of nonlinear components in the driver circuit. Hence a nonlinear ρ-c controller is designed for the system. The ρ-c controller is developed based on mathematical expressions. The input to the ρ-c control are angular position and angular velocity. SMC is acting as a state feedback controller with control law given in Equation (13). Here ζ incorporates the states of the system and control gain ρ and c are the parameters that needs to be varied to obtain optimum performance of the system.

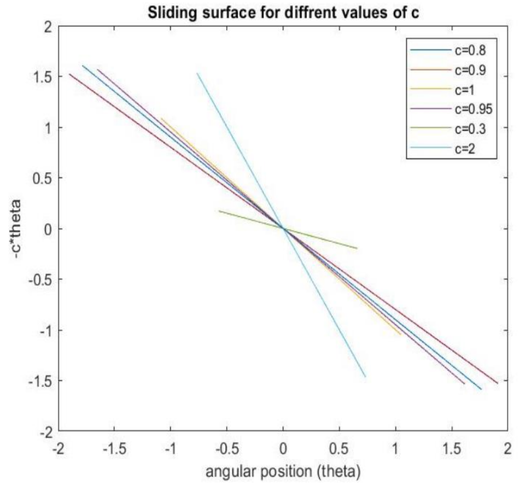


Fig. 2. Sliding surfaces for different values of c

Fig. 3. Hardware in Loop implementation- block diagram

Here optimum performance is considered as minimal integral square error (ISE). Optimal ISE is backed up optimal course length on application of optimization algorithm. The excitation applied to the system is given by Equation (14).

$$v(t) = sgn(\sin \sin \frac{2\pi t}{10}) \qquad (14)$$

The peculiarity of excitation signal gives initial condition as (0,0) and two final states (1,0) and (-1,0). (In radians and by conversion factor final state (1,0) means rotating clockwise by and angle π/4 radians and (-1,0) means rotating counter clockwise by an angle of π/4 radians). The characteristics of phase trajectory shows that the system is trying to settle around final states, but within a time span of 5s excitation varies. Hence the system is toggling between final states. Also, the phase trajectory is in a stable node mode on application of controller.

IV. Harmony search optimization

The phase variables considered are θ and ω. The phase trajectory for the application of the ANFIS-based Sliding mode control is shown below. To determine the optimal phase trajectory optimization is performed. The unique nature of the excitation signal establishes the initial condition as (0,0), with two distinct final states: (1,0) and (-1,0). In terms of radians and using a conversion factor, the final state (1,0) indicates a clockwise rotation of π/4 radians, whereas the final state (-1,0) signifies a counterclockwise rotation of -π/4 radians. Analysis of the phase trajectory reveals that the system endeavors to settle around these final states. However, within a time span of 5 seconds, the excitation undergoes variation. Consequently, the system toggles the two final states. In addition, the phase trajectory exhibited a stable node mode when the controller was applied. The phase trajectories obtained in this study exhibited periodic behavior, as they are influenced by the time-dependent variables of the angular position (θ) and angular velocity (ω). As a result, an analysis based on equilibrium points becomes irrelevant. Moreover, the stability of the phase trajectory is indicated by its closed nature. As the phase trajectory approaches the sliding surface, the phenomenon of chattering occurs which is apparent. Harmony Search (HS) is a metaheuristic optimization algorithm that draws inspiration from the creative process of musicians in the jazz band. The objective is to discover optimal solutions by emulating improvisation and harmonious collaboration among musicians. HS is applicable to a wide array of optimization problems that are capable of accommodating both continuous and discrete variables.

The fundamental steps of the Harmony Search algorithm are as follows:

Step 1: Initialization Generate an initial collection of potential solutions called "harmonies" within the predefined bounds of the decision variables of the problem.

Step 2: Harmony Evaluation Assesses the objective function for each harmony. The objective function encapsulates the optimization criteria specific to the problem being solved. Fitness function: The fitness function is given as in equation (21).

$$fitnessfunction = @(length)sum((\theta * length - \omega)^2) \quad (15)$$

- This fitness function determines best-fitness (length) that minimizes the sum of squared differences between the model predictions and the actual data.
- (θ * *length* - ω) computes the vector of differences between the values obtained by multiplying each element of θ with length and the corresponding elements in ω.
- The sum of squares term in the objective function calculates the sum of squared differences, which represents the fitness function value for a given length.

In the context of the objective function, the length represents the magnitude or value of the objective function output when evaluated with a specific set of decision variables. The algorithm uses this length as a measure of the closeness of the current harmony to the optimal solution. During the optimization process, the HSO iteratively updates and evaluates different harmonies, seeking to converge towards harmony with the shortest (in case of minimization) or longest (in case of maximization) length, which corresponds to the optimal solution for the given problem.

In summary, the "length" in the objective function of Harmony Search Optimization refers to the value of the objective function for a specific harmony, serving as a measure of the harmony's quality and guiding the search for the optimal solution.

Step 3: Harmony Memory Update Selects new harmonies by combining elements from the existing harmony memory. This process mirrors musicians building upon previous melodies during improvisation. Selection strategies can include randomness, probability-based selection, or ranking based on fitness values.

Step 4: Improvisation Generates novel harmonies by perturbing selected harmonies. This involves altering specific components while adhering to the harmony rules. For instance, adjusting the values of the decision variables within predetermined ranges or applying specialized operators for discrete variables.

Step 5: Harmony Evaluation Evaluates the objective function for newly generated harmonies.

Step 6: Harmony Memory Update Update the harmony memory by replacing certain existing harmonies with newly generated ones based on their fitness values. The aim is to enhance the quality of the harmonies stored in memory over subsequent iterations.

Step 7: Termination Criteria Repeat steps 4-6 until the termination condition is met. The termination condition can be defined by the maximum number of iterations, achievement of a desired convergence level, or any other condition tailored to the specific problem.

Step 8: Solution Extraction Once algorithm concludes, extract the best solution from the harmony memory based on its fitness value. This solution represents the optimized solution for the problem at hand.

A. Phase trajectory

The phase variables considered are θ and ω. The phase trajectory for the application of the ANFIS-based Sliding mode control is shown below. To determine the optimal phase trajectory optimization is performed. The unique nature of the excitation signal establishes the initial condition as (0,0), with two distinct final states: (1,0) and (-1,0). In terms of radians and using a conversion factor, the final state (1,0) indicates a clockwise rotation of π/4 radians, whereas the final state (-1,0) signifies a counterclockwise rotation of -π/4 radians. Analysis of the phase trajectory reveals that the system endeavors to settle around these final states. However, within a time span of 5 seconds, the excitation undergoes variation. Consequently, the system toggles the two final states. In addition, the phase trajectory exhibited a stable node mode when the controller was applied. The phase trajectories obtained in this study exhibited periodic behavior, as they are influenced by the time-dependent variables of the angular position (θ) and angular velocity (ω). As a result, an analysis based on equilibrium points becomes irrelevant. Moreover, the stability of the phase trajectory is indicated by its closed nature. As the phase trajectory approaches the sliding surface, the phenomenon of chattering occurs which is apparent. The periodic phase trajectories for different values of θ and ω are plotted in Figures 4,5,6,7.

The concept of a periodic phase trajectory is a fundamental element in the analysis of dynamic systems. It refers to the path followed by a system over time when it displays periodic characteristics. As the system progresses, it traces a trajectory in the phase space, where each point on the trajectory corresponds to a specific state of the system at a particular time. In instances of periodic behavior, the system repeats its states after a defined time interval, meaning it returns to previous states regularly. The recurring evolution of the system characterizes its periodic behavior. The trajectory may manifest as a closed loop or a series of closed loops, reflecting that the system revisits certain states repeatedly during its evolution. It provides valuable insights into the stability, periodicity, and overall dynamics of various physical phenomena, from simple harmonic oscillators to celestial planetary motion. By analyzing periodic phase trajectories, researchers can gain a deeper understanding of the underlying principles governing the systems and make predictions about their future behavior. The optimal phase trajectory was determined by the assessing best fitness based on the Harmony Search Optimization technique. This in turn provides the performance index which accounts for the low value of the Standard deviation and Standard Error.

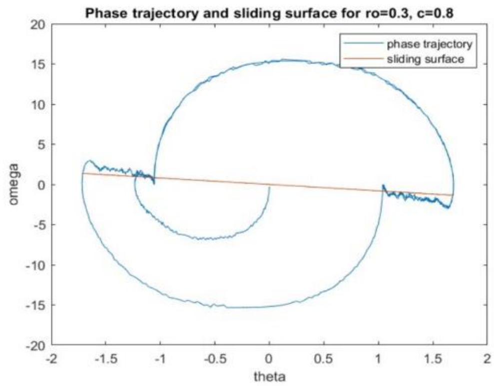


Fig. 4. Periodic Phase trajectory and sliding surface for ρ=0.3 and c=0.8

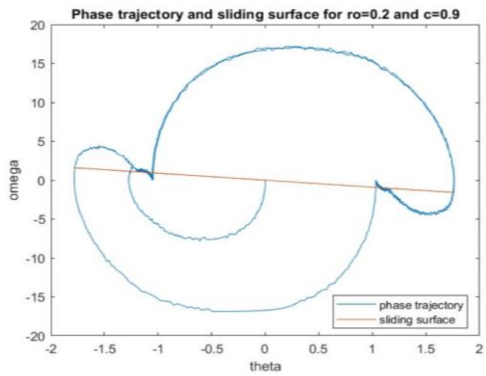


Fig. 5. Periodic Phase trajectory and sliding surface for ρ=0.2 and c=0.9

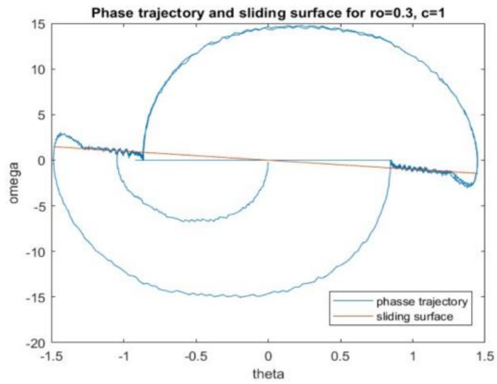


Fig. 6. Periodic Phase trajectory and sliding surface for ρ=0.3 and c=1

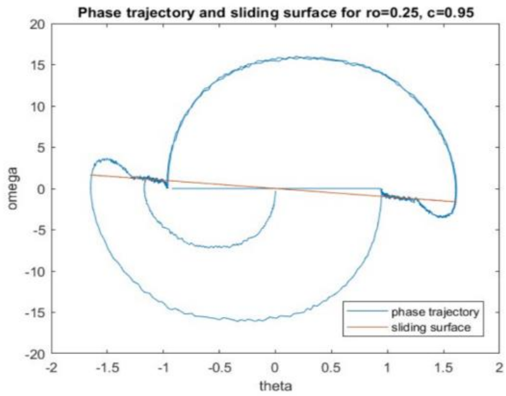


Fig. 7. Periodic Phase trajectory and sliding surface for ρ=0.25 and c=0.95

B. Determination of optimal phase trajectory

The parameters of Harmony Search optimization are listed in Table 1. The significance of choosing an appropriate value for Harmony Memory Consideration rate (hmCR) and Harmony Memory Bandwidth (hmBW) is:

1)The HMCR determines the probability of selecting a value from the Harmony Memory (a repository of previously generated solutions) during the search process. A higher HMCR value increases the likelihood of selecting values from the Harmony Memory, thereby promoting the exploitation of promising solutions. However, lower HMCR value encourages the exploration of new solutions. The appropriate choice of HMCR depends on the characteristics of the optimization problem, balancing the exploration and exploitation trade-off. 2) The HMBW determines the range within which new values are generated based on the values selected from the Harmony Memory. It controls the diversity of solutions generated during the search process. A wider bandwidth allows for a broader exploration of the solution space, potentially finding global optima. Conversely, a narrower bandwidth narrows the search for promising solutions, focusing on the local optima. The selection of an appropriate HMBW value depends on the landscape of the problem and the balance between exploration and exploitation. The Harmony Search algorithm can be

TABLE I
HS PARAMETERS

HMS	Iterations	hmCR	hmBW	Cost function
10	500	0.8	0.4	Standard Error

applied to optimization problems with fitness values that are either maximized or minimized, depending on the nature of the problem. In maximization problems, the algorithm aims to achieve higher accuracy or values for the fitness function by exploring and exploiting candidate solutions. However, from the perspective of the phase trajectory, the Harmony Search Approach is considered a minimization problem, aiming to minimize the length of the trajectory representing the evolution of a system over time. The objective is to find the optimal set of parameters for the most efficient trajectory.

For periodic phase trajectories, the algorithm focuses on toggling equilibrium points to identify parameter configurations leading to a periodic trajectory with minimal error. The fitness function is crucial in Harmony Search Optimization (HSO) for evaluating the quality of harmonies (candidate solutions) and guiding the search process. It serves multiple key roles, including objectively assessing harmony performance, guiding the search direction, determining convergence and termination criteria, evaluating solutions efficiently, providing insight into the problem landscape, balancing exploration and exploitation, and tuning algorithm parameters for enhanced performance and adaptability to specific problems. The fitness function's numerical measure helps quantify solution quality and steer the algorithm toward optimal outcomes while addressing challenges such as local optima and parameter tuning.

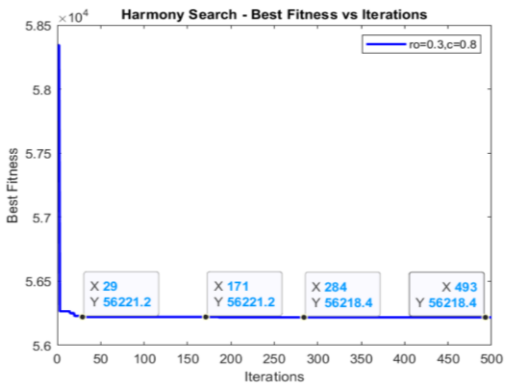


Fig. 8. Harmony Search Best Fitness for ρ=0.3 ,c=0.8

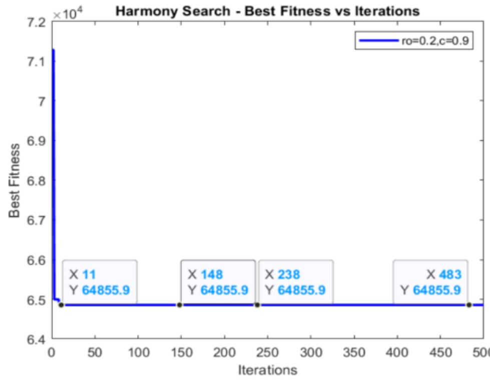


Fig. 9. Harmony Search Best Fitness for ρ=0.2 ,c=0.9

Fig. 10. Harmony Search Best Fitness for ρ=0.3 ,c=1

Fig. 11. Harmony Search Best Fitness for ρ=0.25 ,c=0.95

V. Results and discussions

In essence, the fitness function serves as the driving force behind Harmony Search Optimization, allowing for the evaluation of candidate solutions, guiding the search for optimal solutions, and determining the conditions for termination. As a critical component, it profoundly influences the behavior and effectiveness of the algorithm in addressing complex optimization problems. Tables 2 and 3 show the standard error and standard deviation with respect to the actual and predicted values of the sliding mode and the best fitness value for the same. Table 3 highlights the variation of ISE for different values of control parameters. This finding provides strong evidence to support the claim that the optimal value with the least error for sliding mode control of DC servo systems for position control is achieved when the parameters ρ and c take on the specific values of 0.2 and 0.9 respectively. The Harmony Search Optimization algorithm demonstrated its effectiveness in determining the optimal values for the control parameters. The minimization of standard error and Integral Square error points ρ=0.2 and c=0.9 as the best fit for achieving superior performance in the DC servo position control system.

TABLE II
STANDARD DEVIATION (σ) FOR ρ AND C AND BEST FITNESS

ρ	c	$\sigma\rho$	σc	Best fitness
0.3	0.8	0.177798957	0.065466939	56218.4
0.2	0.9	0.092014492	0.060098391	64855.9
0.5	1	0.202813872	0.093743622	24836.9
0.3	1	0.160961233	0.065565235	45297.8
0.25	0.95	0.056561808	0.034989587	54895.8
0.8	0.3	0.142283932	0.562288408	5471.84
0.9	1	0.023370341	0.158486592	458.17
0.5	2	0.231827368	0.210403524	17356.8

TABLE III
INTEGRAL SQUARE ERROR FOR VARIATION IN CONTROL PARAMETERS

ρ	c	ISE
0.3	0.8	1.439
0.2	0.9	1.407
0.5	1	2.848
0.3	1	1.409
0.25	0.95	1.410
0.8	0.3	6.523
0.9	1	15.25
0.5	2	6.67

VI. Conclusions

This study addresses the precise position control of a real-time dc-servo system. The dc-servo is excited by the periodic reference signal. The response and its derivative form the periodic phase trajectory. The optimal periodic phase trajectory is determined on the basis of the optimal values of the sliding mode control derived through HSO. These techniques supplement each other for the same value of sliding mode control parameters.

In this investigation, the sliding mode control concept, reinforced by ANFIS, is applied to the feedback loop, with the novelty of the state feedback control. By comparing this novel architecture with traditional sliding mode control in the feedback loop, the study demonstrated superior results, highlighting the effectiveness of the proposed approach.

To ensure the optimality of the obtained control parameters, this study employed a meta-heuristic Harmony Search Algorithm. The primary goal of this algorithm is to maximize the fitness value, aligning it with the objective of minimizing the standard error and standard deviation associated with the control parameters. This alignment of the maximum fitness value with minimal error and deviation serves as a crucial performance index, providing valuable insights along with other standard indices to evaluate the system's effectiveness.

In conclusion, SMC contributes to robustness by maintaining the system on the sliding surface despite uncertainties and disturbances. ANFIS adapts to varying system dynamics, enhancing robustness in the face of changing conditions.HSO optimizes the parameters of the combined SMC-ANFIS controller, fine-tuning the system for better response and minimizing error.The combination leverages the strengths of each technique, leading to a controller that is both robust and adaptable, capable of providing optimal performance under different conditions. In this study, ANFIS was carefully fine-tuned to achieve optimal parameters for sliding mode control, and the validation process successfully confirmed the efficiency of the proposed approach. The integration of the periodic phase trajectory concept as a combination of sliding mode control with ANFIS, and utilization of meta-heuristic optimization collectively contribute to a robust and highly effective position control strategy for the real-time DC servo system. By validating the optimality of the control parameters and considering various performance indices, this study substantiates the superiority of the proposed method inachieving precise and reliable position control for a real-time DC servo system.

References

1. Wang X, Wang W, Li L, Shi J, Xie B, 2019. Adaptive control of DC motor servo system with application to vehicle active steering. IEEE/ASME Transactions on Mechatronics, 24(3), pp.1054-1063.
2. Wang J S and Yang G H , 2019. Data-driven output-feedback fault-tolerant tracking control method and its application to a DC servo system. IEEE/ASME Transactions on Mechatronics, 24(3), pp.1186- 1196.
3. Masoud A A, Abu-Ali M, Al-Shaikhi A 2019. Experimental determination of an extended DC servo-motor state space model: an undergraduate experiment. IEEE Access, 8, pp.4908-4923.
4. Kommuri S K, Rath J J, Veluvolu KC, 2017. Sliding-mode-based observer–controller structure for fault-resilient control in DC servomotors. IEEE Transactions on Industrial Electronics, 65(1), pp.918-929.
5. Ratanavilisagul C, 2017, June. Modified Ant Colony Optimization with pheromone mutation for travelling salesman problem. In 2017 14th International Conference on Electrical Engineering/Electronics, Computer,

Telecommunications and Information Technology (ECTI-CON) (pp. 411- 414. IEEE.

6. Ellili W, Samet M, Kachouri A, 2017, February. Traveling salesman problem of optimization based on genetic algorithms. In 2017 International Conference on Smart, Monitored and Controlled Cities (SM2C) (pp. 123-127). IEEE.
7. Agrawal P, Abutarboush H F, Ganesh T, Mohamed A W, 2021. Meta-heuristic algorithms on feature selection: A survey of one decade of research (2009-2019). IEEE Access, 9, pp.26766-26791.
8. Akram I, Awad O A, 2021, March. System identification based PSO approach for networked DC servo motor. In IOP Conference Series: Materials Science and Engineering(Vol. 1090, No. 1, p. 012059). IOP Publishing.
9. Kurode S, Spurgeon S K, Bandyopadhyay B, Gandhi P S, 2012. Sliding Mode control for slosh-free motion using a nonlinear sliding surface. IEEE/ASME Transactions on Mechatronics, 18(2), pp.714-724.
10. Thakar P S, Bandyopadhyay B, Gandhi P S, 2014, June. ρ-c control for a class of underactuated systems using feedforward normal form: A slosh-container system. In 2014 13th International Workshop on Variable Structure Systems (VSS) (pp. 1-6). IEEE.
11. Zaihidee F M, Mekhilef S, Mubin M, 2019. Application of fractional order Sliding Mode control for speed control of permanent magnet synchronous motor. IEEE Access, 7, pp.101765-101774.
12. Min-Xiu Y, Yuan-Wei J, You-Guo H, Ping S., 2009, June. Adaptive ρ-c controller for a class of second-order underactuated systems. In 2009 Chinese Control and Decision Conference (pp. 2782-2786). IEEE.
13. Pyrhonen V P, Koivisto H J, Vilkko M K, 2017, December. A reduced-order two-degree-of-freedom composite nonlinear feedback control for a rotary DC servo motor. In 2017 IEEE 56th Annual Conference on Decision and Control (CDC) (pp. 2065-2071). IEEE.
14. Wang S, Tao L, Chen Q, Na J, Ren X, 2020. USDE-based ρ-c control for servo mechanisms with unknown system dynamics. IEEE/ASME Transactions on Mechatronics, 25(2), pp.1056-1066.
15. Hou H, Yu X, Xu L, Rsetam K, Cao Z, 2019. Finite-time continuous terminal sliding mode control of servo motor systems. IEEE Transactions on Industrial Electronics, 67(7), pp.5647-5656.
16. George J, Mani G, Alexander Stonier A, 2022. An extensive critique of Sliding Mode control and adaptive neuro-fuzzy inference system for nonlinear system. Asian Journal of Control, 24(5), pp.2548-2564.
17. Jim George, Geetha Mani, 2023. HiL Implementation of Harmony Search based Redesigned PI-like Control for DC Servo. Mecharronics Systems and Control, Vol.51, No.1, 2023, pp.1-10

Human Emotions & Conversational Technology

N Dharani Devi
Department of Artificial Intelligence & Data Science

Abstract

Conversational technology have increased chatbot incorporation in many fields. Chatbots are natural language dialogue systems that communicate with humans via text, voice, or multimodal embodied agents. Organizations like chatbots because they save money and help customers immediately. Automation of delivery tracking, reservations, airline information inquiries, and orders is common. Corporations like their 24/7 availability and quick response to general problems. Medical and personal chatbots are increasingly providing social and emotional support.

At the same time, humans are conditioned to think that anything that can speak is human and unprepared to deal with machines that mimic human emotions and opinions. These technologies can affect and isolate businesspeople by assuming social roles they normally play. These linguistic techniques manipulate teens, elderly individuals, and mentally ill persons more. They can control users and reaffirm negative beliefs, causing self-harm and injury to others. The Article focusses on adversarial effect of using AI Chat-bots in Human emotion and behaviour.

Index Terms – Human Emotions, Behaviour, Conversational Technology

Introduction

Artificial intelligence systems that converse with users via text, message, or speech are known as chatbots. Presently, these are extensively implemented in a variety of sectors, including banking and finance, health, and finance. These are utilised in areas including service, processing, payments, and marketing to carry out communicative duties.

Emotional intelligence is the capacity to recognise, constructively utilise, and control one's own emotions. Emotional intelligence within the domain of chatbots would denote the capacity of these digital assistants to identify and react to the emotions of users, therefore augmenting their exchanges and imbuing them with a more human-like quality.

The human mind is hardwired to believe that anything that can speak is human, rendering computers incapable of simulating human emotions and viewpoints. These technologies are capable of isolating and manipulating users during commercial transactions by imitating human social roles.

Conversational systems have a greater capacity to manipulate the young, elderly, and mentally ill. They can mislead users and legitimise harmful beliefs, so encouraging self-harm and harming others.

Conversational AI

Natural language processing (NLP) and machine learning (ML) are what make conversational AI different from a simple chatbot. They take the conversation to the next level and break out of the set flow. This helps conversational AI understand and process human language better, which makes it possible for machines to have smarter conversations with people.

The integration of conversational technology and NLP has led to the development of sophisticated systems capable of understanding user intent, maintaining context in conversations, and providing personalized responses. Ongoing advancements in machine learning and NLP contribute to the continual improvement of conversational technologies, making them more versatile and effective in various applications. Though Conversational AI has developed and made impressive advancements they are still far from replicating the complexity and depth of human emotions and understanding. They operate based on algorithms and data patterns and lack the true emotional intelligence that humans bring to interpersonal interactions.

The Adverse Effects of Conversational Technology

Conversational AI systems might encounter difficulties in precisely deciphering the emotional underpinnings of human inputs. Inadequate or improper reactions may result from misinterpretation, which may induce irritation or bewilderment. It may be difficult for AI systems to identify indicators of emotional distress in users. This may give rise to complications in situations where people are expressing concern or seeking aid, as the AI's response might not be suitable.

Although conversational AI is capable of providing guidance and information, it lacks the capacity to deliver authentic emotional support that would be found in a human connection. When empathy and comprehension are required, the limitations of artificial intelligence become obvious. Frustration may ensue when users attempt to establish an emotional connection or comprehend others due to the constraints of conversational AI. Unheard or misunderstood users may experience bad emotions as a result.

An over dependence on conversational AI to facilitate emotional exchanges could potentially lead to a deterioration in the ability of humans to communicate effectively with one another. Individuals may develop a diminished ability to navigate emotionally difficult interactions as time passes.

Conclusion

It is imperative that businesses and developers be cognizant of these potential hostile consequences and strive to mitigate them during the development and deployment of AI chatbot systems. It is critical to incorporate ethical considerations, transparency, and continuous evaluation in order to minimise adverse effects on human emotion and behaviour.

References:

1. Wach, K., Duong, C.D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J. and Ziemba, E., 2023. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review, 11(2*), pp.7-24.
2. Gkinko, L. and Elbanna, A., 2022. Hope, tolerance and empathy: employees' emotions when using an AI-enabled chatbot in a digitalised workplace. *Information Technology & People*, *35(6),* pp.1714-1743.
3. Wang, P. and Shao, J., 2022. Escaping loneliness through tourist-chatbot interactions. In *Information and Communication Technologies in Tourism 2022: Proceedings of the ENTER 2022 eTourism Conference, January 11–14, 2022* (pp. 473-485). Springer International Publishing.
4. De Freitas, J., Uğuralp, A.K., Oğuz-Uğuralp, Z. and Puntoni, S., 2022. Chatbots and Mental Health: Insights into the Safety of Generative AI. *Journal of Consumer Psychology*

Mesoscopic Modelling of Concrete Using Non-Spherical Aggregates

Shaji M Jamal
Department of Civil Engineering

Abstract

The realization of approximate practical aggregate models is a crucial link to the mesoscopic mechanical analysis for fully-graded concrete. Aggregate generation of concrete is one of the important problems in concrete mesoscopic mechanics. Many of the researchers have used spherical / circular aggregates in their mesoscopic numerical models for the analysis. As the spherical (3D) / circular (2D) aggregate model will not give the exact result, non-spherical aggregate model with different shapes are needed for better results. Compared to 3D models, obviously, 2D models can be easily generated. Geometrical models for concrete are generated taking the random structure of aggregates at the mesoscopic level into consideration. The generation process is based upon Monte Carlo's simulation method wherein the aggregate particles are generated from a certain aggregate size distribution and then placed into the concrete specimen in such a way that there is no intersection between the particles.

Index Terms – Mesoscopic model, numerical modelling, multiscale modelling of concrete.

Introduction

Research on the macro-mechanical properties and the strength of concrete material has important theoretical and practical significance for their ability in designing concrete structures. However, majority of the studies consider concrete as a homogeneous material at the macroscopic scale which did not allow one to establish the microstructure–property for designing better and superior fracture-resistant cementitious materials[1]. At the mesoscopic scale, concrete is treated as a multi-phase composite material consisting of coarse aggregate, mortar matrix with fine aggregate combined with it and interfacial transition zones (ITZ) between the coarse aggregate and the mortar matrix[2]. Mesoscopic model of concrete is believed to be better in the evaluation of mechanical properties such as crack propagation, elastic moduli, strength and failure patterns of concrete.

The main component of Mesoscopic model of concrete is the aggregate volume. Aggregates of random size are generated at random positions – mostly according to a gradation curve – checking the overlapping of each of the particles within the specimen volume or domain. So, generation of aggregate particles inside the mortar matrix is the major challenge in the numerical modelling of concrete in the Mesoscopic scale.

The mostly used shape is spherical but it will not give exact results because in engineering practices we are seldom using spherical aggregates. So there is a need for realistic shapes which will be closer to the actual shape of aggregates [3].

Meso-scale concrete model using circular shape of aggregate tends to give higher strength compared to the model using irregular actual shape of aggregate. This is attributed to the increased levels of stress concentration when modelling the actual shape of the aggregates. Therefore, the effect of the aggregate shape on the meso-scale concrete model is investigated in this study. Other factors like effect of aggregate distribution, volume fraction, etc., will also be considered.

Generation of meso-structure

The evaluation of composite behaviour of concrete at mesoscopic level requires generation of a random aggregate structure in which the shape, size and distribution of the coarse aggregate closely resemble real concrete in the statistical sense. This structure to be generated consists of randomly distributed aggregate particles and mortar matrix filling the space between the particles. The generation of the random geometrical configurations of the aggregate particles must satisfy the basic statistical characteristics of the real material. In order to produce the geometrical configuration which meets the requirements, the random sampling principle of Monte Carlo's simulation method is used. This random principle is applied by taking samples of aggregate particles from a source whose size distribution follows a certain given grading curve and placing the aggregate particles one by one into the concrete in such a way that there is no overlapping with particles already placed [4].

Generation of spherical aggregates

Aggregates generally occupy 50–60% of volume of concrete and greatly influence its properties, mix proportions and economy [3]. Sand, gravel and crushed stone are the primary aggregates used. For most concrete, the coarse aggregate represents around 50–60% of the concrete volume. Gravels constitute the majority of coarse aggregate used in concrete with crushed stone making up most of the remainder [5].

Generation of mesostructure mainly depends on aggregate size.

Aggregate size distribution

Aggregate size distribution is represented by the grading curve. Grading is the distribution of particles of a granular material among various size ranges, usually expressed in terms of cumulative percentage larger or smaller than each of a series of sizes of sieve openings, or the percentage between certain ranges of sieve openings [6]. Different grading ranges have been specified for economical design. In this paper aggregates of size range between 4.75mm and 25mm are considered. It is divided into five segments of size range (4.75- 10mm, 10-12.5mm, 12.5-16 mm, 16-20 mm, 20-25mm). The grading curve of a representative sample is shown in Figure 1.

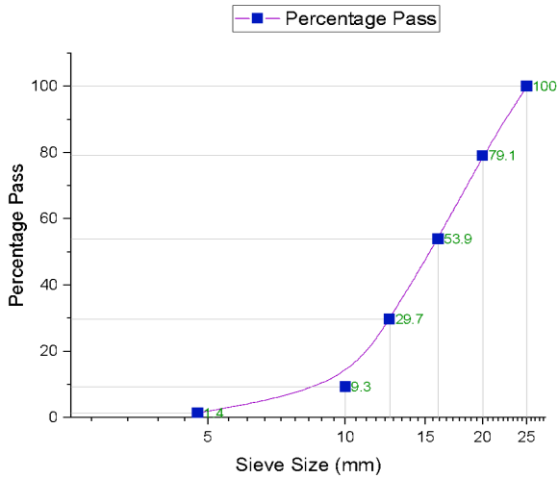


Figure 1 Grading curve of aggregates

Procedure to generate spherical aggregates within the grading segment:

- Step 1: Calculate the area of aggregate to be generated in the grading segment.
- Step 2: Generate a random number defining the size of an aggregate particle.
- Step 3: Calculate the area of the generated aggregate particle and subtract it from the area of aggregate within the grading segment.
- Step 4: Repeat steps 2 and 3 until the area of aggregate left to be generated is less than the area required for generating another particle
- Step 5: Repeat all the above steps for the next smaller size grading segment and then again for successively smaller size grading segment, until the last particle of the smallest size has been generated.

- Area of aggregate to be generated in the grading segment $[d_s, d_s+1]$ can be calculated using the formula,

$$A_p\,[d_s, d_s+1] = \frac{P\,(ds) - P\,(ds+1)}{P(dsmax) - P(dsmin)}\ x\ AR\ x\ A \qquad (1)$$

$P_{(ds)}$ - Cumulative percentage passing through the sieve with aperture diameter d_s, $P_{(ds+1)}$ - Cumulative percentage passing through the sieve with aperture diameter d_{s+1}, AR - Area ratio of aggregate = $\frac{c}{a+b+c}$, where a, b, c represent mix proportion of concrete, A - Total area of the concrete specimen

- Random number defining the size of an aggregate particle can be generated as follows:

Assuming that the size d has a uniform distribution between ds and ds+1, it may be calculated using the following expression:

$d = d_s+1 + \eta(d_s - d_s+1)$, where η is a random number which uniformly distributed between 0 and 1.

Placing process

In order to place an aggregate particle at a free position within the area, three obvious conditions need to be satisfied. Firstly, the whole particle must be completely within the boundary. Secondly, there must not be any overlapping with previously placed particles. To check the above two conditions Euclidean distance formula is used and finally each particle must be coated all around with a mortar film having a certain minimum thickness.

Distance between two point A (x_1, y_1, z_1) and B (x_2, y_2, z_2) is given by Euclidean distance formula,

$$D = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]} \qquad (2)$$

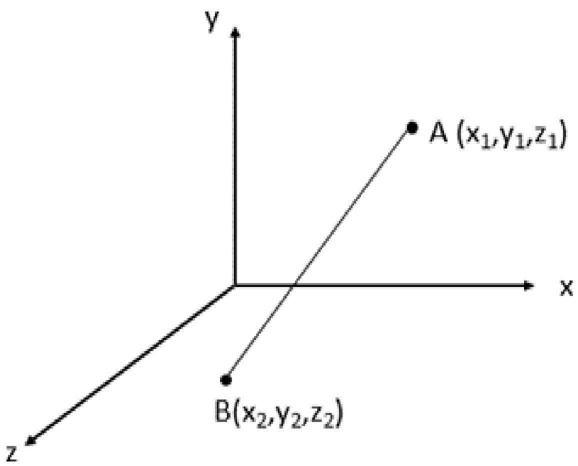


Figure 2 Euclidian distance

Generated aggregates

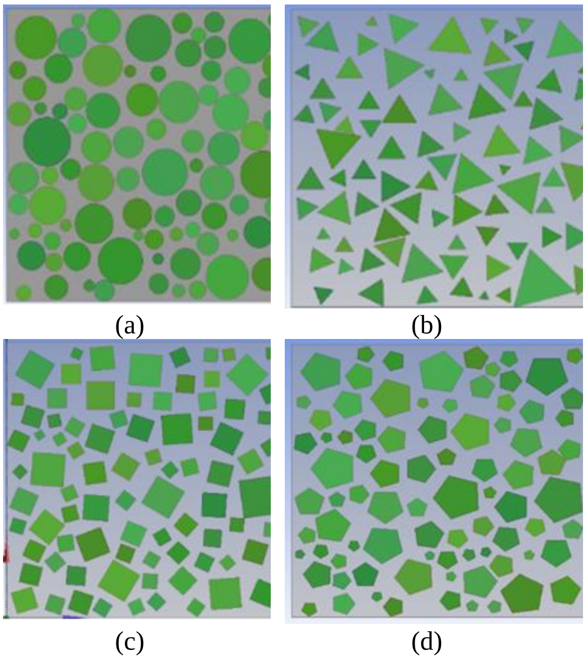


Figure 3 Generated shapes (a) Spherical (b) Triangular (c) Square (d) Pentagonal

Conclusion

The most important features of aggregates that affect the properties of concrete are shape of the aggregates, aggregate distribution and aggregate volume fraction. Conventional studies consider concrete as a homogeneous material at the macroscopic scale in which the stress- strain relationship may be described by non-linear macroscopic constitutive models with strain softening, which did not allow one to establish the microstructure– property for designing better and superior fracture resistant cementitious materials. At the mesoscopic scale, the concrete is treated as a three-phase composite material consisting of coarse aggregate, mortar matrix with fine aggregate combined with it, and interfacial transition zones between the coarse aggregate and the mortar matrix is not explicitly modelled to simplify the process of generation as well as testing of the model. Circular or spherical aggregate shape is commonly used for meso-scale finite element analysis because of simplicity. So we are comparing properties of non-circular aggregates.

References:

1. Jianxin Ding and Qingzhou Yang, 'Three-dimension random aggregate generation of numerical concrete model', 2018
2. Sun-Myung Kim, Rashid K. Abu Al-Rub,'Meso-scale computational modeling of the plastic-damage response of cementitious composites' TX 7784, 2010.
3. A.K.H. Kwana, P.L. Nga, b, Z.M., Wangc, 'Mesoscopic analysis of crack propagation in concrete by nonlinear finite element method with crack queuing algorithm', Modern building materials, structures and techniques, MBMST 2016
4. P. Wriggers, S.O. Moftah, 'Mesoscale models for concrete: Homogenisation and damage behaviour',2005
5. Huaifa Ma, Wenxiang Xu, Yuncheng L, Random aggregate model for mesoscopic structures and mechanical analysis of fully-graded concrete, 2016
6. Hao Jin, Yuliang Zhou, Binglong Wang, and Shunhua Zhou, 'Mesoscopic finite element modeling of concrete considering geometric boundaries of actual aggregates', Shanghai 201804
7. Yijiang Peng, Xiyun Chen, Liping Ying, Ying Chen, and Lijuan Zhang Key, 'Mesoscopic numerical simulation of fracture process and failure mechanism of concrete based on convex aggregate', Beijing 100124.
8. Lin Liu A, Dejian Shen A, Huisu Chen B, Wenxiang XU. 'Aggregate shape effect on the diffusivity of mortar: A 3D numerical investigation by random packing models of ellipsoidal particles and of convex polyhedral particles',2014.

Multiport DC-DC Converters: An Efficient Tool for Integrating Hybrid Energy Sources

Dr. Sivaprasad A
Dept. of Electrical & Electronics Engineering

Abstract

Hybrid energy systems are becoming more and more popular due to the rising demand for renewable energy sources and the requirement for sustainable power systems. These systems combine several energy sources, including wind, solar, and energy storage, to generate power that is dependable and effective. Superior power handling capabilities in steady-state operation and improved dynamic response in transient conditions are guaranteed by a well-designed hybrid energy storage system. However, without a suitable power electronic interface, the HES concept is incomplete. In conventional methods, multiple input sources are integrated through the use of parallel connected Single Input DC-DC (SIDC) converters. However, the SIDC converters have a number of significant drawbacks, including high system complexity, high cost, low efficiency, low reliability, and loss of system compactness. In order to overcome these drawbacks, the idea of multiple input DC-DC (MDC) converters was created. Multiport DC DC converters, or MDCs, are essential to the integration of hybrid energy systems because they offer a common DC bus for connecting different energy sources, loads, and storage devices. There have already been reports of various MDC converter types in the literature. The main disadvantages of the reported MDC converters are their complex structure, lower efficiency, and inability to supply power simultaneously. The topologies, control schemes, and applications of the latest developments in MDCs for hybrid energy integration are reviewed in this article. We go over the benefits and drawbacks of various MDC topologies, including resonant and non-resonant converters, isolated and non-isolated converters, single-stage and multi-stage converters. We also discuss the advantages and disadvantages of the different control strategies used in MDCs, such as model-predictive, decentralized, and centralized control. Lastly, we go over how MDCs are used in different hybrid energy systems, including renewable energy systems, electric cars, and microgrids. The development of new converter topologies with enhanced efficiency, dependability, and flexibility; the improvement of control strategies to achieve optimal power management and dynamic system response; and the investigation of new applications for MDCs in emerging energy systems are some of the future research directions we outline for MDCs.

Keywords: Control strategy, topology, power management, microgrid, renewable energy, hybrid energy integration, and electric vehicles

Introduction

Due to their scarcity and the ensuing environmental problems, countries are currently concentrating on lowering the amount of electricity generated from conventional petrochemical resources. In an effort to address the aforementioned problems, efforts are being made to substitute renewable energy sources for petrochemical-based energy sources. Many applications use highly intermittent renewable energy sources, such as wind, solar PV, micro hydro, etc., as stand-alone systems. However, because of their extreme reliance on unstable weather, stand-alone renewable energy sources such as solar photovoltaics, wind, micro hydro, etc., are not advised for the production of electricity in a reliable manner. Combining multiple non-conventional energy sources to provide dependable power is a feasible solution offered by the Hybrid Energy System (HES)[1].

When compared to a single source energy system, hybrid energy systems operate more efficiently, cleanly, and with greater reliability, which is why they are becoming more and more popular in the field of electric power systems. Better dynamic response during transients and strong power handling during steady-state operation are provided by a well-designed HES. A power electronic interface plays a crucial role in integrating multiple energy sources with varying voltage (V) − current (I) characteristics to form hybrid energy systems (HES).

Multi Input DC-DC Converters

By integrating various energy sources and storage devices to produce dependable and effective power generation, hybrid energy systems present a workable solution. In order to connect different energy sources, storage devices, and loads with a common DC bus, multiport DC-DC converters, or MDCs, are necessary parts of such systems. MDCs are essential to the hybrid system's ability to control voltage, manage power flow, and maximize energy efficiency.Figure 1, depicts the architecture of a hybrid energy system (HES) that integrates multiple energy sources through the use of a MDC converter. When compared to conventional methods, the figure illustrates how the use of MDC converter lessens the requirement for multiple conversion stages. The main advantages of MDC converters are reduced system complexity, easy structure and control, high system reliability, and lower cost. Additionally, regulated output voltage improves system reliability by enabling efficient power distribution.

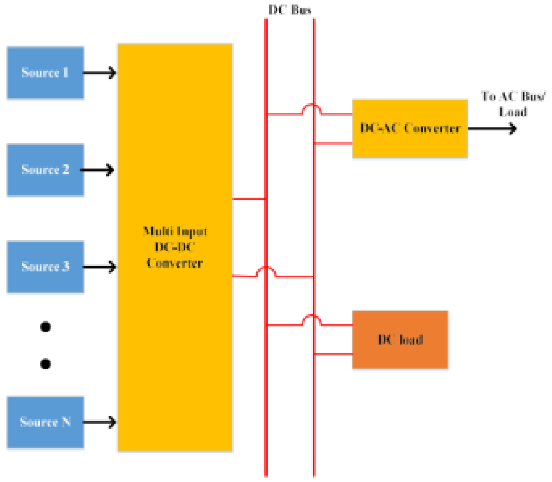


Fig.1. HES with multi-input DC-DC converter

Because MDC converters are superior to single input converters in a number of ways, their use in distributed generation and microgrids is rapidly expanding. For this reason, using MDC converters is required in order to achieve a regulated DC bus voltage. The two main categories of MDC converters are isolated type and non-isolated type.

Based on their topologies, guiding concepts, and port count, MDCs can be divided into a number of groups. Among the most popular topologies are the following:

Different types of Multiport DC-DC Converters

Isolated vs. non-isolated converters: Isolated converters [2] reduce noise and improve safety by offering galvanic isolation between the input and output ports. Non-isolated converters lack galvanic isolation but are more efficient and compact [3].

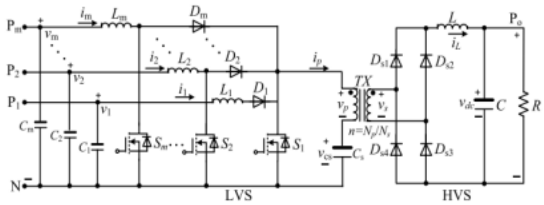
Single-stage vs. multi-stage: Single-stage converters are less complicated and less expensive, but their power density and voltage gain may be limited. Higher voltage gain and power density are achieved by multi-stage converters, but they come with more parts and complicated controls.

Resonant vs. Non-resonant: Resonant converters have low EMI and high efficiency, but stability needs to be ensured by careful design and control. Although simpler, non-resonant converters may have higher EMI and lower efficiency.

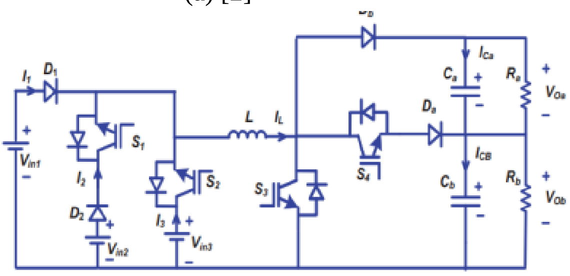
Bidirectional vs. Unidirectional: Bidirectional converters can transfer power in both directions, enabling power flow between different ports. Unidirectional converters can only transfer power in one direction.

The choice of MDC topology depends on the specific requirements of the hybrid energy system, such as voltage levels, power ratings, isolation requirements, and desired efficiency.

Novel Topologies of Isolated and non-isolated multiport DC-DC converters are depicted in Figure. 2(a-b).



(a) [2]



(b) [3]

Fig 2. Circuit diagrams of Multiport DC-DC Converters (a) isolated (b) non-isolated.

Control Techniques for Multiport DC-DC Converters

To manage power flow, control voltage levels, and guarantee stable operation, MDCs need complex control strategies. Typical control tactics consist of:

Centralized Control: In this approach, all of the MDC's ports are managed by a single controller. For systems with many ports, it can be computationally costly, but it provides a straightforward and reliable method.

Decentralized Control: This approach provides more flexibility and modularity by using separate controllers for every port. On the other hand, it might call for more intricate controller coordination and communication.

Model-Predictive Control (MPC): This sophisticated control technique makes use of a system model to forecast behavior and enhance control choices. Although MPC offers excellent performance, it necessitates large computational resources.

Applications of Multiport DC-DC Converters

MDCs are being used more often in a range of hybrid energy systems, such as:

Microgrids: By enabling the integration of renewable energy sources, storage devices, and loads within a microgrid, MDCs help local communities generate power that is dependable and efficient.

Electric vehicles (EVs): MDCs are essential for controlling the flow of power in EVs between the motor, battery, and regenerative braking system.

Renewable Energy Systems: MDCs make it possible to integrate several renewable energy sources, like wind and solar power, into the grid in an efficient manner.

Future Research Aspects

Subsequent investigations ought to concentrate on tackling these obstacles and creating MDCs that are:

More compact: smaller size and fewer components for simpler HES integration. Wider voltage gain: Greater range of input and output voltages can be handled by a voltage gain range that is wider.
Less complicated control techniques: simpler to implement and requiring less computing power.

Conclusion

In hybrid energy systems, multiport DC-DC converters are crucial for the integration and control of various energy sources. Significant benefits from their special features include less complexity, increased effectiveness, and increased flexibility. These converters are anticipated to become more significant in the future of clean and sustainable energy generation as research and development continue. An overview of the fascinating field of multiport DC-DC converters is given in this article. These converters have many benefits and show great potential for future research, which could completely change how we integrate and manage energy.

References:

1. K. Shivarama Krishna and K. Sathish Kumar, "A Review On Hybrid Renewable Energy Systems," in Renewable And Sustainable Energy Reviews, vol. 52, pp. 907-916, 2015.
2. J. Zeng, W. Qiao, L. Qu and Y. Jiao, "An Isolated Multiport DC–DC Converter for Simultaneous Power Management of Multiple Different Renewable Energy Sources," in IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 2, no. 1, pp. 70-78, March 2014.
3. S. Athikkal and K. M. R. Eswar, "A Three Input Dual Output Boost Type Non Isolated DC DC Converter," 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1-6.
4. Y. Han, M. Ranjram and P. W. Lehn, "A bidirectional multi-port DC-DC converter with reduced filter requirements," 2015 IEEE 16th Workshop on Control and Modeling for Power Electronics (COMPEL), Vancouver, BC, 2015, pp. 1-6.
5. L. Solero, A. Lidozzi and J. A. Pomilio, "Design of multiple-input power converter for hybrid vehicles," in IEEE Transactions on Power Electronics, vol. 20, no. 5, pp. 1007-1016, Sept. 2005.
6. V. A. K. Prabhala, M. Khazraei and M. Ferdowsi, "A new control strategy for a class of multiple-input DC-DC converters," 2014 International Conference on Renewable Energy Research and Application (ICRERA), Milwaukee, WI, 2014, pp. 508-513.
7. D. Yang, M. Yang and X. Ruan, "One-Cycle Control for a Double-Input DC/DC Converter," in IEEE Transactions on Power Electronics, vol. 27, no. 11, pp. 4646-4655, Nov. 2012.

Quantum Dots and Their Applications

Dr.Kavitha S.
Department of Basic Science and Humanities

Abstract

Many years have passed since the first works on the reduced dimensionality of semiconductors, which led to the concept of quantum dots (QDs), also known as “artificial atoms”. Semiconductor nanocrystals with diameters in the nanometer range display quantum size effects in their optical and electronic properties. Many QD material systems achieve tunable and efficient photoluminescence (PL) with narrow emission and photochemical stability and core-shell structures[1]. As a result, quantum dots (QDs) have been integrated as active components in various devices and applications. Several of these applications are presently available in the market and are part of our daily lives, as is the case with QD-based displays.

QDs have a variety of applications in light emission, conversion, and detection, thanks to their exceptional optical properties. Biomedical, environmental, and catalytic applications are the focus of this text.

1. Biomedical and Environmental Applications

The unique properties of quantum dots, such as high luminescence, narrow emission, and low toxicity and biocompatibility (depending on the elemental composition), make them an ideal candidate for various applications in bioimaging, diagnostics, and biosensing. These applications are primarily related to the fields of biomedical and environmental sciences. In that sense, fibrous phosphorus QDs were demonstrated as fluorescent labels for human adenocarcinoma bioimaging. $CuInS_2$/ZnS QDs, with low nonspecific binding and possible bioconjugation, were demonstrated as markers for cell imaging[2]. This same application and biosensing have been proposed for MoS_2 QDS produced using a miniaturized continuous multicycle microfluidic process. Bioimaging and biosensing were also demonstrated for QD complexes comprised of zinc chalcogenide QDs, such as undoped and Mn^{2+}-doped ZnS and surface zinc quinolate (ZnQ_2)complexes. Ultrabright graphene QDs obtained under microwave irradiation have been suggested for cell-imaging applications. N−S-doped graphene QDs, highly stable in the presence of metal ions, were also used for cell imaging. Similar nanomaterials have been shown as highly selective probes for Hg^{2+}ion sensing in aqueous solution and therefore evaluation of drinking water quality. S-doped graphene QDs coupled to Au nanoparticles showed the sensitive and selective detection of 4-nitrophenol in water and wastewater. On the other hand, a graphene oxide QD-based adsorbent was found to efficiently bind rhodamine B and methylene blue, suggesting a possible application in pollutant control.

Real water samples were also analyzed for the presence of toxic metal ions using luminescent porous Si with wide-size-distributed Si QDs. Moreover, blue fluorescent WS_2QDs have shown high sensitivity and selectivity for ferric (Fe^{3+})ions, while surface-functionalized $MoSe_2$QDs worked as chemosensors for Cu^{2+}, 2,4,6-trinitrophenol, and melamine. Peroxide sensing was demonstrated for Fe_3O_4QDs by both the nanozymatic activity and fluorescence quenching mechanisms. Regarding clinical diagnostics, a highly sensitive photo-electrochemical immunosensor for an ovarian cancer bio-marker, CA125, was demonstrated using SiO_2@polydopaminecore−shell anoparticles and CdTe QDS. Detection of carcinoembryonic antigen and monitoring of the T_4 polynucleotide kinase activity were demonstrated using Mn:ZnCdS@ZnS and CdTe QDs on SiO_2 microspheres. Functionalized CdSe/CdS/ZnS core−shell−shell QDs and terbium-labeled antibodies were demonstrated as adenosine diphosphate sensors.

2. Catalysis and other Applications

Semiconductor photocatalysis technology converts light to chemical energy; QDs have also been explored to improve existing or create new catalytic routes. Carbon QD-modified graphitic carbon nitride provided photocatalytic H_2 evolution with little or no decay in activity for half a year. Light-driven generation was also obtained with carbon QD-sensitized TiO_2/Pt nanocomposites. N-doped graphene QDs produced by liquid-laser ablation in various aqueous solutions showed high catalytic selectivity for the O_2 reduction reaction and also for the catalytic conversion of 4-nitrophenol to 4-aminophenol under near-IR light[3]. The photocatalytic synthesis of imines was demonstrated with catalysts decorated with CdS QDs. Separately, enhanced enzyme activity was obtained with a strong synergy between DNA cages and CdSe/CdZnS/ZnS core−shell−shell and CdSe/ZnS core−shell QDs. As for different applications, fluorescent CdTe QDs were assembled with maghemite nanoparticles and suggested for magnetic particle inspection in ferromagnetic materials. Encapsulation and nanotags for authentication of goods can also be found among QD applications. Micrometer-sized capsules thermally stable up to 350°C were obtained by dispersing CdSe/ZnS QDs in liquid crystals, while highly stable PbS and PbS/CdS QDs with hybrid coatings were proposed as luminescent nanotags for liquid petroleum products.

In summary, QDs remain an active and intense investigation in applied nanomaterial research.

References:

1. Kubendhiran, S.; Bao, Z.; Dave, K.; Liu, R.-S.Microfluidic Synthesis of Semiconducting Colloidal Quantum Dots and Their Applications.ACS Appl. Nano Mater.2019,2, 1773−1790
2. Mallick, S.; Kumar, P.; Koner, A. L. Freeze-Resistant Cadmium-Free Quantum Dots for Live-Cell Imaging.ACS Appl. Nano Mater.2019,2, 661−666
3. Hu, H.; Quan, H.; Zhong, B.; Li, Z.; Huang, Y.; Wang, X.;Zhang, M.; Chen, D. A Reduced Graphene Oxide Quantum Dot-Based Adsorbent for Efficiently Binding with Organic Pollutants.ACSAppl.Nano Mater.2018,1, 6502−6513

Resistive RAM: A Breakthrough in Non-Volatile Memory Technology

Ancy Joy
Dept of Electronics and Communication Engineering

Abstract

Resistive RAM (RRAM) has emerged as a transformative force in the realm of memory technologies, capitalizing on resistance-based memristor cells to redefine the landscape of non-volatile memory storage. This article explores the distinctive qualities that make RRAM a standout contender in the pursuit of efficient and innovative memory solutions. RRAM's high storage density opens new possibilities for compact gadgets and memory-intensive applications, while its fast write speeds contribute to enhanced overall system performance. Moreover, we examine the promising research suggesting that RRAM advancements may address drawbacks associated with existing low-power memory devices, particularly in terms of power efficiency and latency.

Index Terms - RRAM, Stability, Circuit Design, Noise Margin.

Introduction

Complementary metal-oxide semiconductors have been widely used in microchips and integrated circuits. The main reason behind its wide use is as follows: CMOS suffers from a low noise margin i.e., allowable noise at the input does not affect the output. The reduction in W/L ratio due to the advancements in CMOS makes them feel less voltage, resulting in less noise margin. The static power consumption of CMOS is minimal. The power utilised whenever the transistors are not switching is known as static power. It has a less complicated circuit, CMOS also have some disadvantages, CMOS suffers from scaling issues. As CMOS based devices are shrunk down, their power usage per unit area increases. As a result, scaled devices become progressively heated. This is a significant performance bottleneck for scaled equipment. Scaling also reduces the mobility of charge carriers which thereby reduces the overall gain of the device.

A Memristor is an electrical device that consists of 2 small films which is the abbreviation of memory resistor. Memristor as a core element of the circuit has attracted widespread attention in the international research community of electrical & computer engineers. A key characteristic of the memristor is the potential to remember its state history. Therefore, it will be highlighted in the implementation of VLSI (RRAM) fixed memory. Memristor provides even more amazing environments in electrical circuits opening new horizons in future systems. Memristors are significant because they are non-volatile. It does not lose information when the system is turned off.

The most noticeable application of Memristor is Memristor based non-volatile memory [1]. A statistical memory cell based on memristor compared to SRAM and DRAM can show consistency, good durability, compatible with standard CMOS power, additionally, it has no leakage capacity. CAM (Content Addressable Memory) is a data storage device that stores memory in cells. If any part of the memory is included Content Addressable Memory compares the input with all the stored data. It is a high-speed technology.

RRAM Structure

Resistive RAM is a two-terminal passive device with an insulating layer sandwiched between two electrically controlled electrodes, resulting in a MIM (Metal Insulated Metal) structure [2]. Stacks are made from metals such as Ta, La, W, and TiN. The dielectric constant and thermal stability of the insulating layer are both high. An external voltage put across the MIM structure can change the switching properties and resistance. The topmost electrode TE, the bottom electrode BE, and an insulating oxide layer is shown in figure 1 below. The insulating material can be binary RRAM, which has a two-level stack, or trinary RRAM, which has a three-level stack.

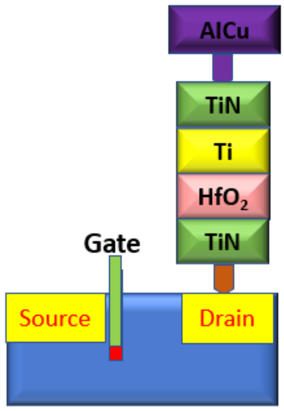


Figure 1: RRAM Structure

RRAM Working

Electrically, resistive memory can be shifted between low and high resistance states. This is accomplished by altering the Resistance of the MIM structure. RRAM can have either bipolar or unipolar switching properties as indicated in the figure 2.

Bipolar Switching: The direction of switching is determined by the polarity of the applied voltage. As a result, the set and reset conditions take place at opposite polarities.

Unipolar Switching: In unipolar switching, however, the polarity of the external voltage provided does not affect the switching direction. It is fully dependent on the magnitude of the applied external voltage. As a result, just one polarity is used to set or reset.

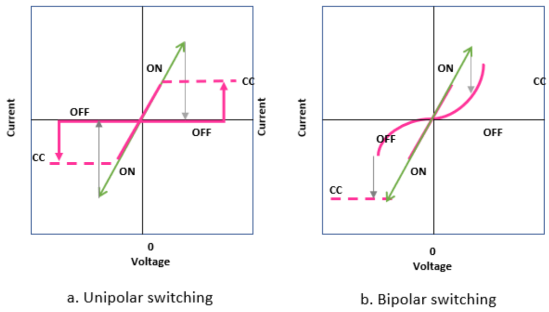


Figure 2: RRAM Working

RRAM States

When a sufficient external voltage is provided, the metal insulated metal structure can be electronically switched between high and low resistance states. The various states of RRAM are shown in figure 3. High resistance state or HRS with logic 0 and low resistance state or LRS with logic 1 are the terms that describe these states. When a voltage is applied to the Resistive RAM, it transitions from HRS to LRS, which is known as the forming state. The RRAM is in a single state once it has been produced. A SET state is an on-the-go state. The potential changes from LRS to HRS as the polarity of the voltage changes and this is referred to as an off state. A reset state is the opposite of the off-state.

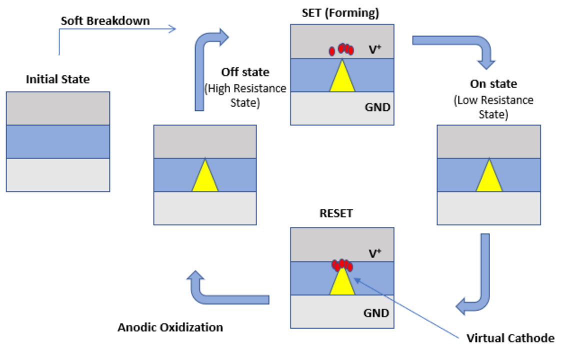


Figure 3: States in RRAM

Memristor modelling technology

Memristors, initially predicted in 1971 and acknowledged as the fourth fundamental circuit element alongside resistors, capacitors, and inductors, have garnered significant attention in recent years. These unique devices exhibit a direct coupling between magnetic flux and charge, offering the remarkable ability to adjust their resistance value when an adequate voltage is applied, which persists even after the voltage is removed. This resistance switching behaviour is characterized by a pinched hysteresis loop observed in the current-voltage characteristics of memristors. The compact dimensions and consistent structure of memristors facilitate their integration into crossbar arrays, facilitating the development of dense memory architectures. Additionally, memristors have been explored for implementing logic functions and interconnections in various studies.

Regarding the effective memristance[3] of the memristor-based device, Equation (1) proposes the formulation;

$$M(t) = RON\frac{\omega(t)}{L} + ROFF\left(1 - \frac{\omega(t)}{L}\right) \quad (1)$$

ROFF : Maximum resistance
RON : Minimum resistance
L : thickness of a memristor.
w(t) : thickness of the doped conductive zone as a function of time.

In the current semiconductor memory environment, RRAM is one of the finest alternatives for memory device manufacturing.

Conclusion

Resistive RAM, a new memory technology under development, stores data in resistance-based memristor cells, whereas SRAM stores charges at the node of cross-coupled inverters. However, challenges such as manufacturing scalability, cost-effectiveness, and integration persist. Ongoing research aims to refine RRAM and other emerging memory technologies for diverse applications in the computing and electronics landscape.

References:

1. E. Vianello et al., "Resistive memories for ultra-low-power embedded computing design," in IEDM Tech. Dig., Dec. 2014, pp. 6-1–6-3.
2. S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in memory with spin-transfer torque magnetic RAM," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 26, no. 3, pp. 470–483, Mar. 2018.
3. B. Giraud, A Makosiej, R. Boumchedda, N. Gupta, A. Levisse, E. Vianello, J. P. Noel, "Advanced Memory Solution for Emerging Circuits and Systems," IEEE Int. Elect. Dev. Meet. Tech. Dig, pp. 473- 475, 2017. DOI: 10.1109/IEDM.2017.8268422.

Seismic Behaviour of Composite Shear Wall

Jereena Jawahar
Department of Civil Engineering

Abstract

Structural design of high-rise building under seismic loading is primarily concerned with the provision of shear wall. Shear walls are vertical element of a system that is designed to resist in-plane lateral forces like wind and seismic loads. Conventional reinforced concrete shear wall and steel plate shear walls are used worldwide for resisting wind and earthquake loads. Composite construction methods have been accepted in the industry in which composite columns, beams and shear walls are started in practice recently. Composite shear walls (CSW) are a combination of steel and concrete sections, which unites the advantages of steel and concrete shear wall systems.

Introduction

Structural design of high-rise building under seismic loading is primarily concerned with the provision of shear wall. Shear walls are vertical element of a system that is designed to resist in-plane lateral forces like wind and seismic loads. Replacing conventional reinforced concrete shear wall and steel plate shear walls composite construction methods have been accepted in the industry in which composite columns, beams and shear walls are started in practice recently. Composite shear walls (CSW) are a combination of steel and concrete sections, which unites the advantages of steel and concrete shear wall systems.

Composite Shear Walls

Composite shear walls (CSW) are a combination of steel and concrete sections, which unites the advantages of steel and concrete shear wall systems. The main components of a composite shear wall are a reinforced concrete shear wall, one or two steel plates, boundary steel beams and columns, and mechanical shear studs or connectors. The steel plate is either attached on one or two sides of the reinforced concrete shear wall (RCSW), or is encased within it. In both cases, shear connectors are welded to the steel plate in order to assure composite behaviour. Various cross-sections for CSW are shown in Fig. 1.

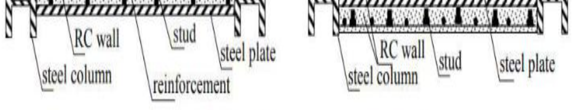


Fig.1 Cross section view of Composite Shear wall

Although composite shear walls are difficult to construct compared to RCC or steel plate shear walls, for the same shear capacity, a CSW will have a smaller thickness, less weight and most likely larger shear stiffness than a RCSW. The lesser the weight of CSW leads to lower seismic forces and smaller foundations. The RC wall of the CSW can be either cast-in-place or precast. In the latter case, RC wall can be bolted to the steel plates at any convenient time during construction. In a CSW, the RC wall restrains the steel plate and prevents its buckling before it yields in shear. Thus, the shear capacity of the steel plate can be significantly greater than its capacity to resist shear by the tension field action as in Steel plate shear walls. The RC wall provides sound and temperature insulation, in addition to fire proofing to the steel plates. The damage of the CSW after a moderate or strong earthquake is repairable with minimal disturbance to its functionality.

Conclusions

Providing RC buildings with composite shear walls instead of RC shear walls is a very efficient and attractive structural decision which can be implemented in regions vulnerable to earthquakes. Composite shear wall system is advantageous over conventional RC and steel shear wall in terms of ductility, stiffness, shear resistance, energy dissipation and overall seismic performance of the building. A significant reduction in the total dead load of the building due to less thickness of the CSW compared to the RC shear wall.

References:

1. Wenyuan Zhanga b, Ke Wanga b, Yong Chenc, Yukun Dinga,b (2019): 'Experimental study on the seismic behaviour of composite shear walls with stiffened steel plates and infilled concrete', Elsevier, June 2019.
2. Reslan N, Masri A, Machaka M (2018): 'Composite Shear Walls an Efficient Seismic Resistant System for MultiStory Buildings', Journal of Steel Structures & Construction, Volume 4, Issue 1, January 2018.
3. Mu-Wang Wei1, J. Y. Richard Liew, Xue-Yi (2017): 'Shear Resistance of Buckling- restrained Steel Plate Shear Walls', International Journal of Steel Structures, September 2017.

Semantic Textual Information Recognition System

Shiji S
Dept. of Artificial Intelligence & Data Science

Abstract

A document can contain a large number of text elements which were placed at defined locations in different formats. The different nature of the document content creates a challenge during the process of editing, extracting and combining the text elements. The content of the document can be viewed in any viewing device and the viewing device is not aware about the organizational structure and it cannot recognize the relationship between the elements and this lack of knowledge creates problems during analysis of the document. To solve the above problem there is a need for an improved and reliable automated system for semantic textual information recognition systems.

Index Terms - Semantic textual information recognition, Semantic Clustering, Machine Learning

Introduction

Documents contain a collection of elements that are drawn at different coordinates on a page. Each document are saved in different formats. This heterogeneous nature of document creates challenges during the processing stage of document. A document viewing device is able to display the document to the user but it has no knowledge in the internal structure of a document [1-2]. For example, a table is displayed as a series rectangle cells with text inside it, which the human viewer can recognizes it as a table. However, the document viewing device which displays this document has no idea about the relationship between the text groupings[3]. This lack of knowledge creates a problem during the processing or analyzing the document using machines or automated means.

To analyze and process the document more effectively by means of machines a semantic textual information recognition system is provided. It consists of mainly two components: processor and memory. The system also includes optical recognition techniques which are capable of extracting text elements or information from any kind of documents like files, images, receipts or pay checks. Memory is used to store the instruction and results. Processor is connected with memory which is capable of receiving text elements along with its respective coordinates. The processor mainly includes spatial reconstruction module, semantic clustering module, rank clustering module and machine learning module.

Spatial reconstruction

The spatial reconstruction module receives the text elements along with its element coordinates. The module initially identifies the text elements based on the text elements coordinates which are associated with each of the text elements. Identification of text elements is based on the cubic spline method or a linear extrapolation method. Identified text elements are sorted or arranged along the information axis based on the element coordinates. Information axis can have different orientations like linear axis, non-linear axis, rows and columns or an angle based orientation. Spatial reconstruction module helps to establish a structured representation of text elements which recognize the spatial relationship of text elements.

Semantic Clustering

In the semantic clustering module, clusters are formed from the structured representation of text elements using the semantic data model. Each semantic cluster includes the text elements having similar properties or some relationship with each other. Similarity of the text elements can be calculated using Euclidean distance, Proximity matrix or kernel method. The semantic clusters are provided to the rank clustering module to rank the semantic clusters using a probabilistic ranking method based on one or more heuristic rules

The generation of rank clusters can also be done using the methods like Jaro - Winkler distance, Levenshtein distance or longest common subsequence distance. The top ranked clusters are considered to have a high similarity of text elements or high relationship between the text elements.

Machine Learning

The machine learning module uses random decision forest or any deep learning algorithm to update the data model based on the feature set extracted from text elements, semantic cluster and ranking cluster. This update allows the system to adapt and improve its semantic understanding of the text element in the document.

Memory of the system also includes one or more dictionaries which are used to store the domain of the file being parsed. The system can use the information in the dictionary as a reference to the files.

Conclusion

Semantic textual information recognition system identifies, organizes and analyses the textual elements in documents. The detailed knowledge of the intended structure of the document along with the finding of relationship of text elements makes the system more accurate and efficient for processing text elements in various kinds of documents.

References:

1. Seungah Suh, Ghang Lee, DaeyoungGil & Yonghan Kim ‘Automated hand-marked semantic

text recognition from photographs' Scientific Report 2023.

2. Xinqi Yang 1,2,3, Wushour Silamu 1,2,3, Miaomiao Xu 1,2,3 and Yanbing Li. Display-Semantic Transformer for Scene Text, Sensors 2023, 23, 8159
3. Muhammad Mansoor, Zahoor Ur Rehman, Muhammad Shaheen,Muhammad Attique Khan, Mohamed Habib. Deep Learning based Semantic Similarity Detection using Text Data Information Technology and Control Vol. 49 / No. 4 / 2020

Unraveling the Potential of FinFET Technology in Sub-5nm Nodes

Jisha Jacob
Dept of Electronics and Communication Engineering

Abstract

The semiconductor industry continues its relentless pursuit of innovation to push the boundaries of transistor scaling. FinFET (Fin Field-Effect Transistor) technology has emerged as a pivotal solution in achieving enhanced performance, power efficiency, and density in integrated circuits. This research article delves into the architecture, challenges, and prospects of FinFETs, specifically focusing on their application in sub-5nm nodes. It explores the intricacies of FinFET design, material considerations, process integration, and the implications of scaling down to sub-5nm dimensions. Moreover, this article investigates the potential limitations and novel approaches to overcome the hurdles encountered in FinFET technology as it ventures into the realm of ultra-miniaturization.

Index Terms - FinFET technology, GAA, Sub-5 nm

Introduction

The relentless pursuit of miniaturization in semiconductor technology has ushered in an era where transistor dimensions are approaching the sub-5nm realm.

Evolution and Implications of Sub-5nm Technology: Unraveling the Nanoscale Frontier

The evolution of semiconductor technology has been a journey characterized by relentless pursuit of miniaturization, pushing the boundaries of transistor scaling to ever smaller dimensions. From the early days of micrometer-scale transistors to the present precipice of innovation, the evolution of semiconductor nodes has been a testament to human ingenuity and technological advancement. The ground-breaking leap into sub-5nm technology heralds a paradigm shift, where transistor dimensions have reached the brink of the unimaginably small. This evolution has been driven by the quest for improved performance, enhanced power efficiency, and increased transistor density, necessitating revolutionary advancements in fabrication techniques, materials science, and innovative device architectures.

The transition to sub-5nm technology marks a critical juncture in the semiconductor industry, where the effects of quantum physics and fundamental limitations are brought to the forefront. As transistor dimensions approach atomic scales, quantum effects such as quantum tunneling become increasingly prominent, posing significant challenges to traditional transistor operation. The evolution of lithography techniques, including the advent of extreme ultraviolet (EUV) lithography, has been instrumental in enabling the production of transistors at these ultra-small dimensions. Innovations in materials and novel transistor structures like nanosheet transistors and gate-all-around (GAA) architectures have provided the necessary tools to mitigate quantum effects and improve transistor performance, ushering in a new era of nanoscale semiconductor manufacturing with profound implications for technology and society.

FinFET, short for Fin Field-Effect Transistor, stands as a revolutionary advancement in semiconductor technology, transforming the landscape of transistor design and paving the way for enhanced performance and efficiency in electronic devices. Unlike traditional planar transistors, FinFETs leverage a three-dimensional structure, resembling a "fin" rising from the silicon substrate, which allows for better control over the flow of electrical current. This architecture provides FinFETs with superior electrostatic control and mitigates issues like leakage currents, enabling unprecedented levels of performance while consuming minimal power.

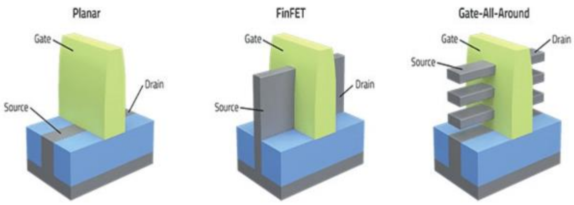


Fig a. Planar vs FinFET vs GAA structures

FinFETs

The introduction of FinFETs has been a game-changer in the semiconductor industry, addressing the limitations encountered as transistor dimensions shrink. By surpassing the constraints of traditional planar designs, FinFETs have enabled significant advancements in transistor scaling, leading to faster and more energy-efficient electronic devices. This transformative technology has found applications across various sectors, from high-performance computing and mobile devices to emerging fields like artificial intelligence and Internet-of-Things (IoT), shaping the future of electronics with its exceptional capabilities and paving the way for continued innovation in semiconductor manufacturing.

The development of FinFET technology utilizing sub-5nm nodes stands as a hallmark achievement in the realm of semiconductor innovation, revolutionizing transistor design and performance. FinFETs, a three-dimensional transistor structure, have emerged as a pivotal solution to address the challenges of scaling down transistors to ultra-miniaturized dimensions. As the semiconductor industry delves into sub-5nm technology, the evolution of FinFETs has been at the forefront, pushing the boundaries of transistor scaling beyond what was previously thought feasible. Leveraging nanoscale dimensions and innovative fabrication techniques, FinFETs have showcased unparalleled control over channel conductance while effectively mitigating issues related to leakage currents and improving overall transistor performance.

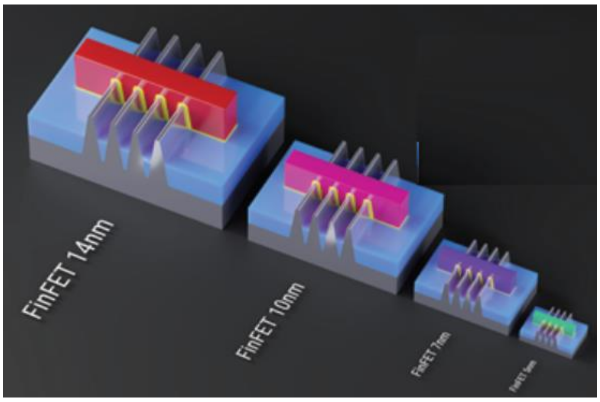


Fig b. Sub 5-nm nodes in FinFETs

The utilization of sub-5nm nodes in the development of FinFET technology represents a culmination of years of research and advancement in semiconductor manufacturing. These ultra-miniaturized transistors have seen a shift from traditional planar structures to three-dimensional architectures, allowing for better electrostatic control and reduced short-channel effects. The evolution of FinFETs at sub-5nm nodes has been fueled by cutting-edge lithography techniques, materials innovation, and meticulous engineering, paving the way for enhanced power efficiency, improved performance, and increased transistor density. As the industry marches forward into this nanoscale frontier, FinFETs utilizing sub-5nm technology are poised to redefine the landscape of semiconductor devices, enabling the creation of more powerful, energy-efficient electronic systems across a spectrum of applications.

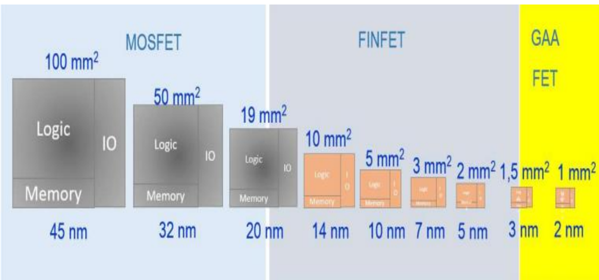


Fig c. Transition from FinFETs to GAAs

The transition from FinFETs to GAA transistors represents a natural evolution to overcome the limitations of traditional FinFET designs. GAA transistors feature a channel completely surrounded by the gate, providing superior control over the transistor channel compared to FinFETs. This 3D structure allows for enhanced electrostatics and better gate control, significantly reducing leakage currents and improving overall transistor performance. GAA transistors also offer increased scalability, enabling manufacturers to continue the trend of miniaturization while maintaining or even improving transistor performance metrics.

Conclusion

In conclusion, the successful implementation of FinFET technology at 5nm nodes represents a pinnacle in semiconductor engineering, embodying the relentless pursuit of innovation and pushing the frontiers of what was once considered technically unattainable. The evolution to GAA transistors represents a strategic shift in semiconductor design, pushing the boundaries of transistor technology to meet the demands of increasingly complex and power-efficient electronic devices. As the semiconductor industry continues to embrace these advancements, the trajectory forward promises a future brimming with possibilities, setting the stage for continued breakthroughs and transformative developments in technology that will shape our digital world for years to come.

References:

1. Etienne Sicard, Lionel Trojman. Introducing 5-nm FinFET technology in Microwind. 2021. hal03254444f.
2. https://semiengineering.com/knowledge_centers/integrated-circuit/transistors/3d/finfet-3/
3. Fundamentals of Ultra-Thin-Body MOSFETs and FinFETs 1st Edition by Jerry G. Fossum (Author), Vishal P. Trivedi (Author)