KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES

ISSN (P): 2616-9592
ISSN (E): 2710-5075

Volume: 5 | Issue: 1 | Jan - Jun | 2022

KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES

Volume 5, Issue 1, 2022
ISSN (P): 2616-9592
ISSN (E): 2710-5075
Frequency Bi-Annual

Editorial Board

College of Computing & Information Sciences
Karachi Institute of Economics & Technology

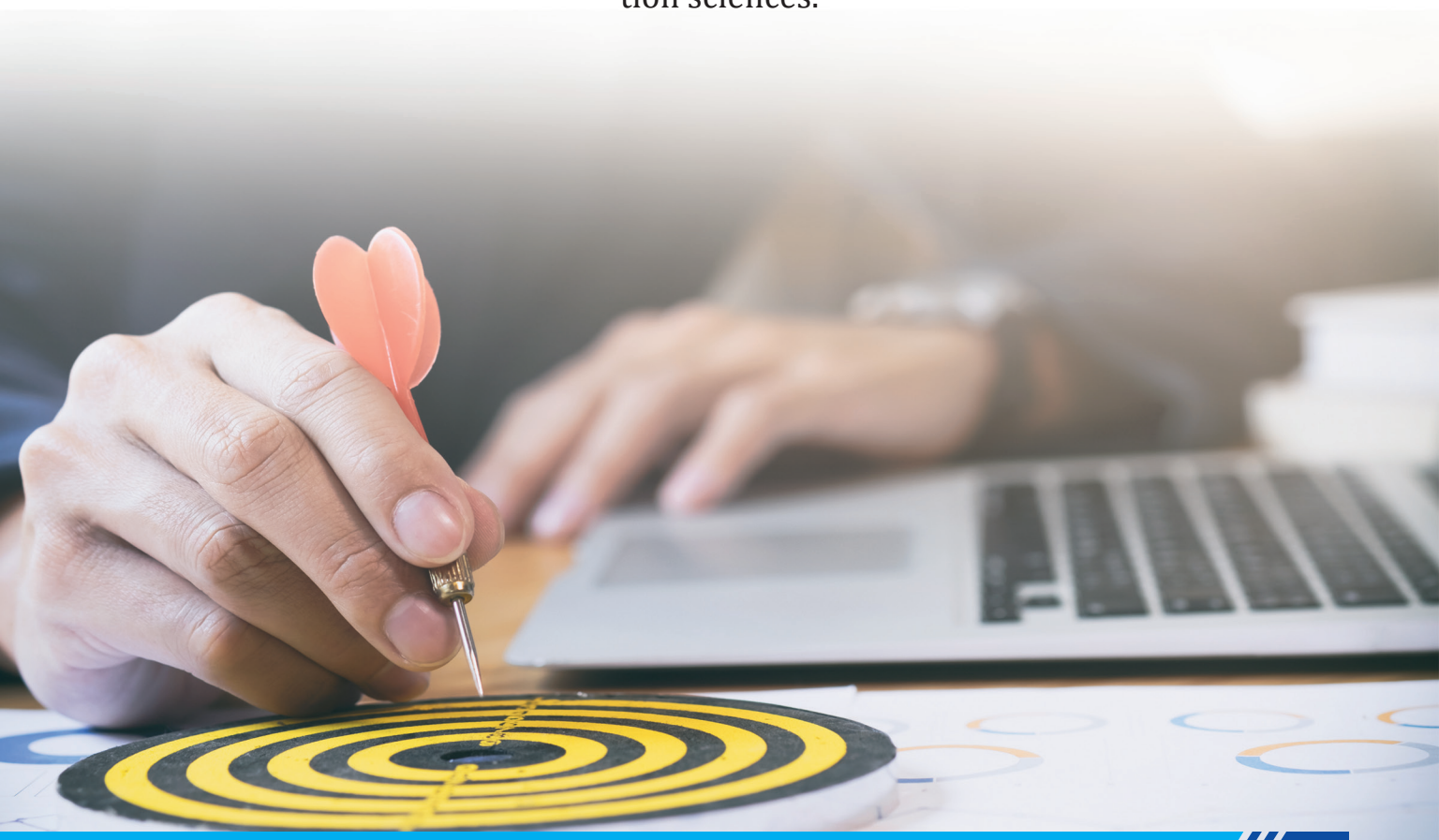
College of Computing & Information Sciences

Vision

To develop technology entrepreneurs & leaders for national & international market

Mission

To produce quality professionals by using diverse learning methodologies, aspiring faculty, innovative curriculum and cutting edge research, in the field of computing & information sciences.



AIMS AND SCOPE

KIET Journal of Computing and Information Sciences (KJCIS) is the bi-annual, multi-disciplinary research journal published by **College of Computing & Information Sciences (CoCIS)** at **Karachi Institute of Economics and Technology (KIET)**, Karachi, Pakistan. **KJCIS** aims to provide a panoramic view of the state of the art development in the field of computing and information sciences at global level.

It provides a premier interdisciplinary platform to researchers, scientists and practitioners from the field of computing and information sciences to share their findings and contribute to the knowledge domain at global level. The journal also fills the gap between academician and industrial research community.

KJCIS focused areas for publication includes; but not limited to:

- Data mining
- Big data
- Machine learning
- Artificial intelligence
- Mobile applications
- Computer networks
- Cryptography and information security
- Mobile and wireless communication
- Adhoc and body area networks
- Software engineering
- Speech and pattern recognition
- Evolutionary computation
- Semantic web and its application
- Data base technologies and its applications
- Internet of things (IoT)
- Computer vision
- Distributed computing
- Grid and cloud computing

OPEN ACCESS POLICY

For the benefit of authors and research community, this journal adopts open access policy, which means that the authors can self-archive their published articles on their own website or their institutional repositories. The readers can download or reuse any article free of charge for research, further study or any other non profitable academic activity.

PEER REVIEW POLICY

Peer review is the process to uphold the quality and validity of the published articles. KJCIS uses double-blind peer review policy to ensure only high-quality publications are selected for the journal. Papers are referred to at least two experts as suggested by the editorial board. All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. We expect our Board of Reviewing Editors and reviewers to treat manuscripts as confidential material. The identities of authors and reviewers remain confidential throughout the process.

COPYRIGHT

DISCLAIMER

The opinions expressed in **KIET Journal of Computing and Information Sciences (KJCIS)** are those of the authors and contributors, and do not necessarily reflect those of the journal management, advisory board and the editorial board. Papers published in KJCIS are processed through double blind peer-review by subject specialists and language experts. Neither the **CoCIS** nor the editors of **KJCIS** can be held responsible for errors or any consequences arising from the use of information contained in this journal, instead; errors should be reported directly to the corresponding authors of the articles.

Academic Editorial Board

Advisory Board

Table of Content

1 01-14 — An efficient Image Processing Technique to Measure and Align Vehicle Wheel Cylinder with Cloud Management System
Shoaib Zaidi, Muhammad Wasim, Lubaid Ahmed, Nauman Ahmed, Mahad Ahmed Khan and Muhammad Usman

2 15-36 — Automatic Taxonomy Generation and Incremental Evolution on Apache Spark Parallelization Framework
Kanwal Aalijah, Rabia Irfan, Umara Umar and Sanam Nayab

3 37-47 — An Ontology Based Approach to Search Woman Clothing from Pakistan's Top Clothing Brands
Shabina Mushtaque, Adnan Ahmed Siddiqui and Muhammad Wasim

4 48-61 — Risk Assessment Approach for Software Development using Cause and Effect Analysis
Abdur Rehman Riaz and Syed Mushhad M. Gilani

5 62-74 — Facial Expression Recognition Using Weighted Distance Transform
Syed Muhammad Rafi, Shahzad Nasim, Sheikh Muhammad Munaf, Syed Hassan Ali and Mohsin Khan

6 75-90 — EEG-Based BCI for Attention Assessment in E-Learning Environment using SVM
Muhammad Bilal, Muhammad Marouf, Safdar Rizvi Fatima Bashir, Muhammad Shahzad, Jawad and Ahmed Bhutta

https://doi.org/10.51153/kjcis.v5i1.90

An efficient Image Processing Technique to Measure and Align Vehicle Wheel Cylinder with Cloud Management System

Shoaib Zaidi[1] Muhammad Wasim[2] Lubaid Ahmed[3] Nauman Ahmed[4]

Mahad Ahmed Khan[5] Muhammad Usman[6]

Abstract

The measurement and alignment of vehicle wheel cylinders (motorcycle wheel hub) is an important and more challenging task for manufacturing companies. Currently in most of the local industries this measurement system is manual, and technicians are using screw gauges and vernier calipers for the measurement of cylinder diameters. There are some issues associated with the manual system to measure the different cylinder diameters of the wheel hub. Some very common issues are time consuming, human error can impact the accuracy of measurements, least count values are changed in different tools, which are used in measurements, and it requires more concentration for converting the decimal places and one of a very important issue is the alignment of centroids for different diameters of wheel cylinder. This centroid problem would never be fixed in a manual system, and it creates a big issue for the alignment of the wheel as well, that causes wobble in the wheel. The automated sensor-based system can resolve these issues and especially centroid issues with accurate measurements of cylinder diameters, but this system is very costly. The proposed system provides a state-of-the-art solution to measure the diameters of the cylinder with the accurate alignment of centroids. The work presented here consists of two modules— an automated vehicle wheel hub measurement and alignment system (VWMAS) using image processing techniques and cloud management. The proposed system is a low cost and effective technique, which resolves the issue of centroid with accurate measurement of diameters of different circles found in the hub with the accuracy (95%) and precision (100%).

Keyword: motorcycle, automated, vehicle wheel, centroid, bubbling, screw gauge, vernier caliper, measurement, diameter, accuracy, precision

[1]*Usman Institute of Technology, Karachi l szaidi@uit.edu*
[2]*Usman Institute of Technology, Karachi l mwaseem@uit.edu*
[3]*Usman Institute of Technology, Karachi l lahmed@uit.edu*
[4]*Usman Institute of Technology, Karachi |naumanahmed449@gmail.com*
[5]*Usman Institute of Technology, Karachi |mahadahmedk98@gmail.com*
[6]*Usman Institute of Technology, Karachi |raousmanhafeez@gmail.com*
**Corresponding author: sohail.iqbal@seecs.nust.edu.pk*

1. Introduction

Motorcycle industry in Pakistan is now one of the fastest growing industries. In the financial year 2020 record motorcycle production and sales is recorded in Pakistan [1-4].

As per the published report of Association of Pakistan Motorcycle Assemblers (APMA), there are around seventy-two motorcycle companies who are registered and actively participating in manufacturing, import and assembling of the motorcycles. Karachi is one of the most important centers to manufacture and produce different physical parts of two-wheel vehicles but the production cost in Karachi is very high compared to the other cities [1]. In late 90s, one of the significant motorcycle companies started the assembly and manufacturing of two-wheel vehicles in Pakistan through a mutual venture with the Yamaha, a world-renowned Japanese company and then other companies opted for the same model. Fateh Hero is one of a leading motorcycle company who produced an average of 21,778.500 Units from Jun 2006 to 2017 in Pakistan. The published data of Fateh Hero during the mentioned period is given here in table 1 [5].

Table 1: Motorcycle Production Rate

Duration (Years)	Production (Unit)
2006	34018.00
2007	25798.00
2008	22519.00
2009	21038.00
2010	35010.00
2011	41972.00
2012	38834.00
2013	20466.00
2014	11525.00
2015	8607.00
2016	2958.00
2017	3012.00

Few years back, different parts of motorcycles were imported and then assembled in local industries but now most of the parts are produced and assembled in the local industries in Pakistan. Some very common locally produced and developed motorcycle parts are listed here and shown in figure1.

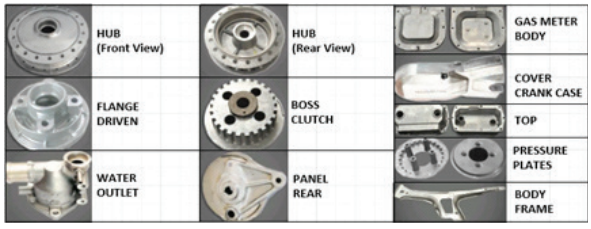


Figure 1: Parts Manufacturing in Local Industries

In most of the industry, the physical inspections and measurements of parts is carried out manually. Numbers of technicians are trained and they use basic tools to measure the components as per the standard engineering drawings. This way of measurement requires more time and there is a higher chance of human error, which can affect the overall accuracy and precision of the components performance.

The work presented here is divided into two key modules, first is related to the automated motorcycle wheel hub measurements and alignment of centroids, which is an important part used to mount the front and rear wheels of motorcycle accurately, secondly the cloud management system, which helps the management to monitor and track the records of good and bad cylinder measurements. In this paper, an automated measurement of wheel HUB is presented by using image processing technique.

2. Related Work

Various authors have published a number of research papers related to motorcycle parts design and manufacturing. This domain is now more attractive and challenging for the industries as well for the researchers. Some key contributions from the researchers are listed here: Ahmed et al. [6], Hussain et al. [7], Niazi et al. [8], Batra [9], Vasuvanich et al [10], Sayeed et al [11], Taneja et al. [12] and Nabi et al. [13], have reported about diverse methodologies, that can increase the volume of trade in the region of Asia. Authors of [14-15] discussed the image contour extraction using pixel-based comparison method. Authors of [16-19] discussed the background subtraction and image enhancement techniques. Authors of [20] describe the method of large-scale image retrieval, which is efficient in time with better efficiency. In [21], authors discussed a 3-D layout estimation technique in a more efficient way. Authors of [22] elaborate the efficient way to image denoising from the image. This way is really help to understand the methods to reduce noise from the captured image. It improves the efficiency of proposed technique.

3. Proposed Method

A. Model Development

The presented work is divided into two key modules, first is related to the automated

motorcycle wheel hub measurements and alignment of its centroids, which is an important part use to mount the front and rear wheels of motorcycle wheels/tire accurately. Second module of the cloud management system will help the management to monitor and track the records of good and bad cylinder measurements of component (HUB).

a. Automated Motorcycle Wheel HUB Measurement and Alignment

This is the first module of the proposed system. Motorcycle wheel hub is an important part of motorcycle to mount the wheel as shown in figure2. In most of the companies, the measurements of wheel hub are based on manual system. Technicians are uses screw gauge and vernier calipers to measure the inner diameters of circular cylinders.

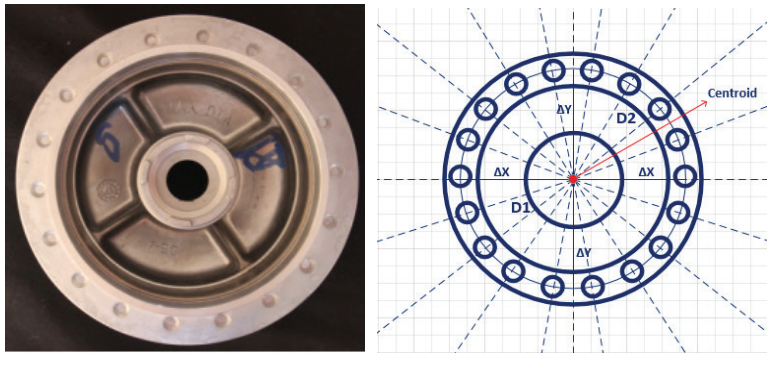


Figure 2: Hub Rear View

As per the standard engineering drawing, the measurements are as follows:

D_1= Diameter of small circle = $(35\text{-}0.5) \leq D_1 \leq (35+0.025)$
D_2= Diameter of large circle = $110 \leq D_1 \leq (110+0.20)$

As far as the alignments of centroids are concerned, it is not possible in the manual system to analyze and measure accurately [23-26]. This is very critical and big issue for the industries. Using the techniques developed in the proposed system, the measurements of diameters and issues of centroid are solved with a very high accuracy and precision.

The system, which is developed in Image Processing Research Laboratory (IPRL) [27] is consists of a wooden base with black colored background. A 'v-shaped' space is created on wooden base to fix the position of wheel hub. A camera was mounted on tripod stand on the top of hub rear face to capture the image. This captured hub rear face then measured and analyzed using the developed application program. The developed application generated the reports about the measurements of diameters and the position of centers of all circles, which needs to be properly aligned. When all the measurements

and centroid are well aligned as per the standard values, green LED will be ON, which indicates the tested wheel hub is 'good', otherwise it will be 'bad' and red LED will be ON. A Raspberry pi module is programmed and used for these indications of good and bad cylinders. The complete hardware set up is shown in figure3.

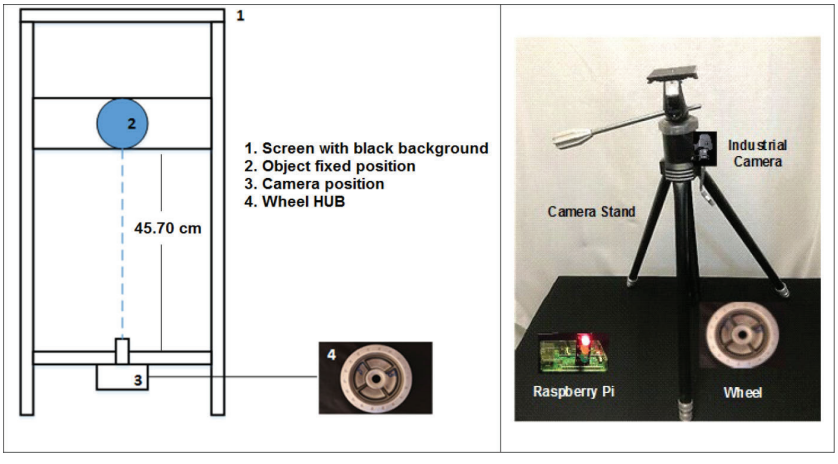


Figure 3: Hardware Set-up

To resolve the issue of centroid multiple radii were calculated for small and large circles in the captured image. All these radii for small and large circles were calculated from a common center red point as shown in figure 3. The system ensures that all the measurement of radii for small circle is equal in magnitude and same for the larger circle. The centroid issue has been resolved using this automated wheel hub measurement and alignment system. The system is capable of identifying the good and bad hubs. Good hub means the drum having valid measurement of the diameter with a well aligned centroid while the bad hub refers to the drum having invalid measurement of diameters or non-aligned centroid [28-33]. All the measurements of diameters are measured and recorded in centimeters (cm) in the system.

b. Cloud Management

This is the second module of the developed system, which is related to the cloud management and monitoring of complete environment [34-38]. All the measurements conducted by the technicians are directly stored in the cloud database[39-40]. Technician just place the wheel hub at the specified positions in hardware set-up and capture the image of rear face of hub by viewing the live streaming on the camera which is connected to the wifi network. The application program read the image and measured the values of diameters from the circular shaped cylinders and validates the centroid of all circles. The decisions about the good and bad measurement are shown using the red and green LEDs controlled by Arduino processor.

Administrator can view all the detail of stored measurement along with the detail of technician IDs, technician shift timing (data is filtered by the day, month and year), status of good and bad cylinder measurements and can generate the reports as well. The complete system flow diagram is shown in figure4.

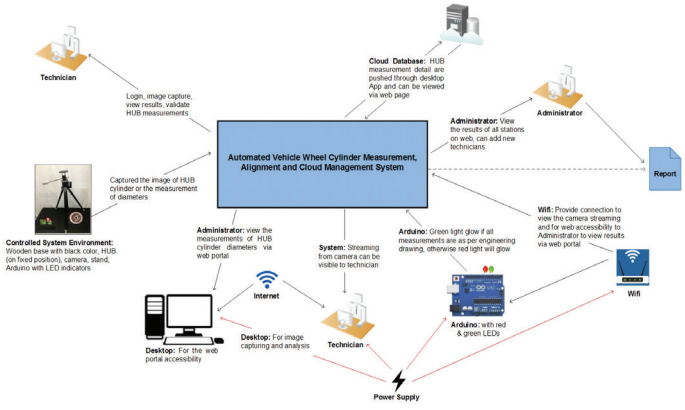


Figure 4: Proposed System FlowDiagram

c. System Application - Explanation of system functions

A desktop application is developed for the technician who can capture the images of the wheel hub (rear face), process the captured image using image processing techniques and results of measurements can be viewed in real time. The admin can view and change the detail of measurements using web portal. The Technician and The admin interfaces in developed application are shown in figure 5.

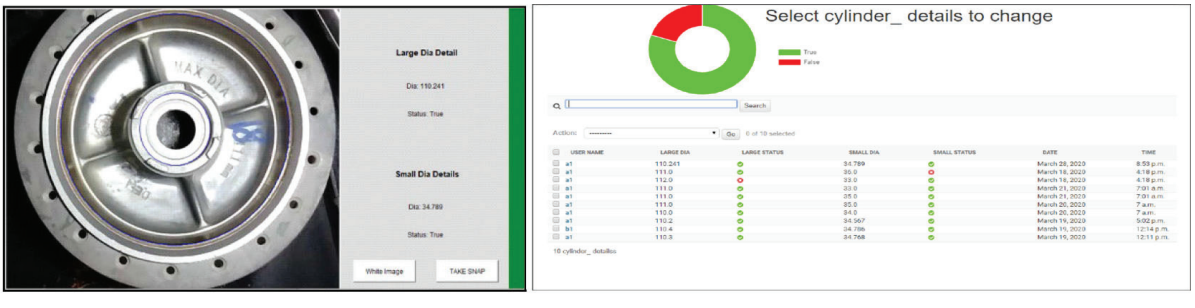
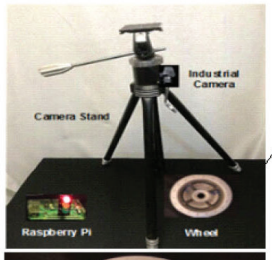
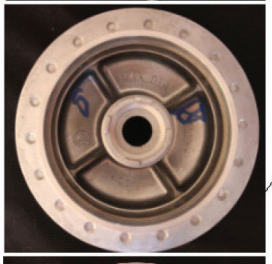


Figure 5: Technician (LH) and Admin (RH) Interface in Application

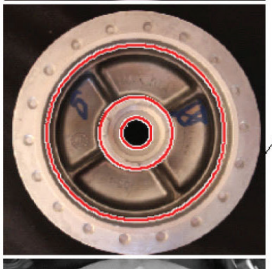
In the following figure 6, image processing techniques and the steps involved to measure the diameters and to find centroid of circles are shown.



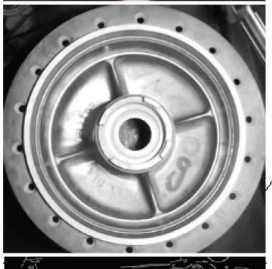
Step 1: Motorcycle wheel hub place at its fix position of setup to capture the hub rear face image.



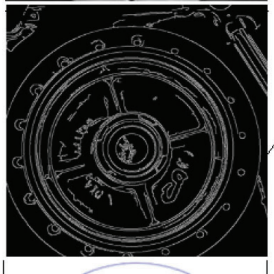
Step 2: Colored wheel hub image is captured using the digital camera, which is mounted of tripod stand. Image then cropped for further processes.



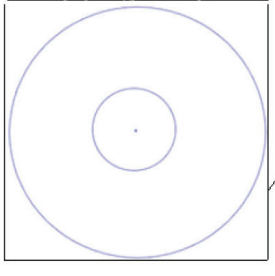
Step 3: Trace all possible circular patterns using image processing techniques.



Step 4: The captured image then converted into gray scale image for better processing and to trace the contours in image.



Step 5: Trace all possible contours from the gray scale wheel hub image.



Step 6: After applying filtering techniques to last image of step 4, diameters $D1$ and $D2$ extracted and remove all other contours from the image.

Fast Finding and Fitting (FFF) Algorithm:

Fast Finding and Fitting (FFF) algorithm was used for multiple circles detection. This algorithm uses Hough transformation.

The Hough Transform (HT):

Hough Transformations along with their extensions are commonly used in multiple circles detection. HT is still one of the most effective techniques because of its high capabilities to remove the noise from the images.

Circle detection using Fast Finding and Fitting (FFF) algorithm is uses the mechanism of the genetic algorithm. The FFF is very efficient and accurate as compared to the other algorithms[41].

Pseudo Code

The pseudo code of the proposed system is given as below:

1. Input the image
2. Apply HT transformation and extract all possible vertical symmetrical axis
3. Locate the pixels of circular shape
4. Measure the diameters of smaller and larger circles and locate the centers of circles to resolve centroid issues.

Figure 6: Steps of Proposed Technique

4. Experimental Results

Table 1 shows the measurement of wheel Hub diameters for both small and large circles. During the manual observations of diameter, two critical values were found for wheel hub 6 and wheel hub 7. For wheel hub 6, the manual value was rejected but in case of system generated values, it was accepted it was found that the wheel hub no. 6 was correct but there was a mistake to observed value via manual system. As for as wheel hub 7 is concerned, manual value was accepted but it was at the maximum allowed limit, which is a critical condition. In case of system generated value, wheel hub 7 was strongly accepted. Wheel hub 6 has centroid issue and it was observed through the proposed technique.

Table 2. Manual and System Generated Readings

S. No	Manual Readings of small circle D_1 (cm)	System Readings of small circle D'_1 (cm)	Manual Readings of large circle D_2(cm)	System Readings of large circle D'_2(cm)
Wheel Hub1	34.93	34.93	110.08	110.03
Wheel Hub 2	34.94	34.93	110.16	110.11
Wheel Hub 3	34.93	34.94	110.06	110.14
Wheel Hub 4	34.93	34.92	110.10	110.18
Wheel Hub 5	34.93	34.93	110.18	110.31
Wheel Hub 6	34.92	34.93	110.22	110.16
Wheel Hub 7	34.92	34.92	110.20	110.13
Wheel Hub 8	34.93	34.93	110.06	110.12
Wheel Hub 9	34.94	34.93	110.16	110.28
Wheel Hub 10	34.94	34.93	110.16	110.22

Table 2 shows the observations of small and large circle diameters and the change between two manual and system generated values. It is very clear in observation that deviations between manual and system generated values are very small and within the range. So all system generated values are highly accepted.

Table 3. Observations of Small and Large Circles

Observation of Small Circle with difference of Manual and System Generated Values			Observation of Large Circle with difference of Manual and System Generated Values		
Manual Readings D_1 (cm)	System Readings D'_1 (cm)	Deviation of Diameters $\Delta D_1=D_1-D'_1$	Manual Readings D_2 (cm)	System Readings D'_2 (cm)	Deviation of Diameters $\Delta D_2=D_2-D'_2$
34.93	34.93	0.00	110.08	110.05	0.03
34.94	34.93	0.01	110.16	110.11	0.05
34.93	34.94	-0.01	110.06	110.04	0.02
34.93	34.92	0.01	110.10	110.18	-0.08
34.93	34.93	0.00	110.18	110.11	0.07
34.92	34.93	-0.01	110.22	110.16	0.06
34.92	34.92	0.00	110.20	110.13	0.07
34.93	34.93	0.00	110.06	110.12	-0.06
34.94	34.93	0.01	110.16	110.18	-0.02
34.94	34.93	0.01	110.16	110.18	-0.02

Table 3 shows the accuracy, precision, sensitivity and specificity of the proposed automated wheel hub measurement systems. In our test run, 40 samples have been tested using edge detection techniques. For edge detection technique, 38 samples of drum were correctly recognized as true Positive. Table 4 shows the facts recorded during test runs for edge detection systems.

Table 4. Key parameters and performance of system

Parameter	Edge Detection	Parameter	Performance (%)
Number of true positive (TP)	38	Accuracy= $\frac{(TP+TN)}{(TP+FP+FN+TN)}$	95
Number of true negative (TN)	0	Precision= $\frac{TP}{(TP+FP)}$	100
Number of false positive (FP)	0	Senitivity= $\frac{TP}{(TP+FN)}$	95
Number of false negative (FN)	2	Specificity= $\frac{TN}{(FP+TN)}$	0

A. Data Consistency

The data consistency of the system can be analyzed using statistical technique. The coefficient of variationfor manual and system generated values is used to analyze the data consistency.

a. Coefficient of Variation for Small Circle

$$\textit{Coefficient of Variation} = \frac{\sigma}{\mu} \times 100 \quad \ldots\ldots\ldots\ldots\ldots (1)$$

σ = Standard Deviation = $\sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ ..... (2)

μ = Mean = $\frac{\sum x_i}{N}$ ..... (3)

After solving the parameters, we have

$$\textit{Coefficient of Variation (Manual)} = \frac{\sigma}{\mu} \times 100 = \frac{0.007}{34.931} \times 100 = 0.020\% \ldots\ldots\ldots (4)$$

$$\textit{Coefficient of Variation (System)} = \frac{\sigma}{\mu} \times 100 = \frac{0.00538}{34.929} \times 100 = 0.015\% \ldots\ldots\ldots (5)$$

The result shows the system generated values are more consistent as compare to manual values.

b. Coefficient of Variation for Large Circle

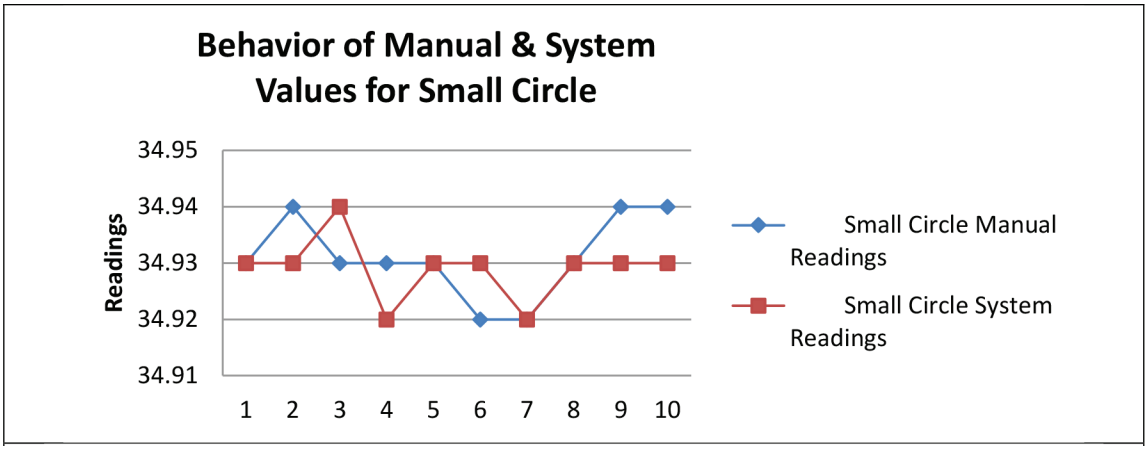
$$\textit{Coefficient of Variation (Manual)} = \frac{\sigma}{\mu} \times 100 = \frac{0.00050}{110.138} \times 100 = 0.00050\% \ldots\ldots\ldots (6)$$

$$\textit{Coefficient of Variation (System)} = \frac{\sigma}{\mu} \times 100 = \frac{0.00044}{110.126} \times 100 = 0.00044\% \ldots\ldots\ldots (7)$$

The result shows the system generated values are more consistent as compare to manual values.

c. Graphical Representation of Manual and System Values

A graphical representation of manual and system generated values are shown in figure7.



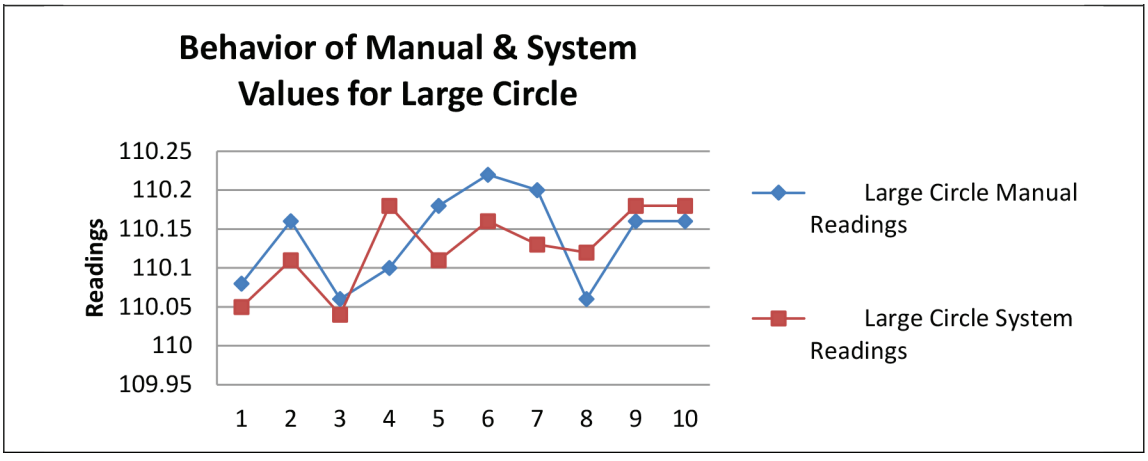


Figure7: Behavior of Manual and System Values for Small and Large Circles

5. Conclusion

The measurement done by proposed system is more accurate as compare to the manual system. The throughput of automated system is higher compared to the manual system. The proposed system has an acceptable range of accuracy, which shows the performance of the proposed system. As in Pakistan, a very large number of motorcycle companies are working so this automated measurement system can be extended for other components of vehicle parts of motorcycles as well.

Acknowledgment

The corresponding author of this paper would like to thank Director, Usman Institute of Technology (UIT), for making available infrastructure and financial support to complete this work. All the test runs and experimental results are obtained in Image Processing Research Laboratory (IPRL) at UIT. Authors are thankful to Dr. Abid Karim (Head Research Committee) and Dr. Talha Ahsan (Head, Department of Electrical Engineering), who provided technical support and guidance. In the last the corresponding author would like to thanks Mr. Pakash Lohana (Head, Department of Computer Science) to provide all departmental support when needed.

References

[1] Imran, Muhammad, and Aaiza Khan. "The Automotive Industry in Pakistan: Structure, Composition and Assessment of Competitiveness with India." Industry and Innovation, Forthcoming, 2015.

[2] Felipe, J. A Note on Competitiveness and Structural Transformation in Pakistan. Economic Working Papers, Asian Development Bank, 2007.

[3] https://www.marketresearch.com/MarketLine-v3883/Automotive-Manufacturing-Pakistan-13901084/, 2020.

[4] ADB, Pakistan: Private Sector Assessment, Asian Development Bank, Country Planning Documents, Manila, 2008.

[5] https://www.ceicdata.com/en/indicator/pakistan/motor-vehicle-production.

[6] Ahmed, Vaqar, and Samavia Batool. "India–Pakistan Trade: Perspectives from the Automobile Sector in Pakistan." India-Pakistan Trade Normalisation. Springer, Singapore, pp. 129-161, 2017.

[7] Husain, Ishrat. Prospects and challenges for increasing India-Pakistan trade. Atlantic Council, 2011.

[8] Pasha, H., and Ismail, Z. "An Overview of Trends in the Automotive Sector and the Policy Framework."Automotive Sector in Pakistan Phase I Report, 2012.

[9] Batra, A. "India's Global Trade Potential: The Gravity Model Approach, Indian Council for Research on International Economic Relations, WP, No. 151, 2004.

[10] Vasuvanich, Saroge, et al. "The Role of Big Data Analytics in Determine the Relationship between Green Product Innovation, Market Demand and the Performance of Motorcycle Manufacturing Firms in Thailand." Int. J Sup. Chain. Mgt Vol 9.1, 37. 2020.

[11] Sayeed, Asad. "Gains from trade and structural impediments to India–Pakistan trade." Karachi, Pakistan: Collective for Social Science Research, 2005.

[12] Taneja, Nisha, et al. "Normalizing India-Pakistan Trade." India-Pakistan Trade. Springer, New Delhi, pp. 13-45, 2015.

[13] Nabi, I. Shaikh, H. Regional Trade Report. Pakistan Business Council Report, 2013.

[14] Rasche, C. "Rapid contour detection for image classification." IET Image Processing 12.4, pp. 532-538, 2017.

[15] Heath, M.D., Sarkar, S., Sanocki, T., et al.: 'A robust visual method for assessing the relative performance of edge-detection algorithms', IEEE Trans. Pattern Anal. Mach. Intell., 19, (12), pp. 1338–1359, 1997.

[16] Babaee, Mohammadreza, Duc Tung Dinh, and Gerhard Rigoll. "A deep convolutional neural network for video sequence background subtraction." Pattern Recognition 76, pp. 635-649, 2018.

[17] Hou, Lu, et al. "Internet of things cloud: Architecture and implementation." IEEE Communications Magazine 54.12, pp. 32-39, 2016.

[18] Rasouli, Mohammad Reza. "An architecture for IoT-enabled intelligent process-aware cloud production platform: a case study in a networked cloud clinical laboratory". International Journal of Production Research 58.12, pp. 3765-3780, 2020.

[19] Aazam, Mohammad, et al. "Cloud of Things: Integrating Internet of Things and cloud computing and the issues involved." Proceedings of 2014 11th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, 14th-18th January, 2014. IEEE, 2014.

[20] Yan, Chenggang, et al. "Deep multi-view enhancement hashing for image retrieval." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[21] Yan, Chenggang, et al. "3D room layout estimation from a single RGB image." IEEE Transactions on Multimedia 22.11 (2020): 3014-3024.

[22] Yan, Chenggang, et al. "Depth image denoising using nuclear norm and learning graph model." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16.4 (2020): 1-17.

[23] Davies, E. R. "The effect of noise on edge orientation computations." Pattern recognition letters 6.5, pp. 315-322, 1987.

[24] Shamuratov, Oleksii, et al. "The methods for Contour Analysis of Images." 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT). Vol. 2. IEEE, 2019.

[25] Maini, Raman, and Himanshu Aggarwal. "Study and comparison of various image edge detection techniques." International journal of image processing (IJIP) 3.1, pp. 1-11, 2009.

[26] Bao, Paul, Lei Zhang, and Xiaolin Wu. "Canny edge detection enhancement by scale multiplication." IEEE transactions on pattern analysis and machine intelligence 27.9, pp. 1485-1490, 2005.

[27] Wasim, Muhammad, et al. "A Comparative Evaluation of Dotted Raster-Stereography and Feature-Based Techniques for Automated Face Recognition." International Journal of Advanced Computer Science and Applications 9.6, pp. 276-283, 2018.

[28] O. E. Okman and G. B. Akar, "A circle detection approach based on Radon Transform," in Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on Vancouver, pp. 2119–2123, IEEE, 2013.

[29] Huang, T. Sasaki, H. Hashimoto, and F. Inoue, "Circle detection and fitting based positioning system using laser range finder," in System Integration (SII), IEEE/SICE International Symposium on Sendai, pp. 442–447, IEEE, 2010.

[30] Zheng, You-yi, Ji-lai Rao, and Lei Wu. "Edge detection methods in digital image processing." 2010 5th International Conference on Computer Science & Education. IEEE, 2010.

[31] Davies, E.R.: The effect of noise on edge orientation computations. Pattern Recognition Letters 6, pp. 315–322, 1987.

[32] Shamuratov, Oleksii, et al. "The methods for Contour Analysis of Images." 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT). Vol. 2. IEEE, 2019.

[33] Maini, Raman, and Himanshu Aggarwal. "Study and comparison of various image edge detection techniques." International journal of image processing (IJIP) 3.1,pp. 1-11, 2009.

[34] Kim, Chulyeon, et al. "A hybrid framework combining background subtraction and deep neural networks for rapid person detection." Journal of Big Data 5.1,pp. 22, 2018.

[35] Bouwmans, Thierry, et al. "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation." Neural Networks 117, pp. 8-66, 2019.

[36] Nandal, Amita, et al. "Image edge detection using fractional calculus with feature and contrast enhancement." Circuits, Systems, and Signal Processing 37.9, pp. 3946-3972, 2018.

[37] Odun-Ayo, Isaac, et al. "Cloud computing architecture: A critical analysis." 2018 18th International Conference on Computational Science and Applications (ICCSA). IEEE, 2018.

[38] Chen, Min, Francisco Herrera, and Kai Hwang. "Cognitive computing: architecture, technologies and intelligent applications." IEEE Access 6,pp. 19774-19783, 2018.

[39] Malhotra, Shweta, et al. "Generalized query processing mechanism in cloud database management system." Big data analytics. Springer, Singapore, pp. 641-648, 2018.

[40] Zhang, Ji, et al. "An end-to-end automatic cloud database tuning system using deep reinforcement learning." Proceedings of the 2019 International Conference on Management of Data. 2019.

[41] Yi, Wei, and S. Marshall. "Circle detection using Fast Finding and Fitting (FFF) algorithm." Geo-spatial Information Science 3.1,pp. 74-78, 2000.

https://doi.org/10.51153/kjcis.v5i1.83

Automatic Taxonomy Generation and Incremental Evolution on Apache Spark Parallelization Framework

Umara Umar[1] Kanwal Aalijah[2] Rabia Irfan[3] Sanam Nayab[4]

Abstract

The term "Big Data" refers to a large volume of information usually in terabytes and petabytes. It includes both structured and unstructured data. Unstructured data is conventionally text-heavy, but may also contain data such as facts, dates, and numbers. To use this unstructured information effectively, it needs to be processed and organized. Taxonomy is considered a powerful way of organizing information. For automatic taxonomy generation, various techniques have been proposed in the past. However, the substantial nature of big data presently crosses the processing abilities of traditional techniques. Thus, to meet this challenge an extensible and scalable technique is required to potentially accelerate the process of taxonomy generation and its evolution upon arrival of new data, hence catering to a large amount of unstructured big data. This paper proposes a technique for both the taxonomy generation and evolution of Apache Spark infrastructure. The proposed technique is evaluated on a text dataset from a computing domain. The evaluation results show that the technique presented in this paper outperformed the existing techniques in terms of time and quality metrics. The time and quality-based evaluation showed that the use of the MapReduce environment has resolved the scalability issues of the current taxonomy generation and evolution process.

Keywords: Big Data, Apache Spark, Unstructured Data, Taxonomy, Map-Reduce, Hadoop, Scalable

1. Introduction

During the past two decades, communication using electronic media has acquired extreme popularity and has gained a significant role in developed societies. Electronic media provides several services such as the World Wide Web (WWW), mobile devices, Internet of Things (IoT)-based devices, social networks, etc. This era is marked by the circulation of the intense amount of data (in petabytes and zettabytes) across the globe. This large volume of data produced from various sources [1] can be both structured and unstructured. This bulk of data is called Big Data-A Technology Giant [2], [4], [5]. The

[1] *Umara Umar l uumar.msit19seecs@seecs.edu.pk*
[2] *Kanwal Aalijah l ksair.mscs8seecs@seecs.edu.pk*
[3] *Rabia Irfan l rabia.irfan@seecs.edu.pk*
[4] *Sanam Nayab l snayab.msit19seecs@seecs.edu.pk*

5V's of big data – volume, velocity, variety, veracity, and value make data management and analytics challenging for the conventional data warehouses. Big data, which is unstructured, is the data with no standard formatting [3] and no definite structure as the name shows. In order to draw useful information from this data, it should be managed, processed, and effectively transformed. In other words, this data needs to be organized into a structured form, like taxonomy.

Taxonomy is a hierarchical structure that organizes the given data in parent-child relationships, based on the inherent concepts present in the data [6]. Taxonomy is an efficient and effective way of organizing and classifying data [7] that also provides standardization in case of the exchange of information. Taxonomy also provides an infrastructure for knowledge management [8]. Taxonomy arranges information in a hierarchical structure that makes navigation and searching for information easier [9] [10].

Automatic taxonomy generation has two types i.e. (1) Incremental (2) Non-incremental. Non-incremental taxonomy generation rebuilds the taxonomy from very scratch on the entry of new documents into the current system. Kashyap et al [11] proposed an innovative method for taxonomy generation that uses the Principal Direction Divisive Partitioning (PDDP) approach [12] to generate taxonomy. Anke et al. [13] suggested a conditional random field classifier for taxonomy generation. Velerdi et al. [14] presented a graph-based method for taxonomy generation. All these techniques successfully resulted in taxonomy generation, but upon the intervention of the new documents into the system, these techniques regenerate the taxonomy from the very basis to get the taxonomy updated, consequently producing non-incremental taxonomy architecture. This approach is very time-consuming when a large dataset is involved. So, there was an extreme need for a technique that generates taxonomy on top of the current taxonomy on the arrival of the new document into the system. This process is known as "Incremental/ Progressive Taxonomy Generation" or may be named as "Taxonomy Evolution".

There are rare techniques which have focused on incremental taxonomy generation like [15], AdaptTaxa [16], IHTCTaxa [17], TIE [18]. The methodology EvoTaxa [15] is especially developed for tagged data. AdaptTaxa [16] focuses on incremental taxonomy generation technique for unstructured textual data. It adopts a supervised approach that requires training data. The technique IHTCTaxa [17] uses an unsupervised hierarchical clustering-based approach by adjusting the newly introduced documents. TIE [18] is an incremental taxonomy generation algorithm that updates taxonomy upon the entry of new documents. All these techniques, be it non-incremental or incremental, provide a more or less good quality taxonomy, however, lacks the focus on rapidly increasing, voluminous big data. With the emerging trend of big data and cloud computing, the data is being produced from varying sources as well as being stored and processed electronically and automatically [19][20].

In the realm of big data, we are always in search of certain techniques and algorithms which prove to be dependable and scalable to negotiate with the varying kind of data. Some progress has already been achieved in the field of hierarchical clustering for huge datasets, such as [21] and [22]. The scope of these studies was limited to hierarchical clustering and they did not adequately concentrate on the idea of taxonomy generation and evolution. Besides, none of these techniques addressed the concept of parallelization for developing a scalable and efficient algorithm for generating and evolving taxonomy.

A new technique has been devised in our work [37] for the taxonomy generation and incremental evolution comprised of the MapReduce paradigm incorporating Apache Spark. MapReduce is capable of minimizing time by parallel data processing. Fault tolerance is also being provided by capitalizing on a distributed file system [19]. MapReduce environment can improve the scalability issues of present taxonomy generation and evolution methodologies. However, our previous work didn't focus on the evaluation of the proposed technique with respect to the parallelization framework, thus, we were not be able to figure out the essence of achieving scalability previously. This paper particularly focuses on this aspect.

The major problem with existing taxonomy generation algorithms was the amount of data it can process. Our algorithm processes the data in a parallel fashion in small chunks applying HAC on each chunk of data. That is where map-reduce comes in. The principle behind map-reduce is you divide the tasks into smaller tasks and then combine them. Exactly in the same fashion, we are making small taxonomies on each chunk of data and once all those taxonomies are made, they are combined. We use HAC on the spark engine which at the backend uses map-reduce to perform HAC.

In our research, we have made the following contributions:

1. The proposed technique provides us a solution for taxonomy generation and evolution in a considerably limited span of time in comparison with the existing techniques, thereby making taxonomy utilization more effective.
2. As clustering is the base of the adopted taxonomy generation algorithm, the clustering quality of taxonomy generated from the proposed methodology is compared and evaluated with the clustering quality of taxonomy generated using the existing taxonomy generation techniques. According to Silhouette's score and Davies Bouldin's score, the clustering quality of the proposed methodology is higher than the present techniques.
3. Zero or no similarity of a document with the current clusters case is being addressed.
4. For the case of evolution, the application of Newick tree graph facilitates the technique to incorporate even a graph-based taxonomy instead of just clustering-based taxonomy.

The salient features of the remaining part of this article are as follows:
Succeeding the Introduction in Section I, Section II discusses the Literature Review. Literature review elaborates the existing techniques for non-incremental and incremental taxonomy generation in detail. This section also throws light on the basic taxonomy generation process that has been used by the existing techniques. Section III presents the background and discusses the preface of big data techniques and tools used in this research work. Section IV explains the proposed technique devised for the processes of taxonomy generation and taxonomy evolution. Section V compares the proposed methodology with the current non-incremental and incremental taxonomy generation techniques and tests the scalability of the proposed technique. Finally, Section VI summarizes the Conclusion and Future Work.

2. Literature Review

This section describes the automatic taxonomy generation in detail. Automatic taxonomy generation process consists of two types: incremental and non-incremental. Be it as a non-incremental taxonomy or an incremental taxonomy generation process, a basic taxonomy generation algorithm is used in order to build the initial taxonomy in both cases. The commonly used steps of taxonomy generation are: data preprocessing, data modeling, hierarchy formation and node labeling. Different works have used different approaches in order to perform these steps. In general, taxonomy generation algorithms first cleanse the data using preprocessing that includes the removal of unnecessary details from the data. Once the data is preprocessed, it is then modeled to bring into a computational form. Using the modeled data, hierarchical relationships are produced, organized, and then labeled to obtain a structure in a hierarchical form of taxonomy.

Non-incremental type of taxonomy generation procedure utilizes the basic process of a taxonomy generation to generate taxonomy and the process runs every time when the newly arriving documents are presented into the system. The work TaxGen [23] presented an automatic taxonomy generation algorithm for unstructured data. The algorithm uses hierarchical clustering algorithm (HCA) for building the underlying structure for taxonomy generation. TaxaMiner [11] was also an addition in the pool of existing non-incremental taxonomy generation techniques. The cluster cohesion is used to extract the taxonomy among the successive levels of the hierarchical clustering tree. TaxoLearn [24] is also a non-incremental taxonomy generation algorithm. In this work, taxonomy hierarchy is built using an unsupervised hierarchical clustering algorithm [25]. On the other hand, an incremental taxonomy generation or taxonomy evolution technique works in a fashion that in-occurrence of the new documents in the system the process does not re-build the entire taxonomy from scratch; instead of that, the new documents are presented in the current taxonomy based upon the similarities with the existing dataset.

AdaptTaxa [16] generates taxonomy incrementally for group profiling problem. EvoTaxa [15] generated taxonomy incrementally for particularly large collection of tags. In this technique, a graph called association rules graph is produced. In an association rules graph, the vertices are tags and based on support and confidence values these tags are connected. Manipulation on the association rules graph is done by taxonomy extraction step. Only those associations are kept which don't add to noisy associations. The technique successfully generated and evolved taxonomy but it does so for tag data only. IHTCTaxa [17] uses unsupervised incremental hierarchical clustering approach to generate taxonomy for unstructured textual data. IHTC (Incremental Hierarchical Term Clustering) algorithm considers the problem of hierarchical clustering as online in contrary to the batch mode non-incremental hierarchical clustering, like HAC [26] and Bisect K-means [27].

TIE [18] algorithm was as advancement in the domain of incremental taxonomy generation. The TIE algorithm takes as an input the following: 1) existing taxonomy 2) respective hierarchical structure (i.e., clusters hierarchical structure) 3) new documents. The nearest cluster of new arriving document is recognized based upon the similarity score. The similarity score range may well identify the level of impact that a new arriving document has on its closest or nearby cluster. For the level of impact, to accommodate the new documents in a current hierarchical structure most of the reorganization operators came into practice. Hence, the current taxonomy develops to identify the change take place in the data [18]. In short, it was observed that the majority of the available non-incremental or incremental taxonomy generation approaches produce the good worth taxonomy. But, these approaches may lack attention on speedily expanding, voluminous and varying natured big data.

Furthermore, it was observed from the analysis of the literature that underlying technique for building a hierarchical structure in a taxonomy generation or evolution technique is mostly clustering-based [28]. Clustering techniques are very useful tools in case an unstructured data needs to be organized in a hierarchy [29]. In our work, we, particularly focus on clustering-based incremental taxonomy generation techniques. However, new challenges of big data make it difficult to apply conventional clustering techniques. Large data volume and time complexity of clustering algorithms lead to the problem of efficient deployment of clustering algorithms for big data to get an outcome in a reasonable amount of time.

Clustering algorithms dealing with big data are generally classified into categories as [30]: partitioning-based clustering approaches, hierarchical clustering approaches, grid-based clustering approaches and model-based clustering approaches. All these techniques have their own advantages and disadvantages. Partitioning-based clustering technique has a disadvantage that it requires a pre-defined value of K parameter to be given by a user.

For a clustering solution the value of K is often non-deterministic [31]. In a hierarchical clustering technique once a stage is completed it cannot be un-done. All the hierarchical clustering algorithms have the limitation stated above [32].

Density-based clustering algorithms contain noisy objects because they work in such a way in which clusters are described as dense areas separated by low density regions [32], therefore, not considered appropriate for very huge size datasets. Clustering algorithms that are based on a model are slow and unsuitable for very large dataset for a classification problem as they utilize the multivariate probability distribution. The grid size is usually far smaller than the database size. In case of highly irregular data distributions, using a single uniform grid might not be a good idea as a single uniform grid will fail to provide the required clustering quality and also is not able to fulfill the required time requirement [31].

Moreover, clustering techniques for big data mentioned here are specifically designed for dealing with big data but to be run on a single machine. New challenges of big data can be solved using multiple machines clustering techniques that can be able to achieve results in a much smaller time. Such parallel algorithms divide the data into various smaller data partitions and distribute them on different machines. This makes the overall running time of the algorithm smaller and increases its scalability. MapReduce algorithm is a task partitioning algorithm designed for distributed execution of a task on many servers which gives a good base for the implementation of such parallel forms of algorithms for data clustering. To understand its working, the next section discusses the MapReduce environment and tools used for big data processing.

3. Background and Preliminaries

This section discusses prominent tools in the world of big data processing: Apache Hadoop and Apache Spark which are based on MapReduce paradigm.

A. MapReduce

Researchers at Google presented a new programming model called MapReduce [33], which was able to solve the challenges of efficient processing of massive datasets using large clusters. MapReduce solves the problems faced in parallelizing the data across the individual machine's clusters [33]. MapReduce gives an easy and simple model for distributed computing by solving the problems of data partition, scheduling of machine failure and decreasing inter-machine communications. MapReduce is a programming paradigm that works by decomposing the problem into multiple map and reduce tasks. An Input is inserted in the form of key or value pairs to the mapper function. This key value pair input is then passed to reducer which then gives it as an input to the reduce

function. Associated with the intermediate key, the reducer merges the intermediate values and finally produced a combined output.
In real world scenarios, several map reduce functions can be applied on various machines individually in order to achieve parallelization. Apache Hadoop and Apache Spark are prominent big data processing environments that uses MapReduce algorithm for processing and analyzing the data [33], which are discussed in the succeeding subsections.

B. Apache Hadoop

Hadoop is a software framework based on MapReduce algorithm. The framework can write applications that can handle huge size of data in-parallel over the large size of clusters. The size of the data to be processed is in multi-terabyte. The size of the cluster is of thousands of nodes. Hadoop provides efficient, reliable and fault tolerance system for processing of big data. There are many different tools and products in Hadoop ecosystem.

The two important components of Hadoop ecosystem are HDFS and YARN. Hadoop Distributed File System [34], commonly referred as HDFS, is a single reliable file system. HDFS is reliable file system as it offers the monitoring of failures of data blocks. Each data block has its replica stored on another block and incase of failure data can be retrieved from other block. This feature of HDFS makes it easier to use commodity hardware for processing of big data. YARN stands for Yet Another Resource Negotiator. It separates MapReduce from resource manager, workflow manager and fault-tolerance. It allows other frameworks to be built on top of it. The original Hadoop framework was modified to use YARN. The initial version of Hadoop had technical deficiencies [34] that the current system is dealing by introducing a structure called linear data flow on the distributed computing programs in the Hadoop cluster. Hadoop gets an input data from the disk, perform mapping function on the data, reduce results of map function, and then finally, stores reduce results on the disk. Everything was to be read and written to disk. This made the implementation of the iterative algorithms difficult [35]. An Iterative algorithm works on dataset multiple times in a loop and then applying data analysis on side. These training algorithms used in systems having machine learning standards. Current version of Hadoop does not have the capability for processing of iterative machine learning algorithms and if processed it will take a lot of time to finish a job. This provided a need of a technology that would solve these issues. This leads to the development of Apache Spark [36].

C. Apache Spark

Apache Spark was developed to facilitate the iterative and machine learning algorithms. Spark was born along with its important component the Resilient Distributed Datasets (RDDs) [36]. The RDDs perform in-memory computations on big data. It runs on large

clusters and nodes. It runs to provide fault-tolerance. Apache Spark also has many discretized streams. The discretized streams were developed in order to provide high-level programming API. An efficient fault tolerance and consistency achieves by high level programming APIs. Spark is one of most primitive high-level systems that not only supported the distributed batch and stream computation but also the iterative querying.

The key feature of Spark is an RDD [36]. RDDs are basically "immutable objects". These objects are usually stored into different partitions. When one RDD is modified, a new RDD is created. A new RDD generation leaves the previous RDD unconverted. It provides fault tolerance due to intelligence that decides when to regenerate and when to re-compute a dataset. The groundwork of a complete project Spark Core may deliver the scheduling, distribute the task, and fulfill essential input output functionalities. They are revealed by an Application Programming Interface (API). This API is centered on RDD abstraction. RDDs consist of two different kinds of operations: transformations and action transformations. Transformations always return pointers to the latest new RDDs. Another transformation called an Action transformation may return results to a driver program. Different transformations and actions can work together in a Spark job. MLlib is a distributed machine learning library framework that operates on the top of Spark core. The spark job operates nine times efficiently and faster than the disk-based implementation due to distributed memory-centered Spark architecture. Various machine learning algorithms has been proposed and transported to MLlib that enables ML large scale pipelines.

D. Comparison of Tools for Implementation

In order to support our choice of Apache Spark, we have demonstrated the suitability of Apache Spark (MLlib) for machine learning applications by comparing the performance of the two parallelization frameworks, i.e. Spark and Hadoop.
The comparison table of Apache Hadoop and Apache Spark with major differences is given in Table 1.

Table 1: Comparison Table

Attributes	Apache Hadoop	Apache Spark
Unified APIs	No	Yes
I/O Operations	Disk-based	Memory-based
Processing Speed	Slow	Fast
Execution	Slow	Fast
ML Support	Limited (for newer machines)	Full

It is also worth mentioning here that Apache Spark fully supports agglomerative clustering being used in the proposed technique whereas Apache Mahout does not

support agglomerative clustering. It supports two algorithms for clustering i.e. 1) Canopy clustering 2) K-means clustering.

The next section contains the well demonstrated discussion of the proposed technique in detail.

4. Proposed Technique

This section is further divided into two subsections. Taxonomy generation algorithm is discussed in the first subsection and the second subsection discusses taxonomy evolution process for updating taxonomy on arrival of new documents in a dataset. Both the taxonomy generation and evolution algorithms are based on a parallelization framework.

A. Taxonomy Generation

The proposed technique performs taxonomy generation on Apache Spark framework and has been divided into six general steps: loading the data, data pre-processing, data modeling, hierarchy formation, node labeling and conversion into tree graph. The process has been explained in detail in our work [37], highlight of which is below:

1. Loading the Data: Resilient distributed dataset (RDD) is used to effectively load text documents as input.
2. Data Pre-processing: In the pre-processing step, stop-word removal using NLTK and stemming using Porter stemmer [38] have been performed.
3. Data Modeling: A feature vectorization method called term frequency-inverse document frequency (TF-IDF) is used in this step.
   In Apache Spark, TF-IDF is performed in MapReduce paradigm whereas, TF-IDF is not calculated in a simple fashion, rather several number of map and reduce tasks are carried out for the implementation of TF-IDF. Apache Spark implements it using hashing trick or kernel trick. Hashing trick is a quick and compact way of vectorizing features. The hash function used here is MurmurHash3$_2$. Figure 1 demonstrates this process of vectorization and hashing.
4. Hierarchy Formation: To form a hierarchy, hierarchical agglomerative clustering approach is used. For implementation of this phase, Parallel prims algorithm is used. This algorithm is available in open-source Spark's library. First the algorithm divides the dataset into multiple sub-datasets. A serial minimum spanning tree algorithm is applied locally on each of the sub problems on it. Spark's programming model supports iterative algorithms because of RDD. In RDD the computation is carried out only in the RDDs that are required at the moment. In an iterative program, RDDs are consumed in a loop. This phase is

called Map phase. Each MST is

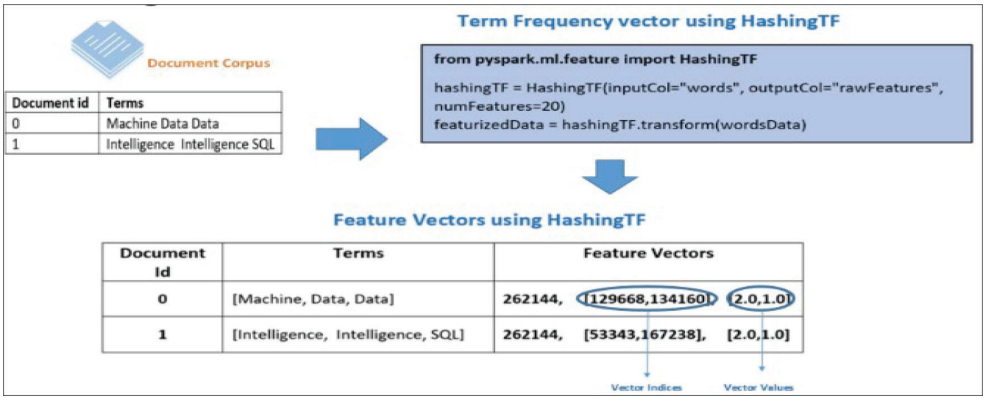


Figure 1: Implementation of TFIDF as HashingTF on Apache Spark

a cluster itself. Now multiple clusters that were created in the map steps are re-arranged in reduce steps based on the distance between two clusters iteratively and arranged in the form one bigger cluster. This is the Reduce phase. The information of distance between the trees clusters are maintained in the similarity matrix S_{gen}.

5. Node Labeling: The hierarchical composition built in the preceding step is unlabeled. This phase acquires labels for these unlabeled clusters. For labeling purpose, titles of documents in a cluster were chosen as labels. The titles were selected as labels because titles are easier to read as compared with the list of top terms in a cluster [18]. The technique basically labels a cluster with the title of the document that is attached to the edge having minimum weight. By the end of this step, the taxonomy T_{gen} has been created.
6. Conversion into a Tree Graph: To use this taxonomy for evolution subsequently, T_{gen} is then transformed into a Newick Tree Graph[3]. Newick is a standard for representing trees in a computer readable form by making use of nested parentheses as shown in Figure 2. The bottom-most node in the tree is an interior node. Matched parentheses represent interior nodes. In between them, there are the images of nodes that are instantly descended from a node which is comma separated. Real numbers are used to incorporate branch lengths. This represents the length of a branch immediately below a node.

```
((((((datamining.9:1.28,3dtech.11:1.28):0.05,(datamining.11:0.92,datamining.10:0.92):0.41):0.42,(computational
geometry.2:0.76,computationalgeometry.1:0.76):0.99):0.33,(((((:1.09,3dtech.13:1.09):0.17,architectureeducation.
7:1.25):0.15,(microarchitecture.2:0.97,microarchitecture.1:0.97):0.44):0.22,(communicationsystems.2:0.70,commu
nicationsystems.1:0.70):0.93):0.45):0.40,((databasesystems.3:0.57,databasesystems.2:0.57):0.31,databasesystem
s.1:0.88):1.59):0.21,(((mobilemultimedia.12:0.74,mobilemultimedia.11:0.74):0.12,mobilemultimedia.10:0.85):0.5
2,(adhoc.10:0.81,adhoc.9:0.81):0.56):1.31);
```

Figure 2: Snippet of a generated Newick Tree

B. Taxonomy Evolution

New documents when added in a dataset for which taxonomy is being maintained, they followed the same process of taxonomy generation as mentioned in the previous subsection. Once taxonomy is generated, a new tree structure T_{evo} is constructed that represents the newly introduced documents. A similarity matrix S_{evo} is also produced. Finally for the taxonomy evolve step, Tree Merge [40] technique is used. The Tree Merge practice takes following as input: the existing T_{gen} , new taxonomy T_evo, the existing similarity matrix S_{gen} and the new similarity matrix S_{evoas} input. After the input has been taken, the next step is the building of a compatibility super tree T_S.

NTMerge algorithm is used for building compatibility super tree [41]. The super tree method constructs trees from smaller trees for overlapping subsets of taxonomies. NJMerge basically runs on an input pair of T_{gen} and $T_{evo,}$ and it also takes similarity matrices S_{gen} and S_{evo} as auxiliary information. As mentioned earlier, in Newick trees numbers are used to represent branch length. NJMerge results the correct neighbors of the tree T_{evo} by comparing and analyzing the branch length of tree structure T_{evo} withT_{gen}. Branch lengths sum achieves for all the branches of both tree structures. The pair having smallest length is called a true neighbor. Once the true neighbors have been identified, the next step is the merging of the two trees. Strict Consensus Merger is used for merging pair of trees in which a merged tree. The proposed technique successfully generates a taxonomy for text documents from a given corpora. The technique also successfully evolves the previously created taxonomy in a very short time. The foundation of the algorithm is a MapReduce in which the capability to minimize the time by parallel data processing and facilitates the fault tolerance feature by using distributed file system. MapReduce environment aids for improving the scalability challenges of an existing taxonomy generation and taxonomy evolution techniques. The algorithm runs on Apache Spark environment. The comparison between our proposed technique and existing taxonomy generation and evaluation with respect to running time and clustering quality has been done. The next section discusses the evaluation of the proposed technique.

5. Evaluation

The technique presented in this research work was evaluated on a textual dataset based upon quality and time parameters. Various experiments were executed using the following experimental configurations:

1. Processor: Intel Core i5
2. RAM: 32 GB
3. Apache Spark version: 2.3
4. Apache Hadoop version: 2.10.0

In the first set of experiments, the generation part of the proposed technique was assessed

by comparing it with a current non-incremental taxonomy generation method TaxGen. In the second set of experiments, the entire algorithm (generation as well as evolution) was evaluated by comparing it with the current incremental taxonomy generation methodology TIE. A textual dataset of ACM scholarly articles, taken from [18] was used for performing these experiments. In the third set of experiments, the scalability of the methodology was evaluated using a separate cluster of Apache Spark, on an individual machine. Due to the limited size of the ACM dataset for testing the scalability PubMed dataset comprising of 17785 documents was used. Last but not least, the focus has been made on the comparison of the two parallelization frameworks namely, Apache Hadoop and Apache Spark. The running time of the parallelization part of the proposed technique was compared on both Apache Hadoop and Apache Spark. The rest of this section will discuss evaluation metrics, experiments, and test results.

A. Evaluation Metrics for Clustering Quality

To evaluate the quality of hierarchical clustering, Silhouette's score and Davies-Bouldin's score are being used as quality metrics [cite our previous work].

1. Silhouette's Score: We can compute the Silhouette's score [42] by using the distance called intra cluster distance and mean closest cluster distance for every data point. The range of Silhouette's score is between [-1, +1]. The values near zero may represent an overlapping cluster. The values that are negative may signify that the data point or document has been assigned to wrong cluster. The higher silhouette value shows that the document matched or assigned to its own cluster and inadequately matched to the other nearby clusters.

2. Davies-Bouldin's Score: Davies-Bouldin's [43] score can be computed as by finding the ratio of sum of within-cluster scatter to the between-cluster separation. In a Davies-Bouldin's score, better clustering quality can be achieved by getting lower score. Zero is the minimum score. If two algorithms are being compared the algorithm with lower score will have well-defined and well-separated clusters.

B. Experiments and Results

Our obtained taxonomy consists of two different kinds of evaluation. The first type is called time-based, whereas the second type is quality-based. We obtained efficiency of time by running time of algorithms for taxonomy generation and evolution. To evaluate hierarchical clustering quality, Silhouette's score and Davies-Bouldin's score are being used. Our evaluation results are given below:

1. Experiments for Generation Process: We compared the generation part of our

technique with an existing taxonomy generation technique, TaxGen [23], by comparing the running time and clustering quality. Initially, 220 documents were involved for taxonomy generation process. Newer documents were then added to evaluate the generation process of taxonomy, which shows better results for the proposed technique [37]. Figure 3(a) & 3(b) shows the hierarchical clustering quality of generated taxonomies, whereas the running time is shown in Figure 4.

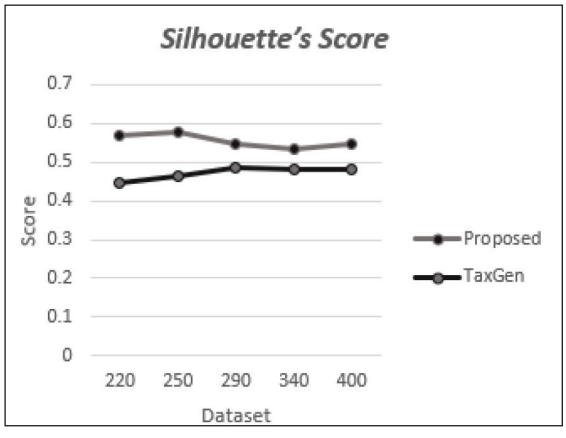


Figure 3(a): Results for Quality-Based Evaluation - Taxonomy Generation Process

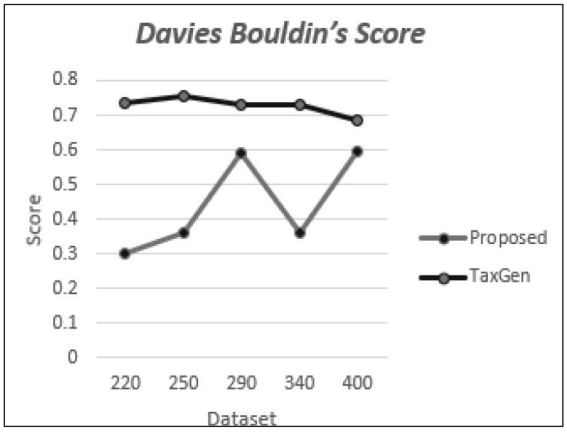


Figure 3(b): Results for Quality-Based Evaluation - Taxonomy Generation Process

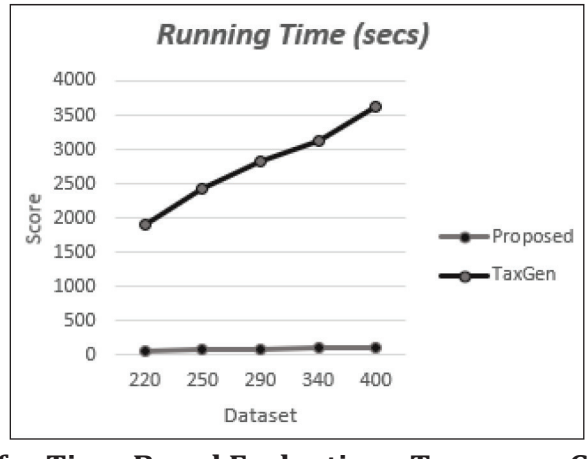


Figure 4: Results for Time-Based Evaluation - Taxonomy Generation Process

2. Experiments for Evolution Process: We compared the evolution part of our proposed method with the evolution of existing method, TIE [18] by performing the result comparisons of hierarchical clustering quality and running rime. To generate the taxonomy, 200 documents were initially used for the taxonomy evolution process using proposed technique and TIE. Then there was a gradually increase in dataset and taxonomy evolution was done using both the methods. The

hierarchical clustering quality scores are shown in Figure 5(a) & 5(b), whereas Figure 6 indicates running time results for both the techniques. The results obtained in both the cases favors the proposed technique [37].

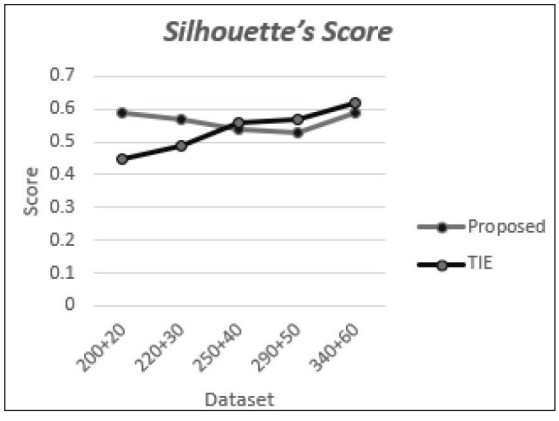


Figure 5(a): Quality Based Evaluation Results-Taxonomy Evolution Process

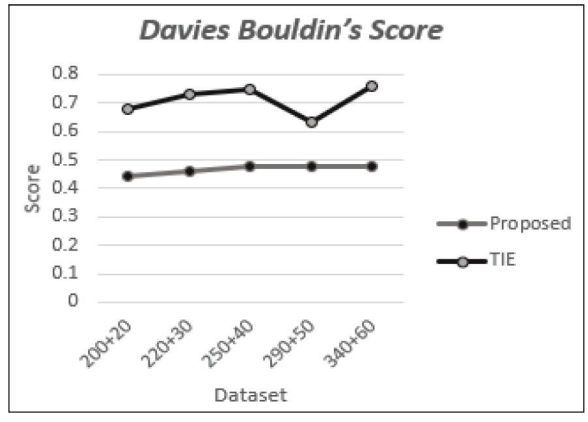


Figure 5(b): Quality Based Evaluation Results-Taxonomy Evolution Process

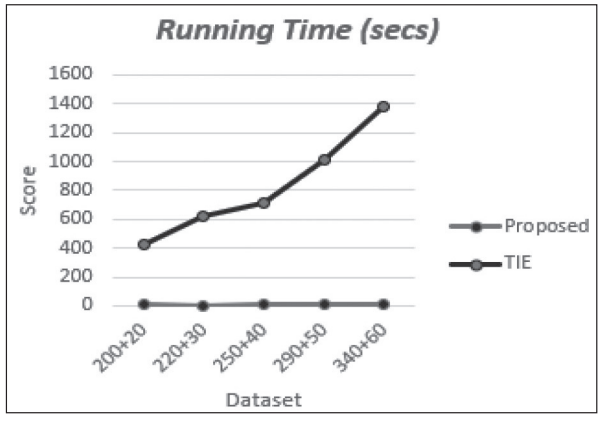


Figure 6: Time-Based Evaluation Results- Taxonomy Evolution Process

3. Experiments for Testing the Scalability: The scalability of the proposed model was tested using a separate cluster of Apache Spark on an individual machine. Due to the limited size of the ACM scholarly articles dataset obtained from [18], for testing the scalability, a dataset namely PubMed was used. This dataset is comprised of 17785 documents. Using 5000 text documents initial taxonomy was generated and after that by adding 5000 documents dataset was gradually increased for the procedure of evolution. Clustering quality for generation and evolution through the proposed methodology was evaluated and the running time was assessed as well. The results of the experiment are shown in Figures 7(a), 7(b) & 7(c), which again show the better cluster.

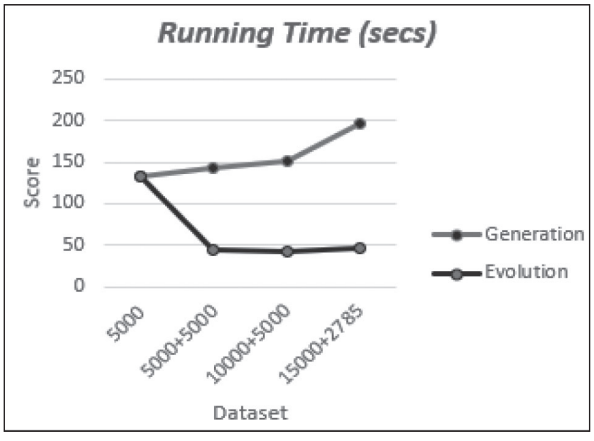


Figure 7(a): Scores showing the Scalability of the Proposed Technique

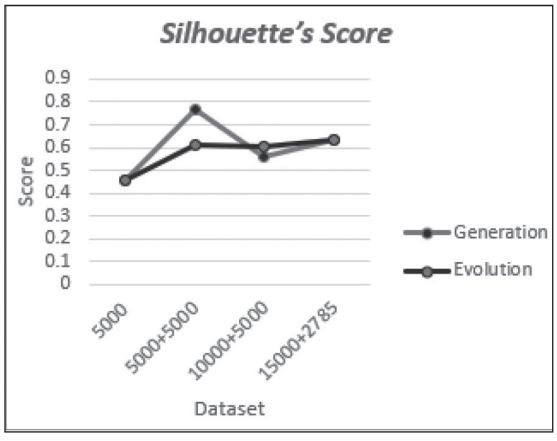


Figure 7(b): Scores showing the Scalability of the Proposed Technique

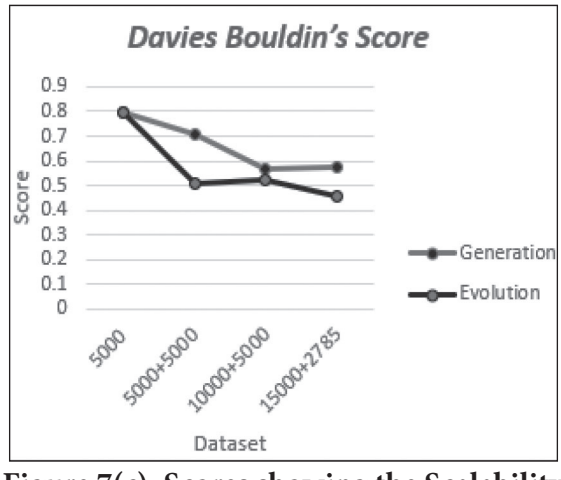


Figure 7(c): Scores showing the Scalability of the Proposed Technique

4. Experiments on Parallelization Framework: In this part of experiment, focus has been made on the comparison of the two parallelization frameworks namely, Apache Hadoop and Apache Spark. Running time of the parallelization part of the proposed technique was compared on both; Apache Hadoop as well as Apache Spark. For the sake of this experiment, we only generated taxonomy, Newick trees were not generated for this experiment. The algorithm was run on 3 cores of Apache Spark and compared it to the processing capability of Apache Hadoop. 3 cores were chosen because that is the minimum number of cores that can be chosen for running of any Spark job. Table 2 shows the running time of the proposed technique on both the environments. It can be observed that the running time of Apache Spark is much smaller as compared with Apache Hadoop.

Table 2: Apache Spark vs. Apache Hadoop Based on the Running Time of the Proposed Technique

Dataset	Running Time (secs)	
	Apache Spark	Apache Hadoop
200	7.97	241.8
300	6.29	369.6
400	8.16	571.8
500	10.11	616.2
600	12.28	673.8
700	14.23	842.4
800	16.68	963
900	18.57	1083.6
1000	19.48	1228.8
1500	42.96	1830.6
2000	45.55	2310

On the basis of running time of the technique on both the environments, their running time ratio, i.e. ⟦RT⟧_r was calculated. The sum of the running time of the algorithm was considered for Apache Spark and Apache Hadoop both and their ratio is calculated as given in (11). According to this time ratio, Spark is 53.05 times faster than that of Hadoop.

$$RT_r = \frac{\sum Running\ Time\ of\ Algorithm\ on\ Hadoop}{\sum Running\ Time\ of\ Algorithm\ on\ Spark} = 53.05 \quad (11)$$

Spark performance, has found out to be optimal over Hadoop as evaluated by processing speed due to following reasons:

1. Spark performs computation using in-memory calculation. It runs a selected part of a MapReduce task and is not bound by input-output concerns every time.

2. Spark's directed acyclic graphs support optimization between steps, whereas any cyclic interconnection between MapReduce steps and levels is not possessed by Hadoop. This means performance tuning cannot be done at that level.

Hence both the theoretical aspects as discussed above and in Section 3.4, as well as the experimental results favor Apache Spark for the case of the proposed methodology.

Apache Spark basically comes with various units and sub-units that aid in the process

of running of Spark jobs. Tuning the resources, parallelism, and using different data representation affect Spark job performance. Schema of data (the way data is arranged) and number of cores for running a job are important factors. The *--executor-cores* specifies the number of cores when submitting a Spark job. A Spark job is submitted by invoking *spark-submit*. In pyspark *--executor-cores* flag are set from the command line. This can also be achieved by using the *spark-defaults.conf* file or a *SparkConf* object and setting the *spark.executor.cores* property. Further experiments were performed to evaluate running time of the evolution process on different number of Apache Spark cores. In the proposed technique running time of taxonomy evolution process against different number of Apache Spark cores is shown in Table 3.

Table 3: Running Time for Taxonomy Evolution on Different No. of Apache Spark Cores

Size of data	Running Time (secs)		
	3 cores	5 cores	8 cores
100+100	7.97	6.79	6.29
200+100	6.29	7.12	6.25
300+100	8.15	8.29	8.08
400+100	10.11	9.72	9.83
500+100	12.28	12.09	11.99
600+100	14.23	14.06	13.68
700+100	16.68	15.98	15.44
800+100	18.57	18.07	17.69
900+100	19.48	19.87	18.46
1100+100	29.73	24.56	23.34
1200+100	34.97	26.78	27.12
1300+100	39.14	30.89	29.87
1400+100	42.96	32.02	30.38
1500+100	44.67	35.44	33.43
1600+100	45.23	37.89	36.32
1700+100	45.76	39.34	38.56
1800+100	46.26	41.56	40.53
1900+100	47.55	43.05	42.67

In Table 3, it can be seen that when taxonomy evolution process was run on different number of Spark cores, the running time of algorithm for 8 cores gives the minimum time. When we specify the number of cores to be 8 that means each executor runs 8 tasks at a given time. It should be noted here that initially when dataset is small the time taken by all three cores to evolve taxonomy is comparable but as the dataset increases the significant difference can be seen in the running time. The impact of using different cores

can be better visualized when dataset is even larger.

C. Discussion

In this section, the proposed methodology is assessed and evaluated by comparing it with an existing algorithm of taxonomy generation i.e., TaxGen. Evolution part of the technique has also been compared with another algorithm of incremental taxonomy generation i.e., TIE. The technique has been assessed and evaluated on the basis of the running time and quality of clustering. Clustering quality was evaluated using two different techniques i.e., 1) Silhouette's score 2) Davies Bouldin's score. It was observed that the proposed technique shows better clustering quality when compared with TaxGen and TIE.
It is evident that the running time of the proposed methodology is significantly smaller as compared to its counter parts. Due to the usage of map reduce framework, the asymptotic complexity of the proposed technique is also reduced from O(n^3) for hierarchical clustering to $O\left(\frac{n^3}{k}\right)$, where k is the number of nodes in which task is divided in map reduce setup.

The technique was also evaluated by running it on Apache Spark and Apache Hadoop both and their running time was compared. It was found that Apache Spark generated taxonomy in much smaller time as compared with Apache Hadoop. So, it can be said that for a clustering problem like taxonomy generation Apache Spark is a better choice. It was also evaluated by experiment that by using how many cores of Apache Spark the proposed technique can evolve taxonomy faster. It was found that when data size is small number of cores do not matter. As the size of data grows using 8 cores can bring significant time improvement. The execution time taken for an analysis to perform is critical in big data applications. The execution time is measured to evaluate the performance. Smaller execution times indicate that the program runs fast and gives good performance. It should also be noted that the proper resource utilization is also crucial in case of large datasets. A good application should give high performance with minimal resource utilization. Since the technique utilizes MapReduce algorithm as its core technique while running on Apache Spark, this makes the technique scalable. The next chapter concludes this research work.

5. Conclusion and Future Work

This research work has reviewed the existing techniques of taxonomy generation and evolution from the perspective of today's data which is particularly fast-evolving and voluminous. It was identified that in the modern era of big data, it is required that there must be some efficient and scalable taxonomy generation and evolution techniques to handle this type of data. Although some work has been done in the field of hierarchical

clustering for substantial datasets, little focus has been made on generating and evolving taxonomy. As per the available information, none of the existing techniques have focused on the idea of parallelization to develop an effective and scalable algorithm for the process of taxonomy generation and evolution. In this research work, a novel and unique technique has been developed for the process of taxonomy generation and evolution which is based upon the MapReduce paradigm using the framework of Apache Spark has the ability to minimize the time by parallel data processing. It also provides the feature of fault tolerance by using the distributed file system (DFS).

The proposed technique is evaluated on the basis of the clustering quality and the time takes to generate and evolve taxonomy in contrast to the present taxonomy generation (TaxGen) and evolution (TIE) methodologies. It is quite clear from the results obtained so far, that the proposed methodology consumes less time for taxonomy generation and evolution. Evaluation based on quality metrics has been done by applying Silhouette's and Davies-Bouldin's scores. When compared with the existing techniques, both the indices verify improved hierarchical clustering for the proposed methodology. Some experiments are also performed for comparing the two parallelization frameworks, namely Apache Hadoop and Apache Spark using 3 cores setting. The running time of the parallelization part of the proposed technique has been compared on both, Apache Hadoop and Apache Spark. Spark's performance is observed to be optimum over Hadoop as measured by processing speed. Furthermore, some specific experiments have been also performed to test the scalability of the suggested technique by using a specifically large dataset. The time and quality-based evaluation have made it clear that the use of the MapReduce environment has improved the scalability issues of current techniques of taxonomy generation and evolution.

There were certain challenges faced during the implementation of algorithms. Initially, we had decided to use Hadoop to perform taxonomy generation and evolution. We faced no issue in performing taxonomy generation on Hadoop but for evolution, we ran into a problem as we are using Newick tree graph technique for the evolution of taxonomy, and Hadoop's scope is limited when it comes to Newick graphs. Hence, we selected Apache spark as it supports map-reduce as well as Newick graph techniques.

This work too is bound to observe some limits. The proposed model is capable of evolving a taxonomy that has been converted into a Tree graph only. Prospectively, we are in the view of working on proposing a more generalized algorithm that can upgrade/evolve any taxonomy being given as an input. The labeling technique and the hierarchical clustering quality of the taxonomy can be further improved. In the future, we also strategize to evaluate our proposed technique using cloud computing to acquire better results in terms of scalability and performance in the spirit of big data.

References

[1] M. Zwolenski, and L. Weatherill, "The Digital Universe: Rich Data and the Increasing Value of the Internet of Things,"Journal of Telecommunications and the Digital Economy, vol. 2, no. 3, pp. 1—47, 2014.

[2] Rajesh Math. "Big Data Analytics: Recent and Emerging Application in Services Industry. Part of the Advances in Intelligent Systems and Computing book series (AISC, volume 654)" SpringerDoi: 978-981- 10-6620-7_21

[3] Coronel, C., Morris, S., Rob, P. (2013). Database Systems: Design, Implementation, and Management, (10th. Ed.). Boston: Cengage Learning.

[4] ImenChebbi, WadiiBoulila, ImedRiadh Farah."Big Data: Concepts, Challenges and Applications"Springer Doi: 978-3-319-24306-1_62

[5] GeorgiosSkourletopoulos, Constandinos X. Mavromoustakis, George Mastorakis, Jordi MongayBatalla, CiprianDobre, Spyros Panagiotakis and EvangelosPallis: Big Data and Cloud Computing."A Survey of the State-of-the-Art and Research Challenges" SpringerDoi: 9783319451435c2

[6] M. S. Paukkeri, A. P. García-Plaza, V. Fresno, R. M. Unanue, and T. Honkela, "Learning a taxonomy from a set of text documents," Applied Soft Computing, vol. 12, no. 3, pp. 1138–1148, 2012.

[7] R. Sujatha, R. Bandaru, and R. Rao, "Taxonomy Construction Techniques–Issues and Challenges," Indian Journal of Computer Science and Engineering, vol. 2, no. 5, pp. 661-671, 2011.

[8] H. Hedden, The Accidental Taxonomist, Information Today, Inc., 2016.

[9] D. Sánchez, and A. Moreno, "Automatic Generation of Taxonomies from the WWW," In International Conference on Practical Aspects of Knowledge Management, pp. 208-219, Vienna, Austria, December 2004.

[10] H. Delgado, Taxonomy Organization of information of Web Content, 2019. https://disenowebakus.net/en/taxonomyinformation-web-content

[11] V. Kashyap, C. Ramakrishnan, C. Thomas, and A. Sheth, "TaxaMiner: an experimentation framework for automated taxonomy bootstrapping," International Journal of Web and Grid Services, vol. 1, no. 2, pp. 240–266, 2005.

[12] D. Boley, "Principal Direction Divisive Partitioning," Data Mining and Knowledge Discovery, vol. 2, no. 4, pp. 325–344, 1998.

[13] L. E. Anke, H. Saggion, and F. Ronzano, " TALN-UPF: Taxonomy Learning Exploiting CRF-based Hypernym Extraction on Encyclopedic Definitions," In Proceedings of the 9th International Workshop on Semantic Evaluation, pp. 949–954, Denver, Colorado, June 2015.

[14] P. Velardi, S. Faralli, and R. Navigli, "Ontolearn Reloaded: A Graphbased Algorithm for Taxonomy Induction," Computational Linguistics, vol. 39, no. 3, pp. 665–707, 2013.

[15] Yao, B. Cui, G. Cong, and Y. Huang, "Evolutionary Taxonomy Construction from Dynamic Tag Space," World Wide Web, vol. 15, no. 5, pp. 581–602, 2012.

[16] L. Tang, H. Liu, J. Zhang, N. Agarwal, and J. J. Salerno, "Topic Taxonomy Adaptation for Group Profiling," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 4, pp. 1–28, 2008.

[17] R. M. Marcacini, and S. O. Rezende, "Incremental Construction of Topic Hierarchies using Hierarchical Term Clustering.," In Software Engineering and Knowledge Engineering (SEKE), pp. 553–558. Redwood City, California, USA, July 2010.

[18] R. Irfan, S. Khan, K. Rajpoot, and A. M. Qamar, "TIE Algorithm: a Layer over Clustering-based Taxonomy Generation for Handling Evolving Data," Frontiers of Information Technology Electronic Engineering (FITEE), vol. 19, no. 6, pp. 763-782, 2018.

[19] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data Mining with Big Data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2013.

[20] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, "Big data: The Management Revolution," Harvard Business Review, vol. 90, no. 10, pp. 60–68, 2012.

[21] M. J. Embrechts, C. J. Gatti, J. Linton, and B. Roysam, "Hierarchical Clustering for Large Data Sets," In Advances in Intelligent Signal Processing and Data Mining, vol 410, pp. 197—233, Springer, 2013.

[22] R. Babbar, I. Partalas, E. Gaussier, M. R. Amini, and C. Amblard, "Learning Taxonomy Adaptation in Large-scale Classification," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 3350–3386, 2016.

[23] A. Muller, J. Dorre, P. Gerstl, and R. Seiffert, "The TaxGen Framework: Automating the Generation of a Taxonomy for a Large Document Collection," In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences HICSS-32, Hawaii, USA, 1999.

[24] E. A. Dietz, D. Vandic, and F. Frasincar, "Taxolearn: A Semantic Approach to Domain Taxonomy Learning," In 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 58-65, Macau, China, 2012.

[25] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," In TextMining Workshop at KDD2000, May 2000.

[26] H. Schütze, C. D. Manning, and P. Raghavan, Introduction to Information Retrieval, Cambridge University Press, 2008.

[27] A. K. Jain, "Data Clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.

[28] R. Irfan, S. Khan, M.A. Abbas, and A. A. Shah, "Determining Influential Factors and Challenges in Automatic Taxonomy Generation: A Systematic Literature Review of Techniques 1999-2016," Information Research: An International Electronic Journal, vol. 24, no. 2, 2019.

[29] V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, "Unstructured Data Analysis on Big Data using MapReduce," Procedia Computer Science, pp. 456–465, 2015

[30] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, "Big Data Clustering: A Review," In International Conference on Computational Science and its Applications, Springer, 2014, pp. 707– 720.

[31] D. Moulavi, P. A. Jaskowiak, R. J. Campello, A. Zimek, and J. Sander, "Density-based Clustering Validation," In Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 2014.

[32] B. Zerhari, A. A. Lahcen, and S. Mouline, "Big Data Clustering: Algorithms and Challenges," In Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15), Tetuan, Morocco, May, 2015.

[33] J. Dean, and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

[34] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," In 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, May, 2010.

[35] J. Lin, "Mapreduce is good enough? if all you have is a hammer, throw away everything that's not a nail!," Big Data, vol. 1, no. 1, pp. 28-37, 2013.

[36] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave et al., "Apache Spark: A Unified Engine for Big Data Processing," Communications of the ACM, pp. 56–65, 2016.

[37] K. Aalijah and R. Irfan, "Scalable Taxonomy Generation and Evolution on Apache Spark," 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2020, pp. 634-639, doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00110.

[38] A. G. Jivani, "A comparative study of stemming algorithms," International Journal of Computer Technology and Applications, vol. 2, no. 6, pp. 1930-1938, 2011.

[39] F. Murtagh, and P. Legendre, " Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?," Journal of Classification, vol. 31, no. 3, pp. 274-295, 2014.

[40] E. K. Molloy, and T. Warnow, "TreeMerge: A New Method for Improving the Scalability of Species Tree Estimation Methods," Bioinformatics, vol. 35, no. 14, pp. 417-426, 2019.

[41] E. K. . W. T. Molloy, "NJMerge: A Generic Technique for Scaling Phylogeny Estimation Methods and its Application to Species Trees," In RECOMB International conference on Comparative Genomics, Magog-Orford, QC, Canada, 2018.

[42] S. Petrovic, "A Comparison between the Silhouette Index and the Davies-Bouldin Index in Labelling IDs Clusters," In Proceedings of the 11th Nordic Workshop of Secure IT Systems, Linköping, Sweden, October 2006.

[43] J. Xiao, J. Lu, and X. Li, "Davies Bouldin Index based Hierarchical Initialization K-means," Intelligent Data Analysis, vol. 21, no. 6, pp. 1327-1338, 2017.

https://doi.org/10.51153/kjcis.v5i1.68

An Ontology Based Approach to Search Woman Clothing from Pakistan's Top Clothing Brands

Shabina Mushtaque[1] Adnan Ahmed Siddiqui[2] Muhammad Wasim[3]

Abstract

The trend of online shopping in Pakistan has been getting more popular since the last decade. Online clothing shopping is also excessively growing and a wider range of clothing items, including stitched and unstitched, are available from different retailers. It has been observed that most women prefer branded clothing whenever they proceed for the online shopping. Many female consumers are faced with the problem of searching for their desired apparel based on the dress attributes from the collection of their favorite brands. The research is aimed to provide a platform for the users to find their desired dresses from the latest available collection of top brands in Pakistan based on different attributes of a dress. A wider range of clothing including stitched and unstitched items will be made available in the system for more precise and accurate search. The system will extract major attributes of dresses, i.e., color, style, and pattern, from the images uploaded by the brands registered in the system. The system will allow consumers to find clothing from their favorite vendors/brands based on major dress attributes using domain knowledge base defined as ontology (OWL/RDF).

Keywords: Women-clothing, brands, top-brands, clothing, ontology

1. Introduction

The trend of online shopping has been drastically increasing in Pakistan since the last few years. Pakistan has one of the largest populations of internet users. The trend of e-shopping for clothing is also growing in Pakistan. For the retailers and marketers, it is very essential to identify the requirements of their customers to fulfill their shopping needs [1]. E-retailers of clothing are more focused on improving the online shopping experiences of their customers [2]. They keep on exploring the factors that directly impact on customers' satisfaction and finding ways to provide their customers a perfect place to find what they want to wear. Men and women may differ in evaluating different attributes of a product before purchasing it [3]. In online shopping, women can be interested in the more detailed information of a particular product, also numerous search options for example color combination, design, and other features like embroidered dress, sleeveless,

[1] *Usman Institute of Technology, Karachi | smushtaq@uit.edu*
[2] *Hamdard University, Karachi | adnan.siddiqui@hamdard.edu.pk*
[3] *Usman Institute of Technology, Karachi | mwaseem@uit.edu*

with-collar, without-collar, long length, short length, etc., can enhance their intentions of purchase.

It has been observed that for the fabric quality, prints, and style, women's consumers mostly prefer their favorite brands. They prefer to purchase their dresses from specific brands to avoid quality related issues as the online shopping cannot provide touch and feel options. Women in Pakistan are more emotional towards a brands status, quality, and design uniqueness [4]. In Pakistan, different clothing brands attracts female by building trust on quality and by providing a wider range of designs, prints, and styling options.

Mostly young females from universities and colleges are more attracted to brands as they are more aware of the designs, quality of the fabric, latest fashion trends, discounts, and sales because of the internet and social media accessibility [5]. Gul Ahmed, Nishat Linen, Junaid Jamshaid, Alkaram Studio, Sana Safinaz, Maria.B, and Khaadi are among the most popular clothing brands in Pakistan and people are more interested in the products' collection of these brands because of their uniquely appealing designs and quality [6].

The aim of the system is to gather all these well-reputed and famous brands on a single platform to display their products with all their features and attributes of the products. So that the system will provide an advanced search option not only based on price range or category but also based on different dress attributes. The system will be using an ontology-based approach to populate its knowledge base with the dress attributes. An already existing approach will be used to extract the attributes of any particular dress from the uploaded image by the registered brand in system.

2. Literature Review

There are some ontology-based frameworks that have been built to provide recommendations for dresses on the basis of their personality and color of garment. The reasoning model in a garment domain is based on the construction of observation model (OM) and recommendation model (RM) that have been designed previously [7]. Knowledge-based systems have been developed to provide suggestions on the basis of colors and mix-match module for dresses and worked as dress advisor [8]. Garment recommendation for occasions after learning personality attributes like body color and body dimensions from the photograph of the user by using domain knowledge described via MOWL (Multimedia Web Ontology Language) [9]. Another approach is employed for improving quality of clothing recommendations by establishing a knowledge graph of user, clothing and context [10].

The semantic description of fashion ontology helps in populating clothing attributes with images. The attributes includes; clothing pattern, major color, sleeve length, Collar

presence, dress category like (tank top, long shirt, short kurti, etc.) [11].

Estimation of human posture will help in extracting clothes from an image by just subtracting the detected body parts from the captured image. Shape matching templates defined in body-model helps in identifying various body parts of human with different postures. The basic body parts template contains descriptions of various posture of legs, hands, head, upper and lower body [12, 13]. Recent research has been conducted on cloth parsing from photographs using novel dataset to perform a precise cloth estimation of a person's outfit and then labeling the outfit using labeling tools for various possible garment types [14].

The domain ontology needs to be populated automatically after the dress attributes extraction. Many mechanisms have been used for the automatic ontology population. Classification of ontology classes and finding instances from the text is considered as an approach for automatic or semi-automatic ontology populations [15].

The above cited papers dealt with a few concerns related to searching in clothing domain, but a more precise and accurate search model for women's clothing on the basis of dress attributes, from the collection of top Pakistani brands with complete knowledge-based information of on women-clothing using ontology (semantic approach), is still needed.

3. Methodology and Approach

A. Development of Domain-Ontology

The ontology of the domain will consist of a few major classes and their sub-classes. Protégé is a tool used for domain modelling or defining ontologies [16]. It is an easy-to-understand tool to describe ontology classes, instances, attributes, and relations among them. The system needs an ontology or a domain related vocabulary to be defined. The ontology for the system is designed on Protégé -5.2.0 as shown in Figure 1.

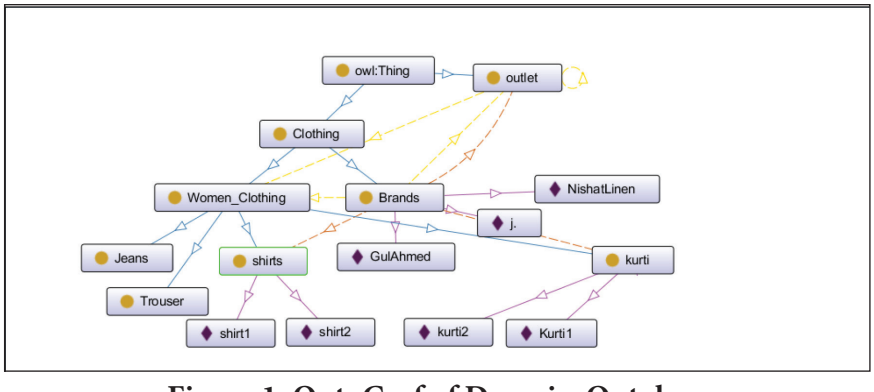


Figure 1: OntoGraf of Domain-Ontology

B. Extraction of Dress Attributes from Image

The system will get the already uploaded images (uploaded by registered brands) from the system's database. A human Pose Estimation, body parts and posture detection technique is used to separate human body parts from the image [17]. This previously defined mechanism will help to get the major part of the dress by subtracting the detected human parts from the dress. After removing unnecessary details the image, the system will work on this separated part of the image that contains only that significant part of dress shown in Figure 2 [18].

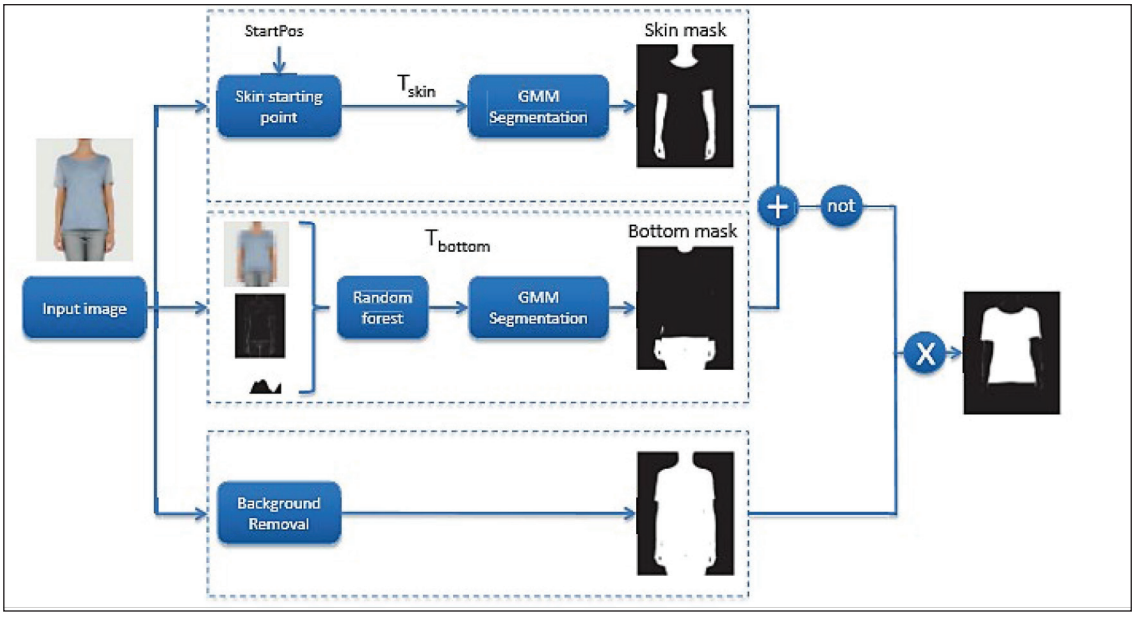


Figure 2: Segmentation of Garment

The already defined automated system that is capable of finding different attributes of the dress from an image will make it easier for the system to separate out the major properties of any dress, shown in Figure 3.

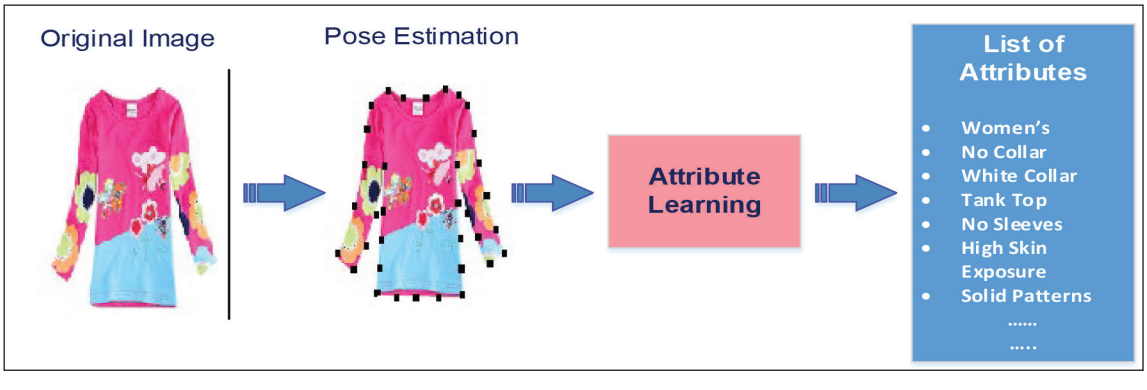


Figure 3: Extracting Dress Attributes

The major attributes of dress defined in the proposed ontology model are:

Color:	*any RGB (0-255)*
Major Color:	*any RGB (0-255)*
Type of Dress:	*Long shirt, Short shirt, Tank top*
Pattern:	*Stripped, Floral*
Is-Collar:	*Yes, No*
Is-Sleeve:	*Yes, No*

After learning the image for these defined attributes, the system will then use an automatic ontology update mechanism to update the ontology after adding individuals to it. Automatic ontology population after learning data sources like images and text, considers as the necessary approach to make such semantic based systems more effective and independent to update their knowledge [15].

For example, the image we are using in Figure 3 lies in the category (class) "kurti", so the system will consider it as an individual for the already defined class "kurti", with any name like "Kurti3" in this example. So, the system will automatically add an individual to the already defined ontology.

The extracted attributes from kurti3:

:Color "white" ;
:is-collar "No" ;
:is-sleeve "yes" ;
:pattern "solid pattern" ;
:type "Tank Top"

C. Automatic population of Ontology

RDF (Resource Description Framework) stores information in the form of Triples:

Subject->Predicate->Object
or
Object->Property->Value

So, the system will automatically insert a triple while updating the ontology. There are many RDF management systems that have been developed by researchers to handle millions of triples defining ontology related to a particular domain [19]. Ontology is about enriching vocabulary related to the domain or about adding triples to the RDF

graph. So, it requires some mechanisms to populate ontology automatically. The different systems have been designed for automatic population of ontology by extracting concepts or classes, relations from text as in [20]. In this system, it is needed to enrich the ontology by adding classes, concepts, individuals and properties as per requirements. SPARQL (SPARQL Protocol and RDF Query Language) pattern matching techniques are also helping in identifying different ontological components. SPARQL queries are also used in mapping patterns to accurately populate ontology [21]. So, we need to execute a SPARQL query for the insertion of an individual with the following five data properties:

i. Color
ii. Type
iii. Is-sleeves
iv. Is-Colar
v. Patterns

After query execution the graph has now populated with an individual named with kurti3.This individual contains five object properties. The graph in figure 4 is visualized on OWLGrED, an online ontology visualization tool that allows a graphical representation of OWL classes in a form of UML classes in order to get more clear view of any domain-specific ontology (OWL) [22].

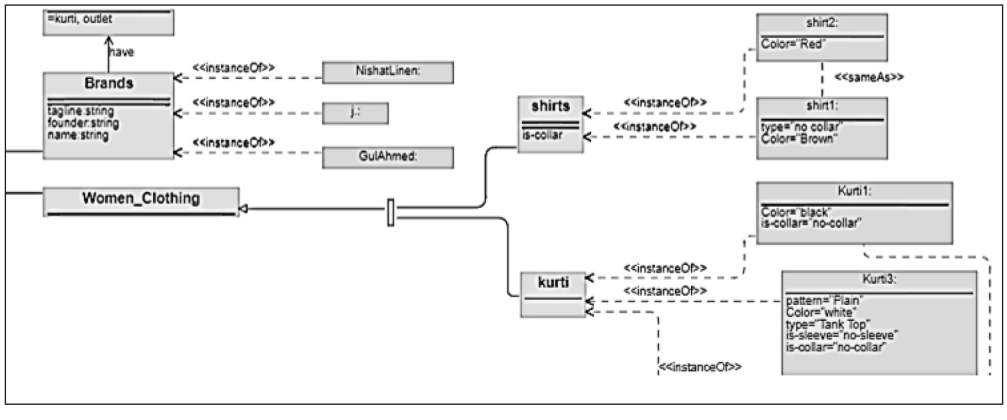


Figure 4: Graphical Representation of OWL and UML

Data properties of the recently added individual 'kurti3' can be shown in figure 5 using Protégé -5.2.0.

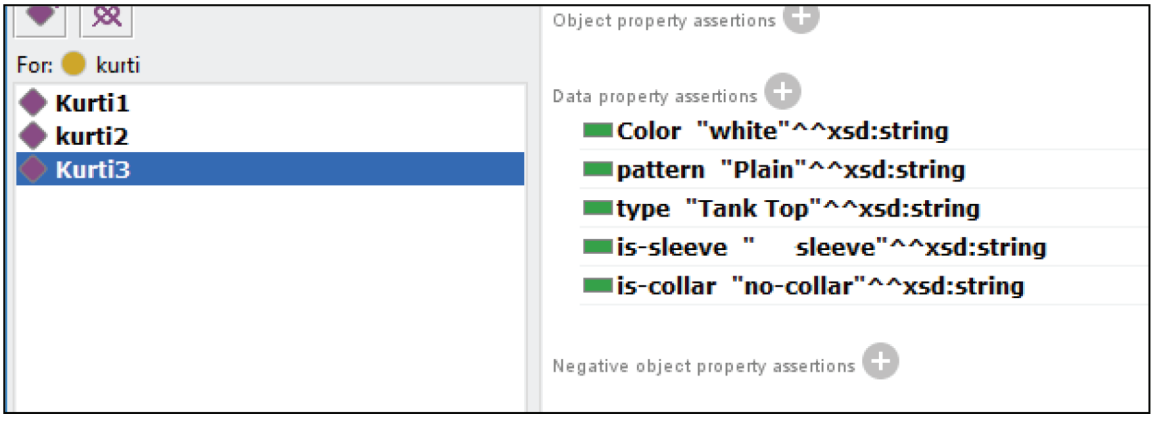


Figure 5: Data Properties of the individual 'kurti3'

D. User search module using SPARQL

SPARQL is used as the standard language for querying RDF or Ontology/Knowledge based systems [23]. A user can search for a dress as per their desire like color, type, or any other specification of the dress like no-collar, sleeveless, etc.
The system will generate query of SPARQL according to the user input or query request.

For example
User enters color= white
PREFIX
dc:<http://www.semanticweb.org/arbab/ontologies/2017/10/Women_Clothing.owl>
SELECT ?subject ?predicate
WHERE {
?subject ?predicate 'white'
}

The search results of the above *query is shown in Figure 6. If the user is searching for a dress with color white, for example, the query will return the result, that is based on triple pattern defined in the WHERE clause.

*Query is executed on Apache Jena Fuseki server to test the results.

	subject	predicate
1	dc:#kurti2	dc:#Color
2	dc:#Kurti3	dc:#Color

Figure 6: Search results of the query as subject and predicate

4. System Frontend

The front end of the system is based on the.NET Technology, C#. It is designed using the Windows Form Application provided by Microsoft Visual Studio using version 2013. DotnetRDF is an open-source .NET library that supports a way to interact with RDF via queries. Different Semantic based systems use this technology for constructing GUI under windows as in [24-25]. It provides a framework to construct a GUI for RDF based system. It contains a lot of supportive classes like IGraph, Graph, Triple, Node, TripleStore, etc. to interact with the triples of any RDF. There are multiple supportive classes with different methods that helps in querying RDF, it also provides support for connection with third-party stores like Jena-Fuseki, Virtuoso, AllegroGraph, store4, etc.

DotnetRDF provides ExecuteQuery() function that helps to execute any SPARQL query. The Following example shows the execution of the query that will return a result in a result set that will contain all the triples available in the provided graph.

Example:
Query string in C# syntax:
String query= "SELECT ?subject ?predicate ?object WHERE {?subject ?predicate 'white'}";

The query is executed with the help of defined DotnetRDF classes.
Results of the query is displayed using Windows form application as shown in Figure 7 user is querying for 'white' kurti and results are returned to the user.

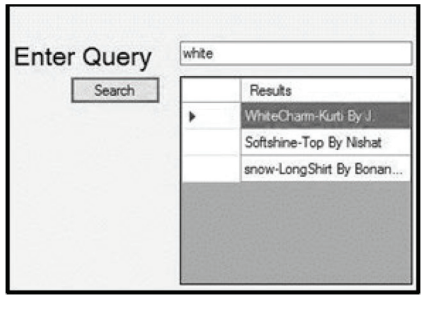


Figure 7: Search Results of the requested query

5. Conclusion

In this article, an ontology-based system is introduced for searching women's clothing from the top brands in Pakistan on the basis of dress attributes like dress color, type, and other specifications of dress like sleeveless, no collar, etc. The Protégé tool is used for defining domain ontologies. The proposed system is able to automatically populate the domain ontology from the knowledge sources. The system uses existing tools for

processing images to extract dress attributes from the already available images in the system in order to populate the domain ontology. A user can make a search query that can include any desired dress attribute or attributes, the proposed system is capable of handling user queries by using SPARQL to find the desired results.

6. Future Work

There is a broader scope of extension in the proposed system by adding more versatility in cloth searching, more specifically, extracting dress attributes of man's and children's clothing and adding them to the domain knowledge base. Further, conceptual advancements in the discussed system are also achievable on the basis of the feedback from users and testing results in order to make searching more accurate.

References

[1] Ahmed, Zahid, et al. "A study on the factors affecting consumer buying behavior towards online shopping in Pakistan." Journal of Asian Business Strategy 7.2 (2017): 44.

[2] Pandey, Shweta, and Deepak Chawla. "Online customer experience (OCE) in clothing e-retail: exploring OCE dimensions and their impact on satisfaction and loyalty–does gender matter?." International Journal of Retail & Distribution Management (2018).

[3] González, Eva M., Jan-Hinrich Meyer, and M. Paz Toldos. "What women want? How contextual product displays influence women's online shopping behavior." Journal of Business Research 123 (2021): 625-641.

[4] M. Khakhan and k. A. Siddiqui, "women's perceptions towards branded clothing in pakistan.

[5] I. Inayat and a. Kamal, "journal of isoss 2017 vol. 3 (2), 253-260 factors affecting young female consumer's behavior towards branded apparels in lahore," journal of isoss, vol. 3, pp. 253-260, 2017.

[6] S. Munir, a. A. Humayon, m. Ahmed, s. Haider, and n. Jehan, "brand image and customers' willingness to pay a price premium for female's stitched clothing," pakistan journal of commerce and social sciences, vol. 11, pp. 1027-1049, 2017.

[7] S. Ajmani, h. Ghosh, a. Mallik, and s. Chaudhury, "an ontology based personalized garment recommendation system," in proceedings of the 2013 ieee/wic/acm international joint conferences on web intelligence (wi) and intelligent agent technologies (iat)-volume 03, 2013, pp. 17-20.

[8] C.I. Cheng, d. S.-m. Liu, m.-l. Liu, and i.-e. Wan, "clothing matchmaker: automatically finding apposite garment pairs from personal wardrobe," the international journal of organizational innovation, p. 78.

[9] D. Goel, s. Chaudhury, and h. Ghosh, "multimedia ontology based complementary garment recommendation."

[10] Wen, Yufan, Xiaoqiang Liu, and Bo Xu. "Personalized Clothing Recommendation Based on Knowledge Graph." 2018 International Conference on Audio, Language and Image Processing (ICALIP). IEEE, 2018.

[11] H. Chen, a. Gallagher, and b. Girod, "describing clothing by semantic attributes."

[12] K. Mikolajczyk, c. Schmid, and a. Zisserman, "human detection based on a probabilistic assembly of robust part detectors," computer vision-eccv 2004, pp. 69-82, 2004.

[13] B. Wu and r. Nevatia, "detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," international journal of computer vision, vol. 75, pp. 247-266, 2007.

[14] K. Yamaguchi, m. H. Kiapour, l. E. Ortiz, and t. L. Berg, "parsing clothing in fashion photographs," in 2012 ieee conference on computer vision and pattern recognition, cvpr 2012, 2012.

[15] C. Faria, i. Serra, and r. Girardi, "a domain-independent process for automatic ontology population from text," science of computer programming, vol. 95, pp. 26-43, 2014.

[16] N. F. Noy, m. Sintek, s. Decker, m. Crubézy, r. W. Fergerson, and m. A. Musen, "creating semantic web contents with protege-2000," ieee intelligent systems, vol. 16, pp. 60-71, 2001.

[17] J. Shotton, a. Fitzgibbon, m. Cook, t. Sharp, m. Finocchio, r. Moore, et al., "real-time human pose recognition in parts from single depth images."

[18] M. Manfredi, c. Grana, s. Calderara, and r. Cucchiara, "a complete system for garment segmentation and color classification," machine vision and applications, vol. 25, pp. 955-969, 2014.

[19] J. Huang, d. J. Abadi, and k. Ren, "scalable sparql querying of large rdf graphs," proceedings of the vldb endowment, vol. 4, pp. 1123-1134, 2011.

[20] W. Wong, w. Liu, and m. Bennamoun, "ontology learning from text: a look back and into the future," acm computing surveys (csur), vol. 44, p. 20, 2012.

[21] I. Haidar-ahmad, a. Zouaq, and m. Gagnon, "automatic extraction of axioms from wikipedia using sparql," in international semantic web conference, 2016, pp. 60-64.

[22] K. Cerans, r. Liepins, a. Sprogis, j. Ovcinnikova, and g. Barzdins, "domain-specific owl ontology visualization with owlgred," in extended semantic web conference, 2012, pp. 419-424.

[23] S. Bischof, m. Krötzsch, a. Polleres, and s. Rudolph, "schema-agnostic query rewriting in sparql 1.1," in international semantic web conference, 2014, pp. 584-600.

[24] G. Barbur, b. Blaga, and a. Groza, "ontorich-a support tool for semi-automatic ontology enrichment and evaluation," in intelligent computer communication and processing (iccp), 2011 ieee international conference on, 2011, pp. 129-132.

[25] S. E. Hamada, i. A. Alshalabi, k. Elleithy, i. Badara, and s. Moslehpour, "updating student profiles in adaptive mobile learning using asp. Net mvc, dotnetrdf, turtle, and the semantic web," international journal of interactive mobile technologies (ijim), vol. 11, pp. 16-38, 2017.

Risk Assessment Approach for Software Development using Cause and Effect Analysis

Abdur Rehman Riaz[1] Syed Mushhad M. Gilani[2]

Abstract

In the industry of software development, the risk is most effective competitor that tries to flop the project at any stage of development. Risk is a critical factor that is obscure and has the potential to wreak massive loss of the project in terms of money, time and resources. It is also harmful to the credibility of the organization. Most of the organizations don't focus on this factor, and as a result, might witness the project failing. Lack of risk assessment is very common in most organizations. This paper introduces risk assessment factors and analyzing the various situations to tackle these. We investigate the most effective key risk variables cost, strategy, technique, operation, and some unknown/unpredictable factors. Based on these variables, survey and interviews are conducted and examined. We applied empirical studies on these variables and map them on the cause-and-effect analysis technique. The proposed technique elaborates the factors behind these risk variables. After that, results and analysis of these variables have been incorporated to scale down the impact of risk.

Keyword: Software Project Risks, Risk Management, Risk Assessment, Risk Analysis, Cause and effect analysis

1. Introduction

Threats are identified during the risk analysis phase that may harm project development. Risk ensures from two factors: prediction of project failure, and its impact on the project [1]. Risk can be of any type like cost, time and scope. Risk analysis is a critical task that needs detailed information about the project so that these issues can be resolved and save assets of the project. During the planning of software project development, some risk analysis methods are used; such as qualitative risk analysis which is applied to the project for tracking down the issues that affect the quality of the project. Software project management is a tricky task to handle project [2] because the whole project depends on software project management. It helps to get profit or loss and demands a lower cost, minimum time, higher productivity, good quality and fine customer satisfaction to deliver the product right time in the market. Risk management is one of the vital part of software project management due to that risk is involved in each software engineering process [3].

[1] *PMAS Arid Agriculture University, Rawalpindi | abdurrehman.ar475@gmail.com*
[2] *PMAS Arid Agriculture University, Rawalpindi | mushhad@uaar.edu.pk*

The basic rule of risk management is defined in ISO 31000 as a state that risk management can create and protect value [2].

In every phase of the project, there are chances of risk, and it is the biggest challenge in software development to remove the risk from these phases. Ignoring the risk factors in the project can lead the project towards failure. Project failure is a loss of time and cost that spoils the image of the developing organization. In the past, many risk mitigation factors identified, but with time, the way technology increases, threats of risk are also increases. So there is always scope for identifying risk in software development. Risk is inherited in every project at every stage of the project. A good analysis of risk plan makes a software project effective and efficient which can fulfill all requirements of the project.

A basic taxonomy of risk analysis is shown in Figure 1. Before doing a risk analysis of a project, some primary questions arised. The risk analysis approach can answer these questions. Past experiences are a proactive approach for tackling risks and you get insight from it. Risk analyst has a knowledge base about the techniques to protect software like hardware base protection, adding watermark, checksum, cryptography for protecting data and guards [4]. Knowledge is needed about the method when risk analysis methods are applied to some projects. The method will be qualitative or quantitative, depends on the type of project [5]. Keeping in mind all the taxonomy, this research provides an approach that made the risk analysis more reliable.

This research paper puts attention on software risk factors. Widely affecting risk features are studied and perceive the four most harmful factors from them. These factors are useful insights for project managers while developing the software. A detail debate on practical problem and practical solution of an application were discussed by using these factors. Empirical studies are done by surveying, and results are extracted. The final results show that what's the main source of these four key risk factors and how to tackle these types of risk. The impact of each one risk factor is also discussed so that the project manager handles these risks based on their consequences respectively.

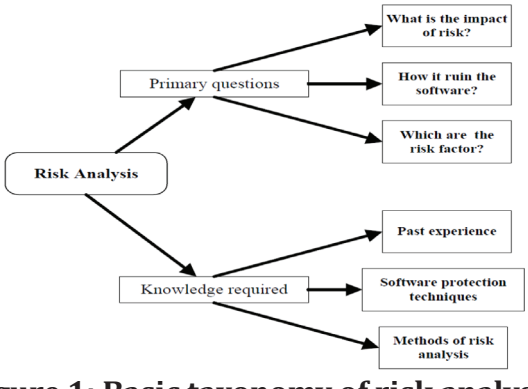


Figure 1: Basic taxonomy of risk analysis.

This paper has further divided into five sections. Section II describes the literature review and what studies are discussed in the past. Section III describes the methodology that how the collection of data can be used qualitatively so that risk analysis has been proved mathematically. In Section IV, results have been discussed and analyzed how these improve the risk assessment. In the final Section V, the paper sums up the whole study and discusses the conclusion and future work.

2. Review of Literature

The reason behind the delay in the completion of the project is mostly the lack of risk management techniques [1]. There is a need to minimize risk from every aspect of the software. Every software has some outcomes which are based on the customer requirements and market needs; which are difficult to handle both at a time. There is a need for empirical study time focus and managing strategy in every process of software development [3]. The studies at [2] have a systematic review type paper and have some limitations that the biases of interviewers and get the bad result from them. There is also a limitation of the area. This survey is done in a city of Denmark so every area has its specific pros and cons. Previously, many research papers published in the domain of software risk analysis. These papers also made a comparison between the different approaches used in the older papers. There is the use of fuzzy logic base modal for the aggregative risk management in the filed software development. It is an easy approach, but it has missing with the use of earlier models [6].

The software industry is based on the management of development. Risk management is one of the challenges of development. In [7] given the way to lessen the risk by using success parameters cost, time, people and process. According to the authors of [8], an architectural approach is used for risk analysis and alignment with agile development; and very few works implemented on it. Traditional risk assessment techniques like qualitative risk modal is bulky and inflexible. From the author of [9], who was inline qualitative risk tool with the phases of the software project. There is always a need for a systematic risk approach in project development. In the paper [10] researchers focus on the root cause of the risk factors which show a negative impact on the project. A fishbone technique is used in it and also conducted a survey to get the results both qualitatively and quantitatively. There is a need for the countermeasure to evaluate the risk control process.

Now a days Agile and its methods are mostly used for software development. As Agile support changes in any stage of the development so chances of risks also increase. Tools were proposed for minimizing the risks in agile project [11]. A framework is presented in [12] named RIMPRO that focuses to connect the product owner and manage its activities to reduce risks. A mathematical model is used for the prediction of risks and this model

is implemented in some real-world scenarios [13]. The prediction is based on similarity analysis of the projects [14].

A model was proposed for the assessment of risks in the phases of software development life cycle and proposed model was also analyzed on SPSS software [15]. A linear programming approach was used for the risk analysis of the project. Risk is a vital part of the project. The team experience is the most effective part of risk management. As risk management is vital, it completely depends on the maturity of the team [16]. Eleven Risk factors are most important and they are the root cause of all the factors. New technology complexity changes incompleteness and ambiguity are the most common. These all factors are an obstacle to the success of the project. Poor planning and bad commitment from the customer are external factors that increase the risk of the project [17]. For increasing the development procedure expelling risk factors in the early phases of development. Risk management was done at the stage of the developed premise of risk possible to happen. The expectation of risk is from the programming designer [18].

TABLE I. Overview of Literature Review

Paper	Contribution	Limitation
[1]	Empirical analysis on selected process modal	Have not a focus on specific risk factor
[2]	How stakeholders perceived important the value of PRM	PRM is a broader term that is not completely covered
[6]	Use Fuzzy logic to evaluate and minimize risk in software	This logic is not used with earlier approaches
[7]	Risk propagates with ripple effects is not identified and eliminated.	The impact of risk on major project success parameters such as Cost, Time, People and Process.
[8]	Risk assessment and management of Agile project by using qualitative tools	Lack of appliance in industry and real projects.
[12]	RIMPRO framework was presented for managing the product owner role	The presented model was not applied to some real project
[16]	Most and less common risk management practices in Scrum projects are identifies	Need to adapt classic risk Management and level of risk aversion, and integrate it with the projects
[17]	Analyzing the 11 most mentioned factors. Relevant most is requirement risks	No assessment of these factors, seeking to reinforce the indicated results
[18]	Introducing a risk assessment step that can be automated	Automated modal cannot link with some tool.

3. Methodology

In our study, we have to find risk assessment factors so that it's made better understanding to risk. The basic hypothesis of this research is finding the dependent factors which are

the cause of independent risks. Several research papers on the topic of risk analysis, and risk management are studied. After analysis, we observe that risks are coming from starting level. For solving this problem, cause and effect analysis is used which helps us to understand the root causes of risk. A survey is also conducted to get the data from real-world and developers who practically work on projects. The proposed work diagram is shown in Figure 2.

For analyzing risk factors, a survey is conducted via questionnaire 1. The questionnaire is shared in different software houses in Pakistan and also send to some other countries where we have our references. The questionnaire was solved by senior developers who had a minimum of 3 years of experience in the development and management of the project. The questionnaire was filled by all types of developers like mobile applications, desktop applications, hybrid applications and web applications. The questionnaire is designed online on Google form so that it can be shared easily in the situation of the pandemic of Covid-19. This questionnaire was shared with almost 18 companies, and 32 developers from 14 companies responded.

The focus of our research is based on some variables that are the key cause for the risks in the project. From studying the literature, it was found that those variables are risk-related, cost, strategy, operations & technique and some unknown risks. The survey is based on these variables. The meaning of the term "unknown risks" is varying from one developer to another. Here is the discussion of responses from the survey.

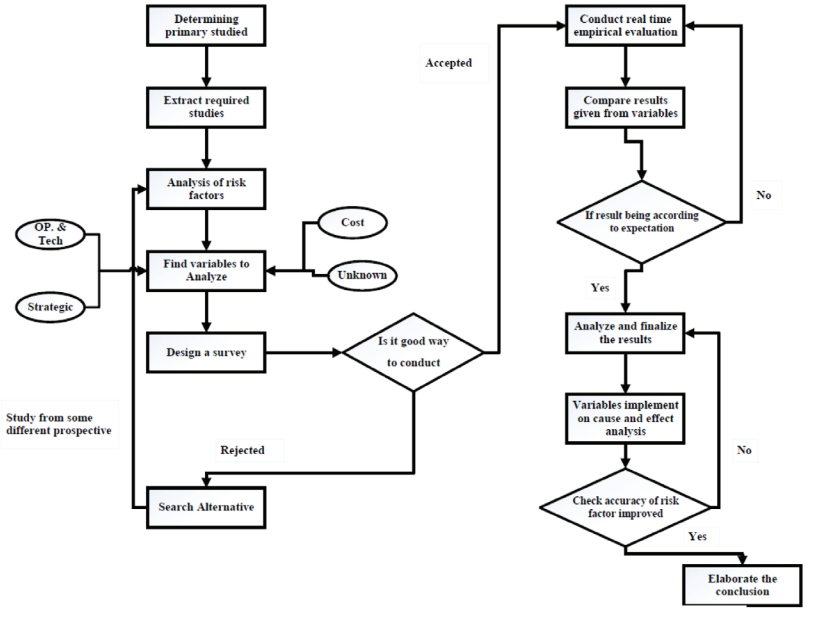


Figure 2: Work flow diagram for the proposed risk analysis approach

A. Cost Risks

As Figure 3 represents that in the response to cost risks questions, most developers think that requirements are not clear and sometimes there is a fault in the machine which can increase our expense. Scope creep is another big reason because changing requirements is a cause of the effort-loss of the developer. Sometimes the client has a low budget and it is difficult to manage the project at a very tight cost.

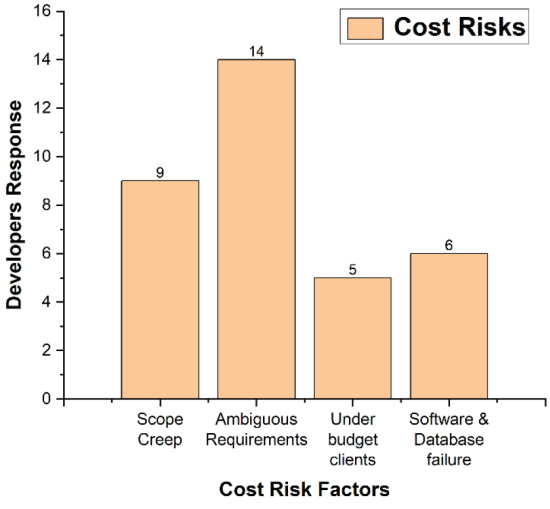


Figure 3: Cost Risks Empirical studies response chart

B. Strategic Risks

The questions linked to strategic risks tell us that information given to the developer was not complete so that could disturb the schedule. There are recurring issues in quality assurance. Sometimes the project looks simple at initial level, but in reality, it is complex due to which entire strategy becomes ruined. Some unexpected circumstances like coding errors could not be solved in the expected time which increases the duration of the project as illustrated in Figure 4.

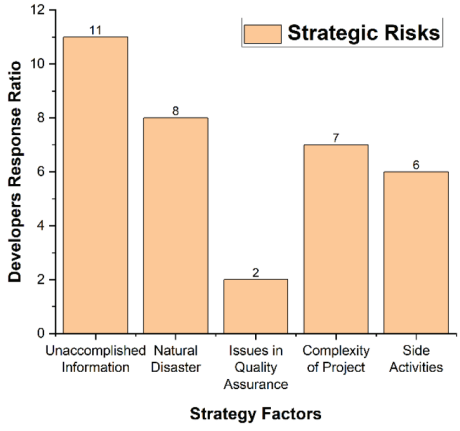


Figure 4: Strategic Risks Empirical studies response chart

C. Technical & Operational Risks

In this type of risk some technical type of risks occurs, like loss of all the program due to crash of hard disk or any drive. There are also network problems: the server become down or any cable is damaged and the project is disabled to save to the cloud, code is not saved due to the shutdown of electricity, and entire work lost. In some cases, an employee leaves the job in the middle of the project and all schedule of the company comes into the trouble. Lack of group work and unprofessional techniques also disturb the schedule of the organization. User interference issues also create risks for the project as shown in Figure 5.

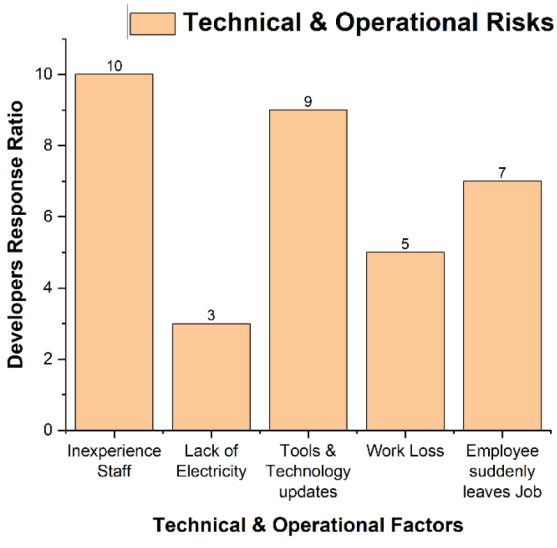


Figure 5: Technical and Operational Risks response chart

D. Unknown Risks

These are the risks which developers identify from their experiences. These risks do not have specific name or definition, and these change from company to company and developer to developer. From our responses, we get the names of risks like market risks, communication risks, risks related to office or organization, risks from the inexperienced staff and how requirement gathering and architecture are designed that increase the risk exponentially as shown in Figure 6.

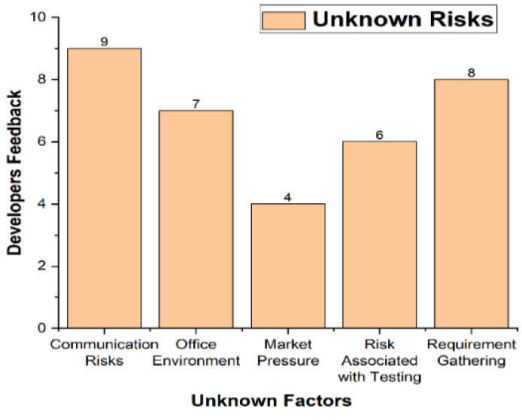


Figure 6: Unknown Risks Empirical studies response chart

4. Results and Discussions

Here are the results of the responses, we got and the analysis we did of those results with cause-and-effect diagram, also known as fishbone diagram. Our research gathers some statistics from the real-time environment of software development. We implement our responses on graphs for more understandable for software engineers. Finally, we analyzed the results of all variables as discussed above. These results are shown by cause-and-effect diagram and make it easy for the software industry to implement them on their projects and mitigate the risks. Overview of all risks is drawn here in a graphical form that shows which risk is more impactful. Figure 7 shows the graphical representation of all the risks.

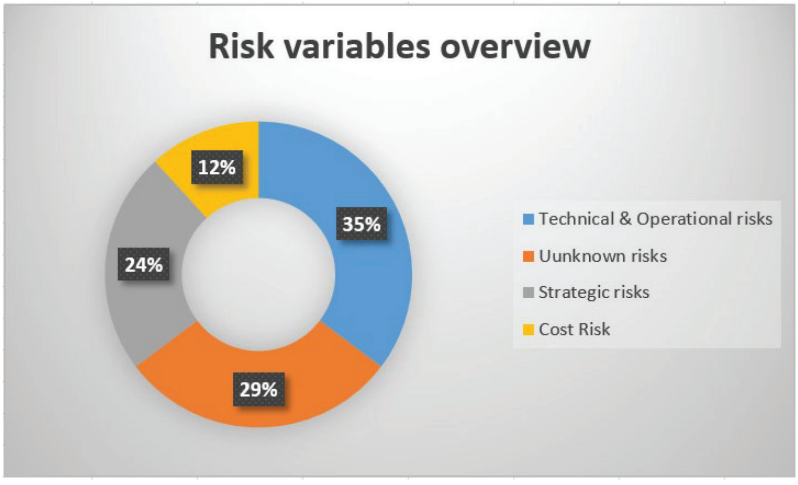


Figure 7: Risk variable overview in regarding their impact

A. Cost Risks

It is one of the basic risks of the project after doing a survey on it and analyzing its root causes which can enhance it. First of all, scope creep is the main cause of client changes requirements time by time. These changing requirements waste a lot of time and money. Time and money are directly proportional. These are waste in parallel. In the same way, ambiguous requirements are also a cause of this. Machine and software crash is also a loss of money. Figure 8 shows the cause and effect diagram. The solution is to always make a copy of the project so that in case of a crash we can save from big risks. In some cases, the client has a very low budget but he/she wants software with all functionalities. To avoid this type of fatigue we have to avoid those clients. Because working on a low budget can create tension in the mind of developers.

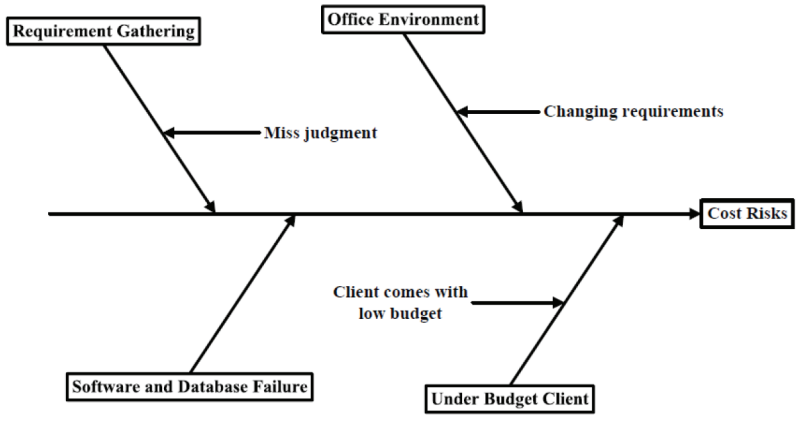


Figure 8: Cause and Effect diagram for Cost Risks

B. Strategic Risks

There are always chances of changing plans. That's the way every reputed organization has more than one plan. Incomplete information was given by the client; stakeholders or developers can create schedule changes. Before starting a project, a clear concept of the software is needed. Those companies who cannot focus on the main project and give time to other projects whose deadline is very far, also have a problem with the schedule. The company needs to have an eye on the project deadline. A very main cause we found on analyzing the cause and effect diagram is natural causes. Sometimes, the developer sees the sudden death of his relative or experiences some illness due to seasonal changes. The organization has no control over this type of cause which is why space for natural causes is always needed. One more cause from the developer's view is that some projects look simple, but they are very complex and disturb all the strategies of the company as represented in Figure 9.

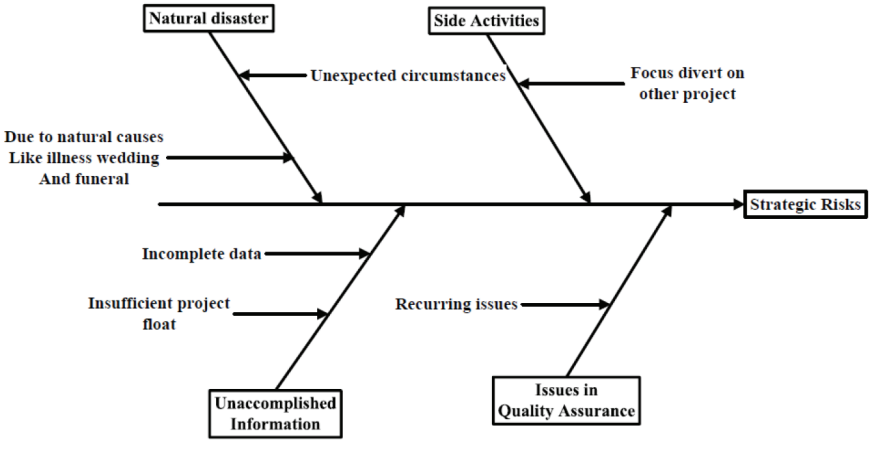


Figure 9: Cause and Effect diagram for Strategic Risks

C. Technical & Operational Risks

The cause and effect diagram shows that technical and operational risks are very impactful for the software. As shown in fig. 6 it has a 35% impact on overall risk factors. In an organization, every employee is a pillar of the company. If an employee leaves the job all operation of the company is out of order. There is a need for an agreement with an employee so that he cannot be in the middle of the project. Due to network problems, lack of electricity and burning of the machine can cause a technical problem for the software. The project always needs a backup to overcome these risks. For working on the main modules of the project, placing experienced staff are optimal solutions. The inexperienced employee can leave some bugs in the project. These are not good for the reputation of the company as shown in Figure 10. Tool and technology usage in software is upgraded

day by day. Technology has innovation every day. Devices and software are a big issue of compatibility. APIs, IDEs, frameworks and languages have a new update after a year. So it needs to use applications and tools which support every version or has some option of update. Because it's a very alarming cause of the risks of the project.

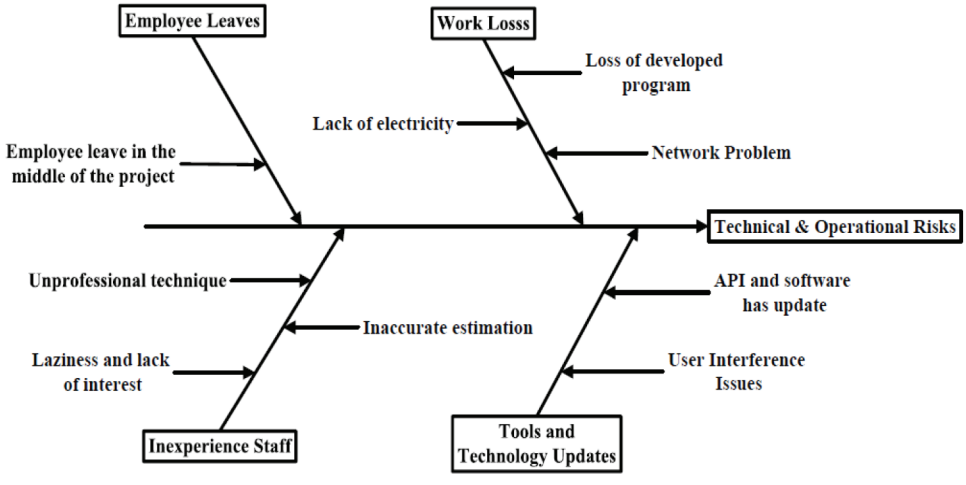


Figure 10: Cause and Effect diagram for Technical & Operational Risks

D. Unknown Risks

Different developers working on different languages have their own experience. Companies have their risks on behalf of their experience. After analyzing this and making a cause-and-effect analysis, most companies focus on the requirement engineering process. Mistakes in the requirement gathering process can cause big risks for the project. Some developers say that the office environment has caused some risks like bad behavior of manager, bad power system of the office, and there is no concept of group work. The manager needs to measure these things and improve on them by taking suggestions from employees and experts as depicted in Figure 11. Lack of communication is a fluently discussed topic in organizations. Developers have an opinion about that, it can divert project direction. Requirements are not clear if there is a lack of communication and ambiguous modules are made. It is very difficult to join these modules. It creates a gap between the stakeholder and developer, and the company cannot get expected outcomes. While testing the software, an experienced tester is needed. Who can test the software and tell the bugs to the developers in a clear way.

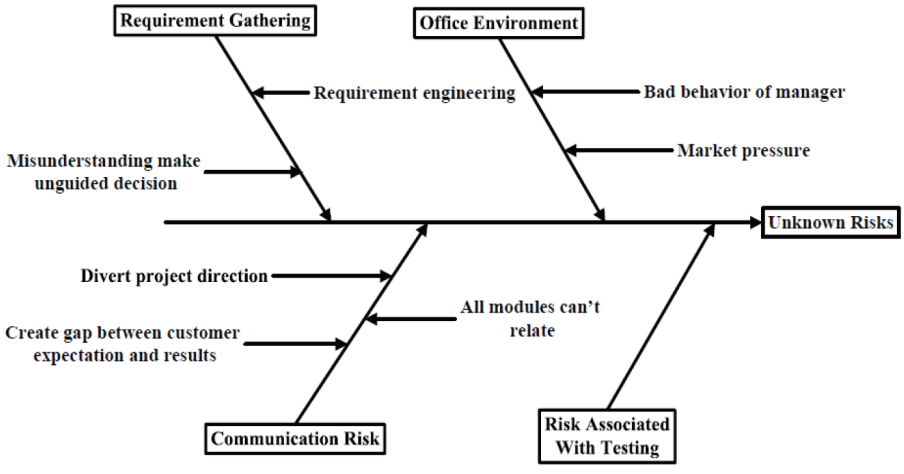


Figure 11: Cause and Effect diagram for Unknown Risks

5. Conclusion

The quality and success of the project are based on the risk assessment of the project. The focus of this research is on risk assessment, finding its root causes and analyze of root risk factors. The paper also describes the consequences of that risk on the project. Cause and effect diagram analysis is used for finding the causes and their effects of different types of risk. Our studies describe four risk factors that are faced by every software project. These four variables help software organizations how they can reduce risk factors, and they can also find causes that are based on this type of risk.

For future work, methodology can be extended and integrated with real-time applications to evaluate risk using these factors. Research presented here find causes and effects of risk factors in different software houses of Pakistan by changing the location results, these might be changed.

Acknowledgment

The Authors elegantly acknowledge, "UIIT, PMAS-Arid Agriculture University" for their help and support.

References

[1] R. Bista and S. Karki, "A New Approach for Software Risk Estimation," 2017.

[2] P. Willumsen, J. Oehmen, V. Stingl, and J. Geraldi, "ScienceDirect Value creation through project risk management," Int. J. Proj. Manag., vol. 37, no. 5, pp. 731–749, 2019, doi: 10.1016/j.ijproman.2019.01.007.

[3] R. K. Bhujang, "Analysis of risk in software process models," no. Iciss, pp. 199–204, 2017.

[4] Memon, Jan M., et al. "A study of software protection techniques." Innovations and Advanced Techniques in Computer and Information Sciences and Engineering. Springer, Dordrecht, 2007. 249-253.

[5] S. S. Muhammad, K. Weyns, and M. Höst. ''A review of research on risk analysis methods for IT systems,'' Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering. 2013.

[6] K. Bansal and H. Mittal, "ANALYSIS AND EVALUATION OF SOFTWARE AGGREGATIVE RISK USING," pp. 172–176, 2014, doi: 10.1109/ACCT.2014.99.

[7] R. K. Bhujang, "Propagation of Risk across the Phases of Software Development," pp. 508–512, 2018.

[8] M. G. Jaatun, "Architectural Risk Analysis in Agile Development of Cloud Software," pp. 295–300, 2019, doi: 10.1109/CloudCom.2019.00050.

[9] V. Anes and R. Santos, "A New Risk Assessment Approach for Agile Projects," pp. 67–72, 2020.

[10] M. T. Riaz, "Risk Assessment on Software Development using Fishbone Analysis," 2019.

[11] Tavares, Breno Gontijo, Mark Keil, Carlos Eduardo Sanches da Silva, and Adler Diniz de Souza. "A risk management tool for agile software development." Journal of Computer Information Systems 61, no. 6 (2021): 561-570.

[12] S. Lopes, R. Gratão de Souza, A. Contessoto, A. Luiz de Oliveira, and R. Braga, "A Risk Management Framework for Scrum Projects," vol. 2, no. Iceis, pp. 30–40, 2021, doi: 10.5220/0010448300300040.

[13] B. Gontijo, C. Eduardo, A. Diniz, and D. Souza, “Risk management analysis in Scrum software projects,” vol. 26, pp. 1884–1905, 2019, doi: 10.1111/itor.12401.

[14] A. S. Filippetto, R. Lima, and J. L. V. Barbosa, “A risk prediction model for software project management based on similarity analysis of context histories,” Inf. Softw. Technol., vol. 131, no. December 2020, 2021, doi: 10.1016/j.infsof.2020.106497.

[15] Nayak, Malaya K., and Arka K. DasMohapatra. "MANAGEMENT OF QUALITY IN IT PROJECTS: A RISK ASSESSMENT MODEL."

[16] K. Adimora, “Application of Linear Programming for the Optimization of Software Risk Assessment Model (Osram),” vol. 5, no. January, pp. 118–126, 2021.

[17] J. Menezes, C. Gusmão, and H. Moura, “Risk factors in software development projects : a systematic literature review,” pp. 1149–1174, 2019.

[18] G. Nasreen, M. Asam, and S. Afsar, “Automated Risk Analysis Software Model for Enhanced Software Development,” vol. 28, no. 6, pp. 5197–5200, 2015.

Facial Expression Recognition Using Weighted Distance Transform

Syed Muhammad Rafi[1] Shahzad Nasim[2]

Sheikh Muhammad Munaf[3] Mohsin Khan[4]

Abstract

Facial emotions of humans transfer non-verbal signals which have a dynamic role in interactive communication. Human-machine interface evolves according to facial expression recognition because both have a significant relationship. Psychology, ethical science, and robotics are necessary applications of facial expression recognition. A lot of work has been done already on feature extraction, face detection, and the famous techniques used for expression recognition. Weighted distance is the basic method of this research. It is used for recognition of all basic human emotions, such as anger, happiness, disgust, fear, neutral, sadness, and surprise. For the extraction of weighted distance paths, a fast-marching algorithm is used, and the seed point is taken on the nose tip of the human face. Diverse number of paths have also been taken, and they have had an effect on facial expression recognition. Intensity variation is the main motivation to use Weighted Distance Transform. Because the facial intensity variations or facial curvatures of most human beings are different, the accuracy of final evaluation may be increased and achieved in a respective manner by applying it. JAFFE (Japanese Female Facial Expression) database is used, and it is composed of 213 facial images of 10 Japanese female models with all seven basic emotions. The dimensions of JAFFE database are 256x256, and all the images are frontal position view. Twenty points are labelled for the calculation of feature vectors. Different mathematical measures are calculated as a feature vector of this geometric representation. Diverse seed locations are also being taken during research. Total four seed locations have been taken, and dissimilar number of points have also been applied for achieving better grades in the final evaluation. In classification, KNN is used and it illustrates reasonable results. In the end, validation is done with famous techniques of facial expression recognition.

Keywords: Weighted Distance, Human Behavior, Psychological Aspects, Euclidean, Fast Marching, HCI, Robotics, KNN

1. Introduction

Human faces are entities of great status in our daily lives. They present us with the

[1,3,5]*Department of Software Engineering, Faculty of Engineering, Science, Technology and Management, Ziauddin University, Karachi*

[2]*Department of Management Science and Technology, Begum Nusrat Bhutto Women University, Sukkur*

personality of the soul we are looking at, and convey information on attractiveness, age, and other traits. Neuroscience, social psychology, and cognitive science are the main applications of human emotions. They play a prominent role in human cognition [1]. These emotions of human expression could also reflect a vital role in one to one communication [2]. By using facial expression, sign languages encode the part of grammar [3]. Darwin proposed the basic rules of expression, and grouped various kinds of expressions into similar categories like hatred-anger, low spirit-dejection etc. He also stated that human expression facial fit in with human evolution [4]. Following figure-1 showed the idea of Darwin's thinking in the 19th century. It illustrated the behavior of a person with electrical equipment against the normal kind of joke.

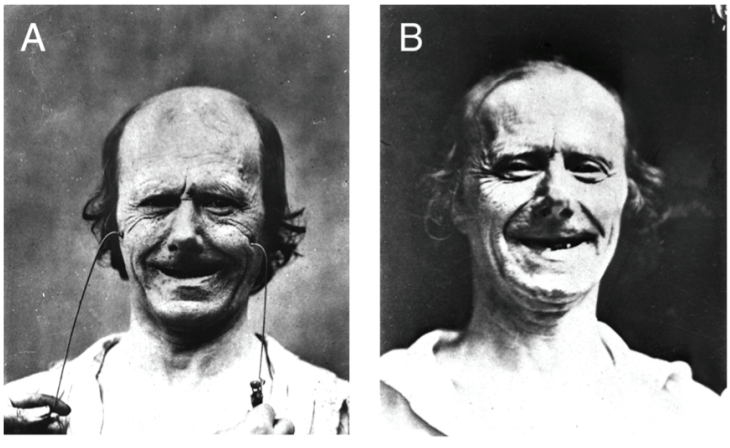


Figure 1: (A) Smile produced when zygomatic major muscles were electrically stimulated (B) Smile generated when subject was told a joke

Facial expression of emotion is much involved in the advancement of many scientific areas. It is because computer vision and machine learning researchers are concerned with developing computational representation of the perception of human face emotion to assist studies in the above sciences [5]. Computerized models of human emotions are very necessary in human computer interaction (HCI) systems, and also the key development of artificial intelligence [6]. Facial expressions are an explicit means by which people accommodate to their social ecology [7]. Face recognition has a vital value of security issues but facial expression recognition always has above hand over face recognition. Psychological research [8], match to distinct universal emotions classified six facial expressions: fear, anger, happiness, disgust, sadness, and surprise as seen in following figure-2.

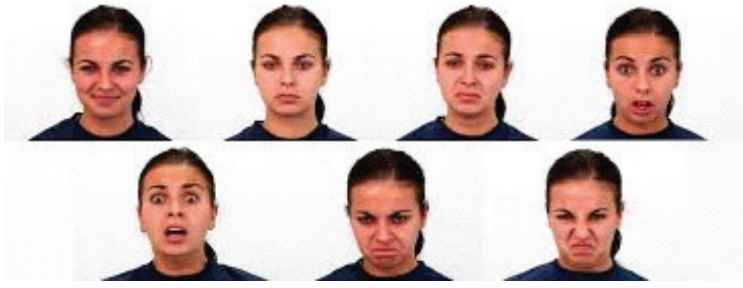


Figure 2: Different Human Emotions

As discussed above, facial expression has been studied by psychologists, clinical practitioners, and also actors and artists who are interested to read about facial expressions to enhance their capability.

The famous book of Le Brun "The Perfect Imitation of Gemini Facial Expression" [9] was the key major book to achieve the best artistic ability in the 18th century. However, over the last quarter century, with the advances in the field of computer vision, animators, computer graphics, and robotics, computer scientists started delivering great interest to explore facial expressions. Shape characterization on microscopic images is also a reasonable approach in medical diseases evaluation [10]. Robotics is also the main factor of facial expression recognition, especially in humanoid robots. Many scientists believe that robots are the final destination of the facial expression recognition system. As the robots begin to interact more and more with humans, they need to develop extra and sharp intelligence in terms of understanding human moods and emotions. This is the basis of human computer interaction (HCI) community to build computers close to humans. Robots and affect sensitive HCI have also opened a new domain to use expression recognition systems in Animations, Telecommunication, Video Games, Automobile Safety, and Education-related Software etc. [11].

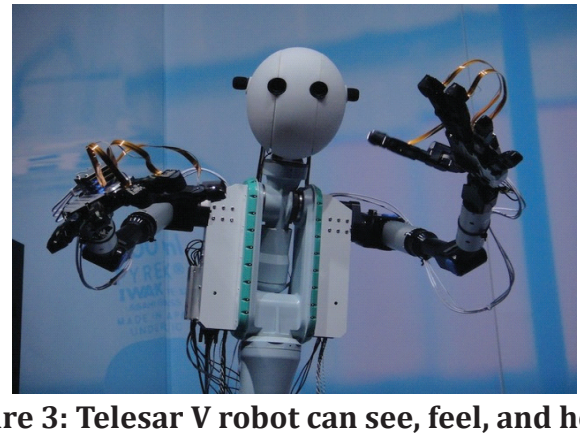


Figure 3: Telesar V robot can see, feel, and hear

For facial expression recognition, there is the available wide range of databases on spontaneous and posed emotions. In this research, JAFFE [12] database is used which

consists of 213 images of 10 Japanese female models. This database covers all major emotions, including anger, disgust, happiness, fear, sadness, surprise, and neutral. These images are captured in the Psychology Department at Kyushu University.

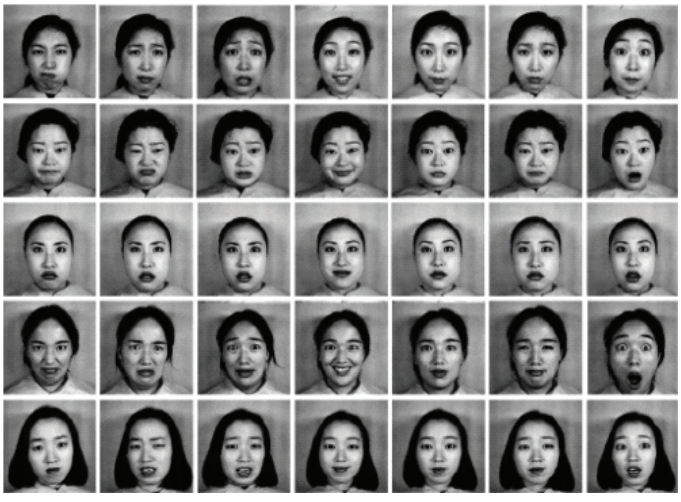


Figure 4: Anger, Disgust, Fear, Happy, Neutral, Sad, Surprise (JAFFE DATABSE)

Weighted distance is widely used for 2D images, surface segmentation, meshes, and feature extraction in different research articles [13, 14, 15]. For the extraction of facial curvatures of 2-D frontal face images, weighted distance is used because it retains intensity curvatures which works in the spatial-intensity domain. It considers the local intensity variations and also uses spatial distance between neighboring pixels. The weighted distance includes redundant information related to facial intensity curvatures. For extracting exact features, multiple paths are drawn using weighted distance transform. For the calculation of weighted distance paths, the starting point is required, and it is called the seed point. Paths are extracted by taking different seed points on the faces.

A wide variety of information is included in these paths. That's why it is very essential to parameterize the algorithm. For this purpose, different feature vectors are extracted. The dimensionality of these feature vectors is also reduced for achieving actual information. After parameterization, these feature vectors are used for the final calculation of recognition rates. For this purpose, we used the Euclidean distance for classification and in the end, comparisons are also done for the validation of method. Later in this research, a literature review and methodology is defined, then the result is shown in a respective manner to comprehensively cover a whole area of research and terminate it with references.

2. Literature Review

Facial expression recognition has attracted researchers because of its diversity. Even face recognition [16] also have respective involvement but facial expressions evaluated in many

aspects. Both Face recognition and facial expression recognition have lots of similarity in concurrence to approach. Many people from other fields, including psychologists, neurologists, and computer scientists are involved in it. Action Unit Detection in particular and emotion recognition in general has been studied broadly in the last few decades. It is not possible to review whole field comprehensively here. That's why only relevant works appear and focusing on related methods. There are two major approaches for human emotion recognition that cover most of the area of facial research. They are geometric based approach and appearance based approaches.

2.1. Geometric based Approaches

Geometric based approaches use location, distance, angle, and other relations between the face components. In this approach, it is important to discover the exact location of the face components [17]. Most of the researches on above approach are mostly about Facial Action Coding Units (AUs)[18]. AUs were mostly based on facial muscle movements. Zheng and Ji [19] also used above approach using Dynamic Bayesian Networks (DBNs). They detected 26 face features by marking around the areas of eyes, nose and mouth. The work of Kotsia and Pitas [20] also shows some significant effect. They used candid grid nodes to the facial landmarks to build a facial wire frame model for human emotion recognition and for classification purpose used Support Vector Machine (SVM). Valster et al. [21, 22] declared that geometric based systems are better than appearance based approaches. They used fiducial point on the face to extract geometrical features.

2.2. Appearance based Aproaches

Appearance based approaches use the texture or color arrangement of whole or some part of the image. In another way, this approach is mostly holistic. The local features of appearance based are much easier to calculate. Ahonen et al. [23] proposed a Local Binary pattern (LBP) method for still images. LBP was proposed by Ojala et al. [24], used texture analysis and achieved better results. Gabor filter is also very famous holistic and appearance based approaches. Gabor filters are time and memory intensive [25] for facial representations. In holistic, whole image is given as an input. Edwards et al. [26] used principal component analysis (PCA) to create Active Appearance Model (AAM). They constructed a multivariate multiple regression for modelling the relationship between the AAM displacement and the image differently. It also matched the AAM to input image in recognition phase. Images in holistic are constrained to be normalized and properly aligned. Holistic approaches also perform better in face recognition techniques [27].

2.3. Hybrid based Approaches

Holistic and local features are merged into the Hybrid approaches. For the representation

of face Yoneyama et al. [34] used a hybrid approach. For a normalized facial image, Yoneyama et al. [34] fit an 8x10 quadratic grid and in the every 8x10 regions. Geometric and Appearance are also combined to make Hybrid based approaches [28]. Weighted distance is also used but only for 3D images. It was never used before for 2D images. As described above, intensity variations are different of every two persons. That's why this algorithm got better results than previous approaches. Following figure-5 shows the implementation steps of the proposed method.

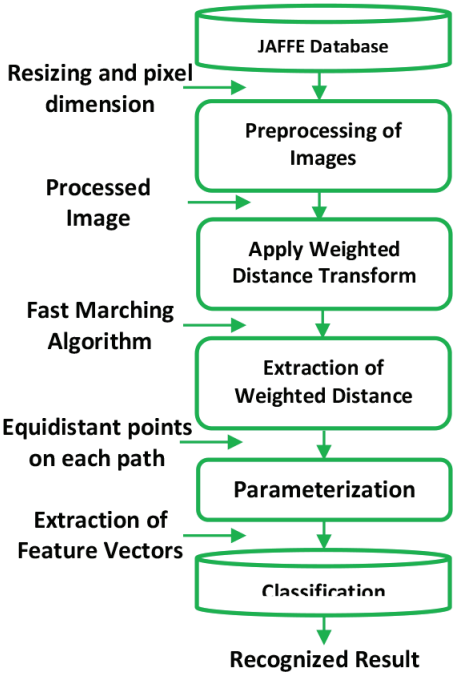


Fig 5: Implementation Steps

3. Proposed Method

The main theme of this research is facial expression recognition. Weighted distance is a very powerful technique and fast marching algorithm is used as described above for the extraction of geodesic paths. The method covered all seven major emotions that are described according to behavioral neurosciences: angry, happy, sad, fear, surprise, disgust, and neutral. The intensity variation of facial curvatures is always different of every two persons. Weighted distance only focuses on intensity variations of facial curvature and that is the requirement of this algorithm. Because of this intensity variation in human faces, we can capture more and more information which is very helpful to recognize images accurately.

3.1. Pre-processing of Image Database

The 2D images of human faces retain the intensity of pixels. Facial intensity curvatures

could be extracted from the 2D image through intensity variations of face. The resizing of the given particular image is necessary if it is not similar to the stored images. But JAFFE database that is used for the algorithm is already contained 256x256 pixels that are the requirement of our algorithm. The calculation of the weighted distance transform requires a point and this seed point is located at the nose tip on frontal faces. Algorithm took 200 images out of total 213 images of JAFFE database. 20 images of 10 persons are characterized to recognize the facial expressions. For all individual emotions, data sets are organized in all expressions separately. In this step, happy, angry, sad, surprise, fear, disgust, neutral are arranged in 3 to 4 images of every emotion along static image in each directory of facial emotions images. All images are set to 256x256 pixels, which is already the original dimension of this database.

3.2. Apply and Extract Weighted

Distance Transform

The weighted distance transform is an effort to extort intensity curvatures from a 2-dimenasional image. So the method depends on facial intensity curvatures. A fast marching algorithm is used for the extraction of paths. This fast marching algorithm [29, 30] was first introduced by James A. Sethian for solving boundary value problems which mostly related to closed curves.

$$F(a)|T(a)|=1$$

The starting point was taken from the nose tip of the frontal face because it is the requirement of weighted distance transform. This starting point is actually the seed location of database images and this work must be done manually. The algorithm took four dissimilar locations on the face to examine the facial intensity curvatures contained in geodesic distance. The weighted distance or geodesic distance calculated from every location of seed point can be seen in fig 6.

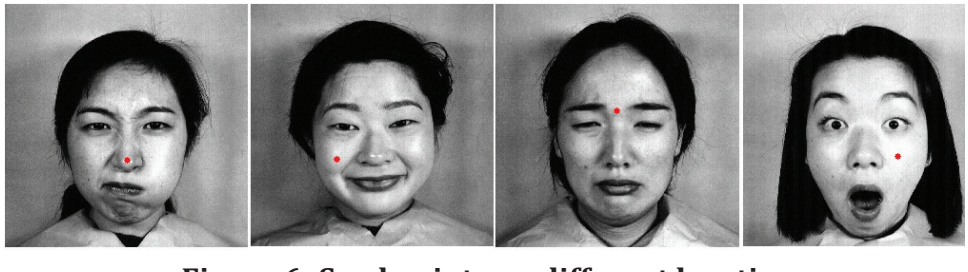


Figure 6: Seed points on different locations

As the distance transform resulted in redundant intensity curvatures, there is a need to extract more weighted distance paths in the face. It is a further attempt that extracts

additional discriminating intensity curvatures. For calculating paths, the algorithm needs ending points in addition to their corresponding seed points. We take distinct loci of end points at the elliptical edge of the face to cover the whole area of the face. Firstly, we have taken 10 paths but for obtaining sufficient information and improving the result we took 20, 30, 40, 50, 60, 70, 80, 90 and 100 paths on each individual emotions of the human face. Weighted distance and calculation of paths on the face are also extracted by using the same fast marching algorithm as seen in fig 7.

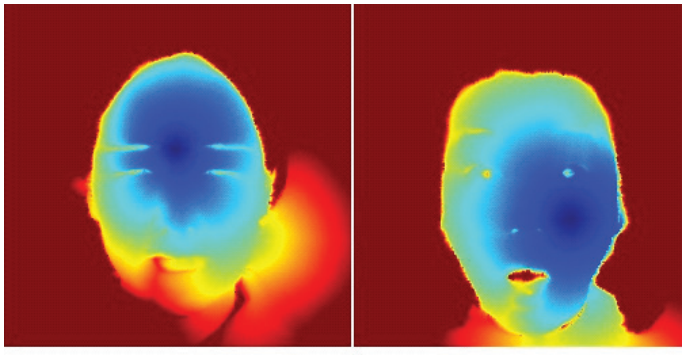


Figure 7: Weighted Distances of above seed points

3.3. Parameterization

For comparison, weighted distance paths needed to be parameterized. Selective curvatures of a path retain the main information regarding facial intensity curvatures, the parametric illustration of a path should effectively reflect the discriminating curvatures in order to enable facial expression recognition. Firstly, the algorithm is parameterized by taking two equidistant points on each path. As the number of points (taken on a path) has a direct relation to the contents of curvatures reflected in parametric representation, that's why more points are taken on paths can be seen in fig 9.

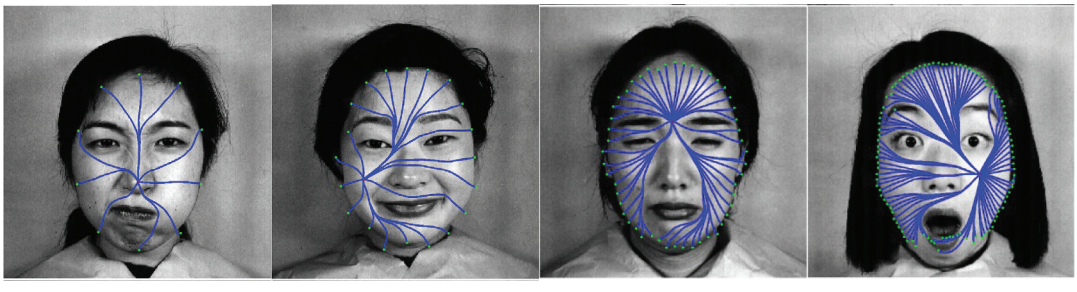


Figure 8: Paths with different seed locations

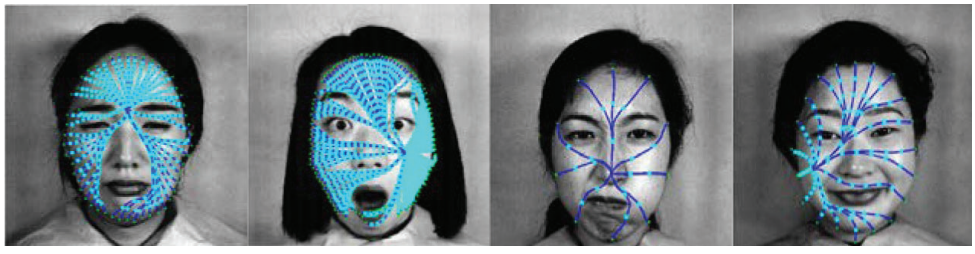


Figure 9: Equidistant points on each path

The algorithm obtained different kinds of information for every point taken from the parametric representation of a path. Coordinates, displacement from seed point, distance along the path between two successive points, and displacement between equivalent points of contiguous paths are calculated after the parameterization process. This collected information is utilized for constructing a feature space of each image.

3.4. Classification

Above feature vectors are used for classification. For this KNN (K-nearest neighbor) is used, in which the classification is done by the matching of neighboring pixels. Euclidean distance is actually following the KNN approach. Following equation represents the Euclidean distance between two points. The result showed the classification in terms of different parameters.

$$d(p,q) = \sqrt{(p_1 - p_{2)}{}^2 + (q_1 - q_{2)}{}^2} \quad (Eq\ 2)$$

4. Results

The method is evaluated on the basis of different parameters. The evaluation is performed using four images for every individual emotion (anger, happiness, disgust, fear, surprise, sad, and neutral) of a person, where two images are taken for training and the remaining two for testing. The intensity curvatures are represented by the facial curvatures of the path retained in weighted distance. By increasing the number of paths, the representation of intensity curvatures are also improved. For this purpose, different numbers of paths are drawn on the face and algorithm took 10 to 100 paths. The representation of paths is also based on the number of equidistant points and in this research 2 to 20 points on each path are selected for showing its parametric representation. The result showed that if the number of points increases, the results became much better. The location of seed points also affected the weighted distance and four different locations for seed points have been taken. Each seed point generated different patterns for each location, as can be seen in figure 8. The effects of different feature vectors that were extracted through weighted distance transform also have significant impact on the algorithm. As discussed above, the feature space of the images are constructed through these feature vectors.

Table.1 showed the classification accuracy of all basic human emotions taken from JAFEE database, and the graph (fig 10) also presented the accuracy proposed method against the LAP (local arc pattern).

Table.1 showed the classification accuracy of all basic human emotions taken from JAFEE database, and the graph (fig 10) also presented the accuracy proposed method against the LAP (local arc pattern).

Table 1: Accuracy of Every individual

Emotions	Proposed Method	Local Arc Pattern (LAP)
Angry	97	100
Disgust	95	93
Fear	87	84
Happy	98	96
Neutral	99	100
Sad	98	87
Surprise	96	96

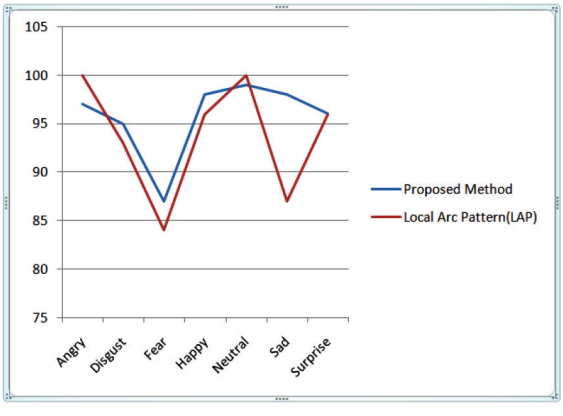


Figure 10: Comparison against LAP

5. Comparison

The Table 2 reflected the comparative evaluation where the better performance of the proposed method in terms of recognition rate can be seen. Local Arc Pattern (LAP) of Shahid et al [31], Gabor filter of Guo and Dyer [32] and M.R. Mahmood et al. [33] are compared with the proposed method.

Table 2: Comparison with LAP and Gabor Filter

Researches	Methods	Classifier	%
Proposed	Weighted Distance	Euclidean	95.71
Shahidulet al. (2013)	LAP	SVM	94.41
Guo and Dyer (2003)	Gabor Filter	Linear programming	91.00
M. R. Mahmoodet al. (2021)	Chi Square	RF/KNN	94.2

6. Conclusion and Future Work

This research showed the facial intensity curvatures could take part in a vital role in facial expression recognition. The extraction of intensity variations of human faces by using weighted distance transform showed capable results. The determination of exact loci of end points of paths created the major problem if using other databases but here it showed better results. Comparison against LAP and Gabor Filter showed comparable results which suggested that local features extraction played important role in facial expression recognition. Robustness and accuracy also increased when taking more reliable database and also some more work on facial face detection methods.

This work based on static or posed images using weighted distance transform while in future it can also be done on spontaneous images, which is actual requirement of psychological aspects. Also for face detection automatic ways are also used to achieve better result. Even if semi-automatic way is used in this research for face detection but more automatic may get better results.

References

[1] Damasio A. R., Descartes error: Emotion, reason, and the huma brain, 1stedn, G. P. Putnams Sons, (1995).

[2] Schmidt K. L. and Cohn J. F., Human facial expressions as adaptations: Evolutionary questions in facial expression research, Amer. J. of phy. anthro., 116(33), 3-24, (2001).

[3] Wilbur R. B., Nonmanuals, semantic operators, domain marking, and the solution to two outstanding puzzles in asl, Sign Lang. & Ling., 14(1), 148-178, (2011).

[4] Darwin C., The expression of the emotions in man and animals, 3rdedn, Oxford Univ. Press, (1998).

[5] Martinez A. M., Matching expression variant faces, Vision Research, 43(9), 1047-1060, (2003).

[6] Pentland A., Looking at people: Sensing for ubiquitous and wearable computing, Pattern. Analy. and Mach. Intelli., IEEE Trans., 22(1), 107–119, (2000).

[7] Martinez A., Du S., A model of the perception of facial expressions of emotion by humans: Research overview and perspectives, The J. of Mach. Lear. Resear., 98888, 1589-1608, (2012).

[8] Andersen E. N., Unconscious processing of emotional content in hybrid faces, Ph.D. thesis, Inst. of Psycho. Univ. of Oslo, 2011.

[9] Montagu J., The Expression of the Passions, final edn, Yale Univ. Press, (1994).

[10] A. Kokou M., and Antoine V., Shape characterization on phase microscopy images using a dispersion indicator: Application to amoeba cells, Res. J. of Comp. and Info. Tech. Sci., 1(5), 8–12, (2013).

[11] Bettadapura V., Face expression recognition and analysis: the state of the art, arXiv preprint arXiv,1203.6722, (2012).

[12] JAFFE database, http://www.kasrl.org/jaffe.html, (Last accessed 30/11/2020).

[13] Peyre G. and Cohen L., Surface segmentation using geodesic centroidal tesselation, Proc. 2nd Intr. Symp. on 3D Data Proces., Visuali. and Trans., (2004).

[14] Toivanen P. J., New geodosic distance transforms for gray-scale images, Pattern Recog. Letters, 17(5), 437-450, (1996).

[15] Ahsraf M., Sarim M. and Shaikh A. B., Raffat S. K., Siddiq M., Face recognition using weighted distance transform, Res. J. of Rec. Sci., (2013).

[16] Sadeghi R., A comparative face recognition algorithm for dark places, Res. J. of Rec. Sci., 2(9), 92–94, (2013).

[17] Shan, C., Gong, S., &McOwan, P. W., Robust facial expression recognition using local binary patterns,ICIP 2005, IEEE Intr. Conf. on Image Proces.,2, II-370, (2005).

[18] Ekman P. and Friesen W. V., Facial action coding system: A Technique for the Measurement of Facial Movement, 1stedn,Consult. Psycho. Press, Palo Alto, CA., USA, (1978).

[19] Zhang, Y. and Ji Q., Active and dynamic information fusion for facial expression understanding from image sequences, Patter. Analys. and Mach.Intelli., IEEE Trans., 27(5), 699-714, (2005).

[20] Kotsia I. and Pitas I., Facial expression recognition in image sequences using geometric deformation features and support vector machines, ImageProces., IEEE Trans., 16(1), 172-187, (2007).

[21] Valstar M. F., Patras I. and Pantic M., Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data,Comp. Visi. and Patter. Recog. Work., CVPR Work., IEEE Comp. Socie., Conf., 76-84, (2005).

[22] Valstar M. and Pantic M., Fully automatic facial action unit detection and temporal analysis,Comp. Visi. and Patter. Recog. Work., CVPR Work., IEEE Comp. Socie., Conf., 149-156, (2006).

[23] Ahonen T., Hadid A. and Pietikainen M., Face description with local binary patterns: Application to face recognition, Patter. Analys. and Mach. Intelli., IEEE Tran., 28(12), 2037-2041, (2006).

[24] Ojala T., Pietikäinen M. and Harwood D., A comparative study of texture measures with classification based on featured distributions, Pattern Recog., 29(1), 51-59, (1996).

[25] Bartlett M. S., Littlewort G., Frank M., Lainscsek C., Fasel I. andMovellan J., Recognizing facial expression: machine learning and application to spontaneous behavior,Comp.Visi. and Pattern Recog., CVPR 2005, IEEE Comp.Socie. Conf.,2, 568-573, (2005).

[26] Cootes T. F., Edwards G. J. and Taylor C. J, Active appearance models,Comp.Visi., ECCV,Springer Berlin, 484-498, (1998).

[27] Muhammad Sharif S. M. M. R., Shah J. H., Sub-holistic hidden markov model for face recognition, Res. J. of Rec. Sci., 2(5), 10–14, (2013).

[28] MendozaD. S., Masip D., Baró X., andLapedriza À., Emotion Detection Using Hybrid Structural and Appearance Descriptors,Model. Deci. for Artifi.Intelli., 105-116, (2013).

[29] Sethian J. A., Fast marching methods, SIAM review, 41(2), 199–235, (1999).

[30] Sethian J. A., A fast marching level set method for monotonically advancing fronts, Proc. of the National Academy of Sciences, 93(4), 1591–1595, (1996).

[31] Islam, M. S., and Auwatanamongkol S., Facial Expression Recognition Using Local Arc Pattern, Asian J. of Inf. Tech., 12(4), 126-130, (2013).

[32] Guo G., and Dyer C. R., Simultaneous feature selection and classifier training via linear programming: A case study for face expression recognition.Comp.Visi. and Patter.Recog., Proc. 2003 IEEE Comp.Socie. Conf., 1, I-346, (2003).

[33] M.R. Mahmood, M.B. Abdulrazzaq, S.R. Zeebaree, A.K. Ibrahim, R.R. Zebari, and H.I. Dino, "Classification techniques' performance evalutation for facial expression recognition." Indonesian Journal of Electrical Engineering and Computer Science, vol. 21 no.2, pp.176~1184. 2021.

https://doi.org/10.51153/kjcis.v5i1.107

EEG-Based BCI for Attention Assessment in E-Learning Environment using SVM

Muhammad Bilal[1] Muhammad Marouf [2] Safdar Rizvi[3]

Fatima Bashir[4] Muhammad Shahzad[5] Jawad Ahmed Bhutta[6]

Abstract

In this era, technology has paved the way to new dimensions of research in academia in terms of students' performance, learning outcomes, and capability. Brain-Computer Interface (BCI) has shown to be essential in monitoring students' brain activity through electroencephalogram (EEG) signals. Attention is a prerequisite to the evaluation of the student learning process. This paper proposed recognition of attention level in an e-learning environment. It was divided into two states, attention, and inattention(distracted). EEG signals were extracted using the non-invasive device (Emotiv Insight) and processed data for noise removal through the Finite Impulse Response (FIR) filter. A machine learning approach has been used for the classification of data. The data acquired through the channels is continuous for which Support vector machines (SVM) have been used for classification. The selected features are then classified. The obtained accuracy for attention level is 90.07% in an e-learning environment.

Keyword: Brain-Computer Interface, Electroencephalogram, Attention, Machine Learning

1. Introduction

Brain-Computer Interface (BCI), the understanding intended to explain the purpose through which the signals are produced. An extensive study has been done to design valuable and efficient systems that provide an effective and interactive service using human biological signals [1],[2]. The individual's neural system can help to yield these signals. Whereas the human heart can obtain the Electrocardiograph (ECG) signals [3], an individual's hand muscle can attain Electromyogram (EMG) signals and the scalp of the human brain can achieve Electroencephalogram (EEG) signals [4]. The EEG is used by BCI systems for monitoring and analyzing the signals of the human brain [3],[5].

[1]*Department of Computer Science, Bahria University, Karachi | bilalvohra.bukc@bahria.edu.pk*
[2]*Department of Computer Science, Bahria University, Karachi | mmarouf.bukc@bahria.edu.pk*
[3]*Department of Computer Science, Bahria University, Karachi | safdar.bukc@bahria.edu.pk*
[4]*Department of Computer Science, Bahria University, Karachi | fatimabashir.bukc@bahria.edu.pk*
[5]*Department of Computer Science, Bahria University, Karachi | muhammadshahzad.bukc@bahria.edu.pk*
[6]*Department of Computer Science, Bahria University, Karachi | jawadbhutta.bukc@bahria.edu.pk*

For several years, the brain-computer interface (BCI) has been a theme of research that connects the human brain and computer [6]. From previous decades, Researchers have made an outstanding pathway in practical BCI applications and made sure that these interfaces are part of several technological visions [8]. BCI is mainly contributing to attention analysis, which is a basic part of human productivity.

Attention is an important aspect of the educational environment [9] and may affect learning processes positively or negatively according to its level. In the twenty-first century, the increase of digital media has led to very rapid advancement in the use of videos in an educational medium [10]. Video-based education has become a progressively popular technique in e-learning. E-learning means learning activities through internet tools outside the traditional classroom. Due to its usefulness, learners can learn from online video via mobile devices, tabs, or computers beyond the time-space limitation [11]. Several various video-based learning platforms, such as TED, Coursera, YouTube, and MOOCs, started to facilitate different video courses for learners. The platforms which are based on education purposes usually allow teachers to upload well-prepared videos based on their teaching plans. This mode of learning formed a new learning method, which is different from the customary classroom-based or text-based learning. Thus the purpose of the research is to use EEG technology to study that in what way video-based lectures in an e-learning environment setting affect the attention level and its impact on the learner, and further apply feature extraction and machine learning technique. The outcomes of the study show the contribution by analyzing the significance of the attention factor in an e-learning environment.

2. Background Knowledge

A. Brain-Computer Interface

The Brain-Computer Interface system is schematically shown in Figure 1. It is widely used as an integrated diagnostic tool to analyze brain signals and patterns by placing electrodes on the scalp. It gives real-time data. The mental syndromes and brain patterns are identified by the general info of functional, physical, and pathological status of the brain contained by EEG for diagnosing and recording brain activity. The first prototype of BCI came out in 1973, in the laboratory of Dr. Vidal [7].

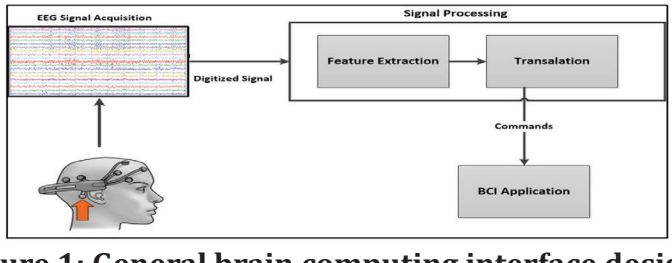


Figure 1: General brain computing interface design

In the earlier demonstration, it has reported that the EEG signals had been noticed with the fusion of different frequencies that are alpha, beta, delta, and theta signals. These types of signals are also known with multiple names like; EEG signals or brainwaves and spectral components. All these signals are visualized in Table 1 [12].

Table 1: Type of EEG spectral wave with defined ranges and their basic functions

Wave	Frequency Range	Major Functions
Delta	Up to 4 Hz	Sleep, found in attention tasks
Theta	4 Hz to 7 Hz	Halting, incompetence, related to ADHD
Alpha	8 Hz to 15 Hz	Rest, eyes closing
Beta	16 Hz to 31 Hz	Concentration, uneasy thoughts
Gamma	Above 32 Hz	Cognition

3. Literature Review

Attention imitates various activities of the human body and that is why an essential constraint of the brain. It is an activity of keeping the brain active and thoughtful as well as processing the things in the environment for particular tasks. Attention helps in gaining information for the things of better interest as well as the prediction of attention level is also a major research area nowadays [13], [14].

The attentiveness of students while learning significantly affects their learning results. The teacher's difficultly in observing the student's attentiveness to online learning can also cause a problem in determining students' attention level. The experiment on the students with EEG-related data in a controlled e-learning environment can be feasible in determining the attentiveness or inattentiveness of a student [15]. Unlike a psychologist, many BCI applications are there to analyze the attention level, which has been reported in many studies under controlled and uncontrolled environments [16], [17].

There are extensive applications and researches on determining the attention level, like exploring the learning performance and behavior of university students in English listening courses by determining their attention level through brainwave signals [18]. Provided some active and inert indications in an electronic learning environment to determine whether their attention is affecting in different circumstances studied by [19]. The excessive use of cell phones in the class can also influence the attention level of students in the learning environment [20]. Brain-Computer Interface (BCI) has been used and researched in every aspect of human life and carried out by different researchers to accomplish the research problem. This section explains different techniques used in many studies for the evaluation of students' attention via video lectures in an e-learning environment through BCI.

In [21], a review on electroencephalography (EEG) evaluation was done to observe the effects of challenge-skill balance on flow experience and the effect of flow experience on learning performance in a computer-aided environment. The outcomes determined that the challenge-skill balance of learning materials was the basis of a flow experience of learners. The classification of attentiveness and non-attentiveness in the subjects while watching the lectures via multimedia in different situations can be done using. The EEG power spectral density (PSD) features were extracted from preprocessed EEG signals in 5 frequency bands of the delta, theta, alpha, beta, and gamma and then the attentiveness and non-attentiveness were classified using the extracted features with the change in the span of a period. SVM, kNN, Ensemble, and CNN were the classifiers to identify the attentiveness and non-attentiveness situations EEG [22].

A combination of data mining algorithms like correlation-based feature selection (CFS) and KNN for the classification system has been used by [23]. The CFS+KNN algorithm has been evaluated against multiple classification algorithms such as CFS+C4.5. They measured the performance of classification using the different 3-fold cross-validation data. They collected data from 10 individuals while learning the material in a virtual distance-learning environment. The attention has been assessed on high, neutral, low levels using a self-assessment model of self-report. The interactive Brain Tagging system (IBTS) has been used to assemble the learner's attention developed. The EEG data has been used by IBTS to transform it into evaluable attention value. They recorded the individuals' attention level every second while watching a video. The constant level of attention and the variation level of attention have been envisioned while watching a video. The distinct and cooperative attention period was the outcome of this study [24].

Age is a major factor for the inspiration of visual attentiveness and reading time. The mobile electroencephalography device can measure one's consideration and attentiveness level in the reading environment [25]. In a study of [26], a system was developed to assess brainwave data. A Massive Open Online Courses (MOOCs) system and conventional techniques have been used in applicants learning. Fourier represented the brainwaves and symmetry elements in Fast Fourier transform have used to obtain Power Spectral Density (PSD) values for the analysis of data. They determined that MOOCs system in teaching methods were efficient for increasing the attention of the applicants as compared to the conventional techniques as well as providing comfortable learning for them. As an overall assessment and a part of the study, we can see in Table 2, a summary of the existing research.

Table 2: Attention level in the e-learning environment

Environment	Ref.	Sample Size	Type	Experimental Finding
E-Learning	[21]	20	Multimedia Content	Assess Challenge-skill balance and flow experience.
	[22]	8	Video	Assess attention level while applicants were instructed to watch lecture videos through multimedia.
	[23]	10	Video	Measured attention (High, Neural, and Low) based on 20 minutes of learning task and self-assessment model.
	[24]	31	Video	Analysis of proper time-period through video learning to increase the attention level of learners in learning content.
	[25]	55	Video	Measure sustained attention while students watched the same video lecture for 16 min.
	[26]	15	Video	Measure effects of attention level in learning through MOOCs system.

4. Methodology

The proposed methodology is based on an e-learning environment. The basic framework of the presented research is shown in Figure 2. The data was collected, and the learning model was designed to classify attention levels. Finite Impulse Response (FIR) filter was used to remove the noises that affect signals. Further independent component analysis (ICA) was applied to extract attention features from the noise-free signals. In the end, features were analyzed with EEG signals and tests of the participants in e-learning environments. Further, the data is classified using Support Vector Machine (SVM).

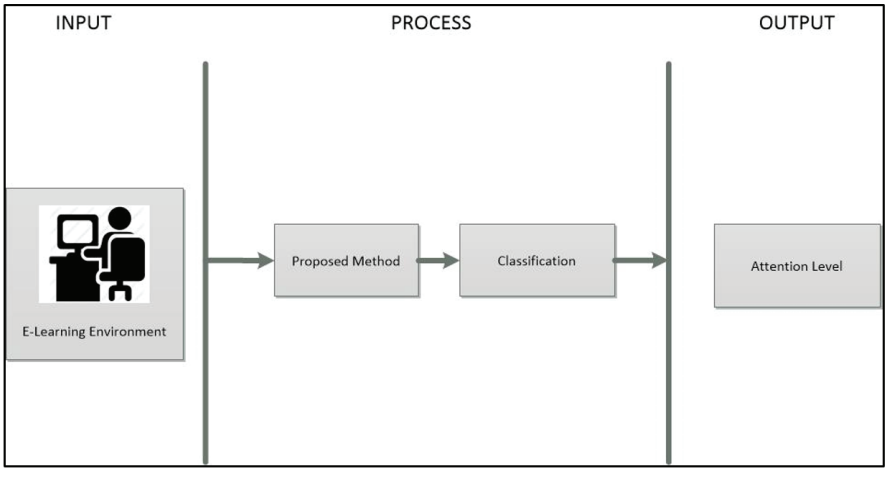


Figure 2: Basic Framework

The dataset is generated by acquiring a signal of participants. For the e-learning environment, 15 participants were selected. EEG data gathering device which has been used and the participants which were involved in the e-learning environment are described in subsections in detail.

A. Proposed Approach

Figure 3 shows the workflow of the proposed research. The first step in the proposed approach is signal acquisition. For recording signals, we used the EEG technique. EEG technique is commonly used for collecting data by the noninvasive method in which the participant wears the emotive insight headset.

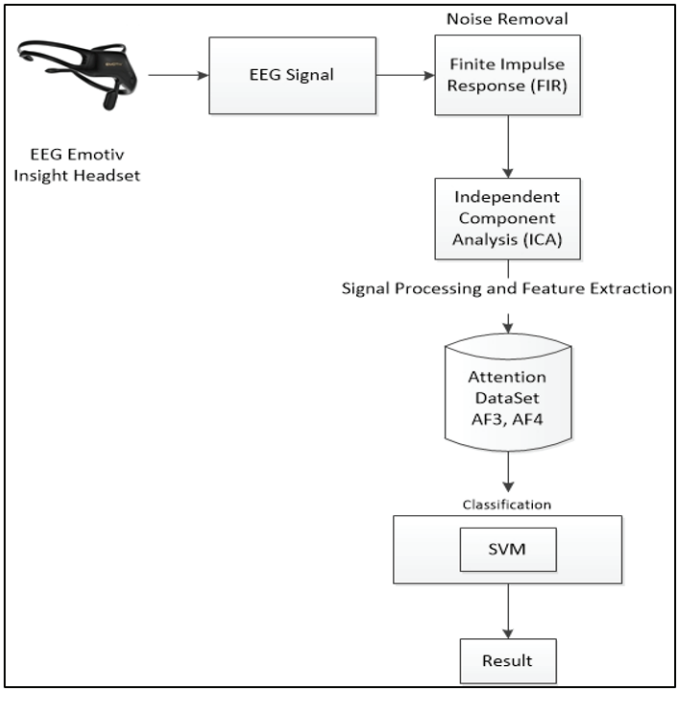


Figure 3: Workflow of the proposed methodology

B. Participants and Dataset Preparation

In this study, 15 computer science students of the 3rd year from Bahria University Karachi campus were recruited, all of whom were right-handed. Among the 15 participants, 11 were male, 4 were female, and the age was between 18 and 21 years old (Mage=18.8, SD=0.98). According to the investigation of the participants, they had no mental diseases such as epilepsy, depression, and hyperactivity disorder or did take psychoactive drugs for a long time. At present, they neither had used any drugs to change their thinking nor had any history of head injury or brain injury. The experimenter introduced the scope and

procedure of the experiment to the participants and informed them that the experiment would not cause any risk to their health, to ensure that the participants could participate in the experiment voluntarily and sign the informed consent before the experiment. Because this experiment was based on learning through videos via computer devices, the participants needed to ensure that they can see the learning content. The participants are tested one by one in the laboratory environment, each was asked to sit in a comfortable chair and watch the video-based lecture using a computer and headphones for noise-free audio. During the experiment, the participants' EEG signal was simultaneously recorded as described in Figure 4.

Figure 4: E-learning environment

C. Signal Acquisition

The brain signal has been accomplished by EEG noninvasive technique. This is a widely used signal acquisition technique due to its greater temporal resolution [28]. A portable EEG headset device (Emotiv Insight), introduced five main data-collection electrodes and two reference electrodes for signal acquisition. The electrodes were positioned at AF3, AF4, T7, T8, and Pz based on the international 10-20 systems. 128Hz was set as the sampling rate. During the experiments, each participant has been informed that they must be relaxed and calm. All healthy participants have been seated in a quiet room i.e. lab and the LCD monitor screen has been placed and the armchair to make the gazing level equal for the participant.

D. Artifacts Handling in E-Learning Environment

For the removal of artifacts in the E-Learning environment, EEGLAB has been used in the proposed research study. The data set is labeled with the names of the electrodes. The entire data set was labeled with the x-axis as time framework in-unit seconds and the y-axis for the frequency spectral power that is measured in micro-volts (μV). Artifact removal in the e-learning environment is performed by removing the entire unnecessary channel's data. Therefore, no further filter is needed to apply as it is a noise-free and

controlled environment. In the end, this new representation of the data has been used for the preprocessing of the signals.

E. Feature Extraction

E-learning environment signals consist of numerous electrodes channels including; AF4, AF3, Pz, T7, and T8. After the ELE signal enhancement phase, each component has different spectral powers with respect to its sampling rate that are measured at 128 Hz frequencies.

Signal AF4 is the spectral power that ranges from 4100 to 4300 microvolts with time series measured at 8 seconds. It is a further projectile in milliseconds for the frequency domain per subject. This time framework is constantly used on all the electrodes channels for a single subject. Signal AF3 has spectral power that ranges from 4100 to 4250 microvolts. Signal Pz has spectral power that ranges from 4155 to 4175 microvolts. Signal T7 has a spectral power that ranges from 4140 to 4180 microvolts. Signal T8 has spectral power that ranges from 4100 to 4250 microvolts.

Features are extracted based on the component spectral power, which is measured in microvolts' unit where the unnecessary components power is reduced based on dimensionality reduced algorithm as discussed above. After applying the model, a new visualization of the spectral power of all components is shown at an ELE state that results in the adjacent representation of the electrodes but has different spectral power. These electrodes signal include; AF4, AF3, Pz, T7, and T8 components. Signal AF4 has a spectral power that ranges from -100 to +100 microvolts with time series measured at 8000 milliseconds. This time framework is constantly used on all the electrodes channels for a single subject. Signal AF3 has a spectral power that ranges from -100 to +100 microvolts. Signal Pz has spectral power that ranges from -5 to +10 microvolts. Signal T7 has spectral power that ranges from -100 to +100 microvolts. Signal T8 has spectral power that ranges from -10 to +10 microvolts.

F. Classification

Classification is one of the major problems assigning one of N labels to an input signal which is newly generated, given labeled training data of inputs along with consequent output labels of them. To learn and recognize the EEG pattern, a machine learning algorithm is implied for classification which can be explained as a method for understanding the mapping or relation between EEG data classes against mental tasks such as a hand's movement [36]. Application of supervised learning is, however, a difficult task as the EEG data is noisy, and the selection of an optimum frequency band and evaluating a suitable set of characteristics are areas that still need to be solved. In addition, various degree of

attention influences the data quality, which then changes in their concentration.
Initially, the data set is recorded based on the delta, theta, alpha and beta, and gamma waves for each participant with 5 channel electrodes. For attention, two channels AF3 and AF4 data with beta (β) wave frequency [37] were extracted for further classification. Mean is calculated for each electrode from each participant's recorded data and calculate the average mean from each calculated mean of electrodes and this average mean have been utilized as labeled data for classification of an individual subject. Doing the same procedure, we have calculated all participants' average mean and finally, we have labeled the dataset with a multi-label data classification.

5. Results

We evaluated the SVM classifier by accuracy, precision, recall, and F-Measures rates. Figure 6 shows the confusion matrix, and the definitions of these performance measures are listed as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

The scores were achieved through an SVM classifier using a linear kernel. The equation for prediction for a new input using the dot product between the input (t) and each support vector (ti) is calculated as follows:

$$f(x) = B(0) + sum(bi * (t,ti)) \quad (5)$$

Individual accuracy of participants was measured by breaking the dataset into 30% testing and 70% for training while for overall accuracy for creating the SVM model, all files were merged to make a single file and separate it with 30% of the data as testing and 70% for training. Support vector machine generates in total 8703 support vectors 4351 for attention class boundary and 4352 for inattention (which is labeled as "Distracted"), a class boundary that lies on this margin to separate attention and distracted classes with the cost of 0.1. Table 3 shows the confusion matrix of the e-learning environment through which precision, recall, and F-measure were calculated. The purple color shows the attention data and inside attention class red spots are misclassified data Furthermore, distracted data is represented by blue color and the misclassified data is also present in

it which is shown as black spots as shown in Figure 5.

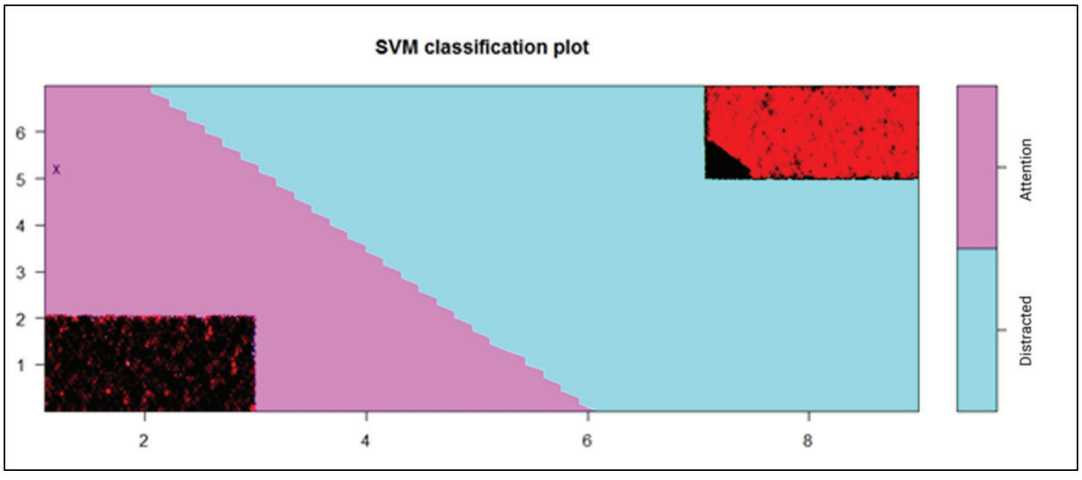


Figure 5: Attention level in e-learning environment support vectors

The value of cost (C) is large then the model chooses more data points as a support vector and thus it gets the higher variance and lowers bias, which may lead to the problem of overfitting.

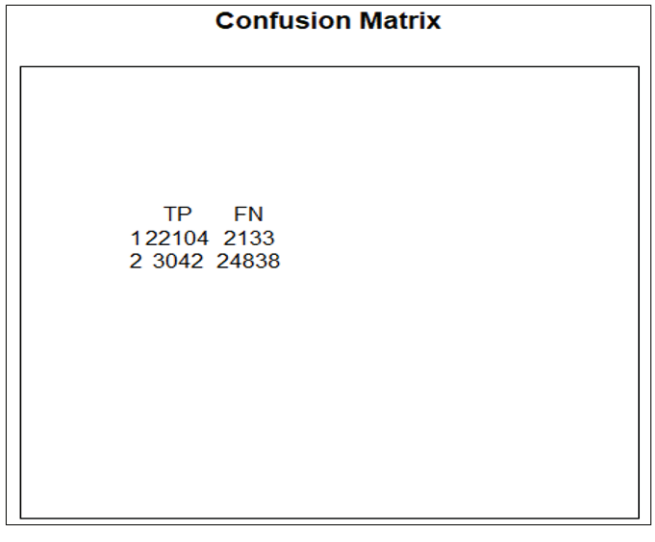


Figure 6: Confusion matrix E-learning environment

The overall accuracy shows the results of the SVM algorithm which is 90.07% with 0.879 precision, 0.911 recall, and 0.894 f-measure, and the mean of participant's data accuracy, precision, recall, and f-measure are 90.43%, 0.883, 0.897, and 0.890 respectively as shown in Table 4.

Table 4: E-learning environment results

	SUPPORT VECTOR MACHINES				
	Participants	**Accuracy**	**Precision**	**Recall**	**F-Measure**
E-LEARNING ENVIRONMENT	P1	90.25%	0.897	0.903	0.900
	P2	89.79%	0.884	0.895	0.889
	P3	91.46%	0.901	0.912	0.906
	P4	89.14%	0.881	0.894	0.887
	P5	90.94%	0.895	0.901	0.898
	P6	89.86%	0.899	0.895	0.897
	P7	89.43%	0.917	0.891	0.904
	P8	90.58%	0.874	0.891	0.882
	P9	91.89%	0.898	0.997	0.945
	P10	89.32%	0.842	0.855	0.848
	P11	90.43%	0.896	0.908	0.902
	P12	89.03%	0.789	0.881	0.832
	P13	91.56%	0.876	0.879	0.877
	P14	90.87%	0.897	0.884	0.890
	P15	91.86%	0.898	0.871	0.884
	Mean	90.43%	0.883	0.897	0.890
	SVM Model	90.07%	0.879	0.911	0.894

6. Conclusion

A Brain-Computer Interface (BCI), derived from the cognitive area of Human-Computer Interaction (HCI) is an efficient and successful emergent field [1]. The understanding is intended to explain the purpose through which the signals are produced. BCI is widely used as an integrated diagnostic tool to analyze brain signals and patterns by placing electrodes on the scalp. It gives real-time data. The EEG is used by BCI systems for monitoring and analyzing the signals of the human brain [3], [5]. Attention is an important term in educational settings. It is assessed by the cognitive mind by assisting the variety of inbound perceptual knowledge and preventing the external incentives managed by constrained cognitive minds for avoiding congestion [38], [39].

The proposed study focuses to evaluate the e-learning method and its efficiency by analyzing students' attention through EEG signals during the e-learning method and will evaluate the method if it is efficient for learning. The model was based on the following major steps to achieve the desired outcome: signal acquisition, noise (artifact) handling, pre-processing of data, and attention and cognitive load detection by feature extraction and classification. In the proposed approach, a learning model has been designed to assess attention in students, filtration of signals, and improve the accuracy of the BCI

system. For recording signals, we used the EEG technique. EEG technique is commonly used for collecting data by the noninvasive method in which the participant wears the emotive insight headset.

The data was collected, and the learning model was designed to classify attention levels. Finite Impulse Response (FIR) filter was used to remove the noises that affect signals. Further independent component analysis (ICA) was applied to extract attention features from the noise-free signals. In the end, features were analyzed with EEG signals and tests of the participants in both environments. The Labeled training data is used to prepare the learning model and a Support Vector Machine (SVM) is used for classification. The experiment is based on attention analysis in learning, for these purposes, undergraduate students are selected for the experiment that generally studies in an e-learning environment. Acquiring a signal of participants generates the dataset. For this purpose, 15 participants were selected. EEG emotive headset is used on participants for signal acquisition.

The accuracy, precision, and recall for attention levels in an e-learning environment are achieved through the SVM classifier using a linear kernel. The individual accuracy of participants was measured by breaking the dataset into training and testing through the holdout technique, while for overall accuracy for creating the SVM model; all files were merged to make a single file. The efficiency of the SVM algorithm is 90.07% with 0.879 precision, 0.911 recall, and 0.894 f-measure. The overall accuracy also confines our hypothesis of better learning outcomes in the e-learning environment.

References

[1] Gu, X., Cao, Z., Jolfaei, A., Xu, P., Wu, D., Jung, T. P., & Lin, C. T. (2021). EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. IEEE/ACM transactions on computational biology and bioinformatics.

[2] Tenório, K., Pereira, E., Remigio, S., Costa, D., Oliveira, W., Dermeval, D., ... & Marques, L. B. (2021). Brain-imaging techniques in educational technologies: A systematic literature review. Education and Information Technologies, 1-30.

[3] Wang, H., Yan, F., Xu, T., Yin, H., Chen, P., Yue, H., ... & Bezerianos, A. (2021). Brain-Controlled Wheelchair Review: From Wet Electrode to Dry Electrode, From Single Modal to Hybrid Modal, From Synchronous to Asynchronous. IEEE Access, 9, 55920-55938.

[4] Nooreldeen, H., Badawy, S., & El-Brawany, M. A. (2021). EEG Signal Analysis Based Brain-Computer. Menoufia Journal of Electronic Engineering Research, 30(2), 34-38.

[5] Stefanidis, V., Anogiannakis, G., Evangelou, A., & Poulos, M. (2015). Stable EEG Features. In Optimization, Control, and Applications in the Information Age (pp. 349–357). Springer.

[6] Brumberg, J. S., Pitt, K. M., Mantie-Kozlowski, A., & Burnison, J. D. (2018). Brain-computer interfaces for augmentative and alternative communication: A tutorial. American Journal of Speech-Language Pathology, 27(1), 1–12.

[7] Vidal, J. J. (1973). Toward Direct Brain-Computer Communication. Annual Review of Biophysics and Bioengineering, 2(1), 157–180. https://doi.org/10.1146/annurev.bb.02.060173.001105

[8] Lebedev, M. A., & Nicolelis, M. A. (2017). Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation. Physiological Reviews, 97(2), 767–837.

[9] Akyurek, E., & Afacan, O. (2013). Effects of Brain-Based Learning Approach on Students' Motivation and Attitudes Levels in Science Class. Online Submission, 3(1), 104–119.

[10] Joo, Y. J., Kim, N., & Kim, N. H. (2016). Factors predicting online university students' use of a mobile learning management system (m-LMS). Educational Technology Research and Development, 64(4), 611–630.

[11] Zhang, D., Zhao, J. L., Zhou, L., & Nunamaker Jr, J. F. (2004). Can e-learning replace classroom learning? Communications of the ACM, 47(5), 75–79.

[12] Lin, F.-R., & Kao, C.-M. (2018a). Mental effort detection using EEG data in E-learning contexts. Computers & Education, 122, 63–79. https://doi.org/10.1016/j.compedu.2018.03.020

[13] Li, Y., Li, X., Ratcliffe, M., Liu, L., Qi, Y., & Liu, Q. (2011). A real-time EEG-based BCI system for attention recognition in a ubiquitous environment. Proceedings of 2011 International Workshop on Ubiquitous Affective Awareness and Intelligent Interaction - UAAII '11, 33. https://doi.org/10.1145/2030092.2030099

[14] Nanda, P. P., Rout, A., Sahoo, R. K., & Sethi, S. (2017). Work-in-Progress: Analysis of Meditation and Attention Level of Human Brain. 2017 International Conference on Information Technology (ICIT), 46–49. https://doi.org/10.1109/ICIT.2017.53

[15] Liu, N.-H., Chiang, C.-Y., & Chu, H.-C. (2013). Recognizing the Degree of Human Attention Using EEG Signals from Mobile Sensors. Sensors, 13(8), 10273–10286. https://doi.org/10.3390/s130810273

[16] Ko, L.-W., Komarov, O., Hairston, W. D., Jung, T.-P., & Lin, C.-T. (2017). Sustained Attention in Real Classroom Settings: An EEG Study. Frontiers in Human Neuroscience, 11. https://doi.org/10.3389/fnhum.2017.00388

[17] Takeda, Y., Sato, T., Kimura, K., Komine, H., Akamatsu, M., & Sato, J. (2016). Electrophysiological evaluation of attention in drivers and passengers: Toward an understanding of drivers' attentional state in autonomous vehicles. Transportation Research Part F: Traffic Psychology and Behaviour, 42, 140–150. https://doi.org/10.1016/j.trf.2016.07.008

[18] Kuo, Y.-C., Chu, H.-C., & Tsai, M.-C. (2017). Effects of an integrated physiological signal-based attention-promoting and English listening system on students' learning performance and behavioral patterns. Computers in Human Behavior, 75, 218–227. https://doi.org/10.1016/j.chb.2017.05.017

[19] Ilgaz, H., Altun, A., & Aşkar, P. (2014). The effect of sustained attention level and contextual cueing on implicit memory performance for e-learning environments. Computers in Human Behavior, 39, 1–7. https://doi.org/10.1016/j.chb.2014.06.008

[20] Mendoza, J. S., Pody, B. C., Lee, S., Kim, M., & McDonough, I. M. (2018). The effect of cellphones on attention and learning: The influences of time, distraction, and nomophobia. Computers in Human Behavior, 86, 52–60. https://doi.org/10.1016/j.chb.2018.04.027

[21] Wang, C.-C., & Hsu, M.-C. (2014). An exploratory study using inexpensive electroencephalography (EEG) to understand flow experience in computer-based instruction. Information & Management, 51(7), 912–923. https://doi.org/10.1016/j.im.2014.05.010

[22] Lee, H, Kim, Y., & Park, C. (2018). Classification of Human Attention to Multimedia Lecture. 3.

[23] Hu, B., Li, X., Sun, S., & Ratcliffe, M. (2018). Attention Recognition in EEG-Based Affective Learning Research Using CFS+KNN Algorithm. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15(1), 38–45. https://doi.org/10.1109/TCBB.2016.2616395

[24] Shen, Y. T., Chen, X. M., Lu, P. W., & Wu, J. C. (2018). Use BCI to Generate Attention-Based Metadata for the Assessment of Effective Learning Duration. In P. Zaphiris & A. Ioannou (Eds.), Learning and Collaboration Technologies. Learning and Teaching (Vol. 10925, pp. 407–417). Springer International Publishing. https://doi.org/10.1007/978-3-319-91152-6_31

[25] Lin, Y.-T., & Chen, C.-M. (2019). Improving the effectiveness of learners' review of video lectures by using an attention-based video lecture review mechanism based on brainwave signals. Interactive Learning Environments, 27(1), 86–102. https://doi.org/10.1080/10494820.2018.1451899

[26] Liao, C.-Y., Chen, R.-C., & Tai, S.-K. (2019). EVALUATING ATTENTION LEVEL ON MOOCS LEARNING BASED ON BRAINWAVES SIGNALS ANALYSIS. 13.

[27] Mert, A., & Akan, A. (2018). Emotion recognition from EEG signals by using multivariate empirical mode decomposition. Pattern Analysis and Applications, 21(1), 81–89. https://doi.org/10.1007/s10044-016-0567-6

[28] Abdulkader, S. N., Atia, A., & Mostafa, M.-S. M. (2015). Brain-computer interfacing: Applications and challenges. Egyptian Informatics Journal, 16(2), 213–230. https://doi.org/10.1016/j.eij.2015.06.002

[29] Tang, Z., Li, C., Sun, S. (2017). Single-trial EEG classification of motor imagery using deep convolutional neural networks. Optik-International Journal for Light and Electron Optics, 130, 11–18.

[30] Urigüen, J. A., & Garcia-Zapirain, B. (2015). EEG artifact removal—State-of-the-art and guidelines. Journal of Neural Engineering, 12(3), 031001.

[31] Rogasch, N. C., Thomson, R. H., Farzan, F., Fitzgibbon, B. M., Bailey, N. W., Hernandez-Pavon, J. C., Daskalakis, Z. J., & Fitzgerald, P. B. (2014). Removing artifacts from TMS-EEG recordings using independent component analysis: Importance for assessing prefrontal and motor cortex network properties. NeuroImage, 101, 425–439. https://doi.org/10.1016/j.neuroimage.2014.07.037

[32] Pruim, R. H. R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. NeuroImage, 112, 267–277. https://doi.org/10.1016/j.neuroimage.2015.02.064

[33] Yildiz, M., Bergil, E., & Oral, C. (2017). Comparison of different classification methods for the preictal stage detection in EEG signals.

[34] Sridhar, C., Bhat, S., Acharya, U. R., Adeli, H., & Bairy, G. M. (2017). Diagnosis of attention deficit hyperactivity disorder using imaging and signal processing techniques. Computers in Biology and Medicine, 88, 93–99. https://doi.org/10.1016/j.compbiomed.2017.07.009

[35] Wessel, J. R. (2018). Testing Multiple Psychological Processes for Common Neural Mechanisms Using EEG and Independent Component Analysis. Brain Topography, 31(1), 90–100. https://doi.org/10.1007/s10548-016-0483-5

[36] Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for EEG-based brain-computer interfaces: A 10-year update. Journal of Neural Engineering, 15(3), 031005.

[37] Xu, J., & Zhong, B. (2018). Review on portable EEG technology in educational research. Computers in Human Behavior, 81, 340-349.

[38] Krauzlis, R. J., Bollimunta, A., Arcizet, F., & Wang, L. (2014). Attention is an effect, not a cause. Trends in Cognitive Sciences, 18(9), 457–464.

[39] Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J. H., & Welsh, T. N. (2019). No one knows what attention is. Attention, Perception, & Psychophysics, 81(7), 2288-2303.

Call for Papers/Authors Guideline

KIET Journal of Computing & Information Sciences (KJCIS) is biannual publication of College of Computing & Information Sciences, Karachi Institute of Economics and Technologies. It is published in January and July every year. We are lucky to have onboard prominent and scholarly academicians as part of Advisory Committee and reviewers.

KJCIS is a multi-disciplinary journal covering viewpoints/ researches / opinions relevant to the non-exhaustive list of the topics including data mining, big data, machine learning, artificial intelligence, mobile applications, computer networks, cryptography and information security, mobile and wireless communication, adhoc and body area networks, software engineering, speech and pattern recognition, evolutionary computation, semantic web and its application, data base technologies and its applications, internet of things (IoT), computer vision, distributed computing, grid and cloud computing.

The authors may submit manuscripts abiding to following rules:-

- Certify that the paper is original and is not under consideration for publication in any other journal. Please mention so in case it has been submitted elsewhere.
- Adhere to normal rules of business or research writing. Font style be 12 points and the length of the paper can vary between 3000 to 5000 words.
- Illustrations/tables or figures should be numbered consecutively in Arabic numerals and should be inserted appropriately within the text.
- The title page of the manuscript should contain the Title, the Name(s), email address and institutional affiliation, an abstract of not more than 200 words should be included. A footnote on the same sheet should give a short profile of the author(s).
- Full reference and /or websites link, should be given in accordance with the APA citation style. These will be listed as separate section at the end of the paper in bibliographic style. References should not exceed 50.
- All manuscripts would be subjected to tests of plagiarism before being peer reviewed.
- All manuscripts go through double blind peer review process.
- Electronic submission would only be accepted at kjcis@pafkiet.edu.pk
- All successful authors will be remunerated adequately.
- The Journal does not have any article processing and publication charges.

Submission is voluntary and all contributors will find a respectable acknowledgement on their opinion and effort from our team of editors. Submission of a paper will be held to imply that it contains original unpublished work. In case the paper has been forwarded for publication elsewhere, kindly apprise in time if the paper has been accepted elsewhere. Manuscripts may be submitted before September and May to get published in Jan & July issues respectively. We encourage you to submit your manuscripts at kjcis@pafkiet.du.pk

Editorial Board KJCIS
College of Computing & Information Sciences
KIET Institute of Economics and Technology

Karachi Institute of Economics and Technology
Korangi Creek, Karachi-75190, Pakistan
Tel: (9221) 3509114-7, 34532182, 34543280 Fax: (92221) 35009118
Email: kjcis@pafkiet.du.pk
http://kjcis.pafkiet.edu.pk