

KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES



ISSN: 2616-9592



Volume: 2

Issue : 1

Jan - June

2019



KIET
JOURNAL
OF COMPUTING AND
INFORMATION SCIENCES

Volume 2, Issue 1, 2019

ISSN: 2616-9592

Frequency Bi-Annual

Patron

Air Vice Marshal Tubrez Asif (Ret'd PAF)

Chief Editor

Dr. Muzaffar Mahmood

Managing Editor

Dr. Muhammad Khalid Khan



College of Computing & Information Sciences
Pakistan Air Force - 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 **Pakistan Air Force - Karachi Institute of Economics and Technology (PAF-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

All rights reserved. No part of this publication may be produced, translated or stored in a retrieval system or transmitted in any form or by any means; electronic, mechanical, photocopying and/ or otherwise the prior permission of publication authorities.

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.

Editorial Board

Dr. Mohamed Amin Embi University Kebangsaan, Malaysia	Dr. Rauf Shams Malik National University - FAST, Pakistan
Dr. Ibrahima Faye University of Technology Petronas, Malaysia	Dr. S. M. K. Raazi Muhammad Ali Jinnah University, Pakistan
Dr. Anh Nguyen-Duc Norwegian University of Technology, Norway	Dr. Shaukat Wasi Muhammad Ali Jinnah University, Pakistan
Dr. Mohd Fadzil Bin Hassan University of Technology Petronas, Malaysia	Dr. Imran Khan Institute of Business Administration, Pakistan
Dr. Tahir Riaz Data Architect, Sleeknote ApS, Denmark	Dr. Tariq Mahmood Institute of Business Administration, Pakistan
Dr. Ronald Jabangwe University of Southern Denmark	Dr. Aqeel-ur-Rehman Hamdard University, Pakistan
Dr. Syed Irfan Hyder Institute of Business Management, Pakistan	Dr. Muhammad Saeed UBIT, Karachi University, Pakistan
Dr. Bhawani S Chowdry Mehran University, Jamshoro, Pakistan	Dr. Muhammad Aamir Sir Syed University of Engineering & Tech, Pakistan
Dr. Hussain Pervez PAF-KIET, Pakistan	Dr. Kamran Nishat PAF-KIET, Pakistan
Dr. Sameer Qazi PAF-KIET, Pakistan	Dr. Khurram Junejo PAF-KIET, Pakistan
Dr. Arsalan Jawed PAF-KIET, Pakistan	Dr. Taha Jilani PAF-KIET, Pakistan
Dr. Ayaz Khan PAF-KIET, Pakistan	Dr. Maaz Bin Ahmed PAF-KIET, Pakistan

Language Editors

Mr. Ayub Latif, Mr. Sohail Imran, Ms. Solat Jabeen, Ms. Huma Tariq
CoCIS, PAF-KIET, Pakistan

Production Management

Mr. Muhammad Furqan Abbasi
CoCIS, PAF-KIET, Pakistan

Table of Content

1 1- 11	A Study of SAR Despeckling Methods <i>Muhammad Haris, Mahmood Ashraf, Faraz Ahsan, Ahmed Athar, Memoona Malik</i>
2 12-25	<i>Achieving High Availability in Cloud through Live Migration</i> <i>Abdul Wahab Khan, Shaikh Adnan Ahmed Usmani</i>
3 26-38	Detection of Myocardial Infarction in ECG Base Leads using Deep Convolutional Neural Networks <i>Awais M. Lodhi, Adnan N. Qureshi, Usman Sharif, Zahid Ashiq</i>
4 39-48	Mining Diagnostic Investigation Process <i>Sohail Imran, Dr. Tariq Mahmood</i>
5 49-60	Ontology Based System for Expert Searching in Academia using SWRL and SPARQL <i>Furqan Hussain Essani, Quratulain Rajput</i>
6 61-75	Optical Character Recognition Engine to extract Food-items and Prices from Grocery Receipt Images via Templating and Dictionary-Traversal Technique <i>Ali Sohani, Rafi Ullah, Faraz Ali, Athaul Rai, Richard Messier</i>
7 76-86	The Role of SEO Techniques to Enhanced Performance and Improved Rankingfor Intelli-Web Shop <i>Muhammad Noman khalid, Hira beenish, Muhammad Iqbal, Kamran Rasheed, Muhammad Talha</i>

A Study of SAR Despeckling Methods

Muhammad Haris¹ Mahmood Ashraf² Faraz Ahsan³ Ahmed Athar⁴ Memoona Malik⁵

Abstract

To capture geographic images, optical sensors were used in past to analyze and extract information for numerous purposes ranging from mineral resources to surveillance. But the limited availability of optical sensors use led to radar based sensors for continuous data capturing. In which, neither time constraint was involved nor the hinderance by atmospheric factors. The image captured with the synthetic aperture radars undergo the phenomena of speckle which is a granular noise triggered by the constructive or destructive supersition of received signals. The result of this interenference affects the image quality and the retrivel of the information. To cope with this issue, numerous strategies were defined from local filters to non-local filters, transforms and combined solution through transforms and non-local filters as hybrid techniques. The studies of local, non-local and transforms were efficient but each having their limitation posed the question for improvent in denoising the images. The experiments conducted showed that hybrid solutions outperform their predecessors for which this study is conducted to analyze and evalute the techniques and explore the factors in hybrid techniques that can efficiently despeckle the synthetic aperture radar images by maintaing the image quality and details.

Keyword: Synthetic Aperture Radar (SAR), Optical Sensors, Radars, Speckle, Noise, Local Filters, Non-Local Filters, Wavelets, Transforms

1 Introduction

Optical sensors image capturing is a passive phenomenon that is dependent on the visible spectrum of sunlight which is captured by the sensors. The SAR imaging synthesizes the working of actual radar and is an active image capturing process in which signals are transmitted by the carrier and received by the same radar. The distance carrier flies between transmission and receiving of signals is the reason its synthetic aperture radar else the hundreds of meter long radar have to be carried which would make the process ineffective and costly. The figure 1 explains the working of active and passive image acquisition processes. In the upper portion of the image, optical sensor captures the reflected sunrays from objects and transfers it to the ground station. The reason of it being obsolete is the visible spectrum constraint to be used only in the day time and shorter wavelength limits the application for only the objects above the ground surface.

^{1,4,5} COMSATS Institute of Science & Technology, Islamabad, Pakistan

² Federal Urdu University of Arts, Sciences & Technology Islamabad, Pakistan

³ University of Islamabad, Institute of Engineering and Sciences

¹ m.haris@comsats.edu.pk

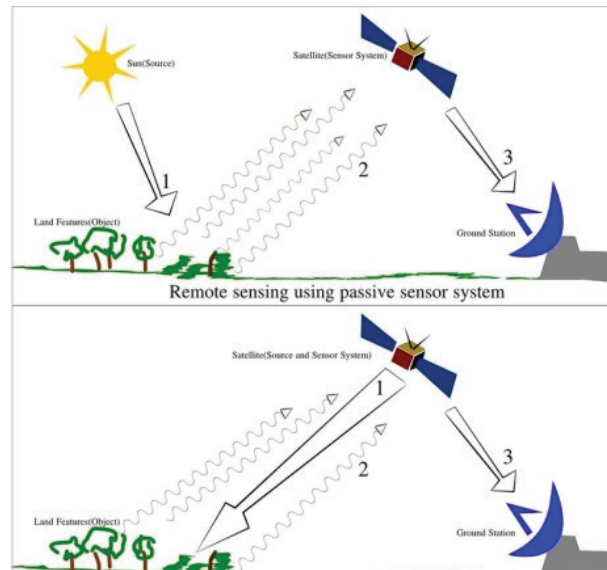


Figure 1: Remote sensing using optical and radar

Whereas the active sensing techniques self illuminates the area of interest in a side looking direction flew by the radar carrier and covering the distance to synthesize the diameter of radar to capture the reflected signals from the object [1-3]. Active sensing supersedes the passive by three areas that are all day and night image acquisition because of microwave signal instead of visible spectrum which having higher wavelength strengthens the image acquisition process to capture high quality images and not restricted to the ground objects. The higher wavelength of microwave signals can penetrate the earth surface enabling it for exploring the underlying structure and resources. The microwave signals are also not hindered by the atmospheric factors like cloud, rain etc. The synthesizing of radar in active image capturing the images encounter speckle that degrades image quality and obstructs the extraction of information [7-8]. The table 1 presents a brief comparison among the two remote sensing techniques.

Table 1: Comparison of Optical and Radar based Image Acquisition.

	Optical Sensors	Radar Based
1.	Passive Image	Active Imaging
2.	Use of Visible spectrum	Use of Microwaves
3.	Non penetrable	Can penetrate the Earth Surface
4.	Works only in Daylight	Can Work both in Day and Night
5.	Effected by atmosphere	Not Effected by atmosphere
6.	Not efficient results	Efficient results

2 Material And Methods

Radar airborne carrier satellite emits the microwave signals towards the earth surface in a side looking direction and travels the distance to capture the transmitted signal. The transmitted microwave when received by the radar encounter reflection and scattering of the signals. When

the incidence angle of the signal is same as reflection angle it is called reflection and scattering is when the signals deviate from their course due to the object. These two factors together cause disturbance in the received signals either constructive or destructive interposition known as speckle. The effect of speckle is difficulty in understanding the images for information extraction and features of images. To overcome the speckle issue in the radar images, various solutions have been applied that are filtering the images through local, non-local filter, transforms and combination of transforms with non- local [9-10]. In the early 1980s, Lee filter was proposed by Jong Sen Lee that served as the foundation in local filters for despeckling techniques where the regions were uniform and performed well, but the regions with non-uniform structure Lee filter smoothed the structures of region that altered the image. Lee compelled the filtering by utilizing mean and neighborhood change of pixels keeping up edge sharpness and refined points of interest. That favored its application in exploiting the image insights [11]. Kuan filter was the next advancement in local filters that did not take any approximation while denoising the images and produced better results than the Lee filter. In Kuan filter a weighted function was used for which an additive noise model was generated for the existing multiplicative noise. [12]. Frost filter was proposed in local filters category with the objective to limit the mean square between the captured and denoised synthetic aperture radar image [13]. To overcome the limitations of local filters non-local filters were introduced with the principle to take the information an image provides that is pixels similarity and prediction before filtering the captured image. To overcome the limitations of local filters, non-local filters were introduced with the principle to take the information an image provide that is pixels similarity and prediction before filtering the captured image. The non-local filters executed to perform better results than the local filters with the requirement of more powerful computation machines. With this principle, Buades presented a non-local solution to denoise based on the non-local average algorithm. The non-local mean algorithm by Buades is weighted mean function of neighboring pixels in place of average, after which weighted average of pixel similarity is calculated. To extend the work of Buades, Charles Alban et al presented an enhancement to the means with nonlocal as noise distribution model. In Charles model probabilistic patch filtering is done for which weighted maximum likelihood estimation of synthetic aperture radar image calculated and weights are extracted in a data-driven way. Probabilistic model runs iteratively solves the problem by computing the similarity among the calculated noised patches and patch from last iteration to enhance the weights of the function at each iteration. Iteratively solving and updating the weights results in better performance. The same process can be used for Gaussian and additive noised images. The iterative process on hand increases the performance but suppresses the thick and dark details in a regularized image. [14-15]

While the efforts being done to denoise, the images having Gaussian noise was successful but image details were not preserved. For which Zhong proposed non-local Bayesian filter to denoise and preserve image details. The work done by Zhong was based on Lee sigma filter an extension of earlier Lee local filter and Charles Alban work of the patch probability phenomena [16]. Bayesian addressed speckle elimination by the use of improved sigma filter [17] for preservation of image details with moving window having size from 5×5 , 7×7 , 9×9 , 11×11 and above; along with scanning rectangular windows. The advantage of doing so provides filtering images with the azimuth pixel spacing higher than the range

pixel spacing. In general, higher speckle reduction is achieved using a larger window, but computational load will also be increased. This overcomes the bias in the estimates, unfiltered black pixels, and smearing of strong targets and addition of MMSE estimator for the adaptively filter the image [18]. Another contribution to non-local filters is by Bin Xu which uses the nonlocal sparse model by iterative filtering mechanism for synthetic aperture radar images. The captured image is at first transformed to logarithmic image domain. The transformed images are then denoised by nonlocal sparse model and iteratively regularizing. At each turn of the iteration, noise factor is updated and variance is calculated. For a patch in noised region similar to those are found, grouped and passed through sparse coding. The filtered patches are combined to obtain the image without noise. Iterative regularization and non-sparse model, result into higher computational complexity of model in comparison to the techniques explained before [19].

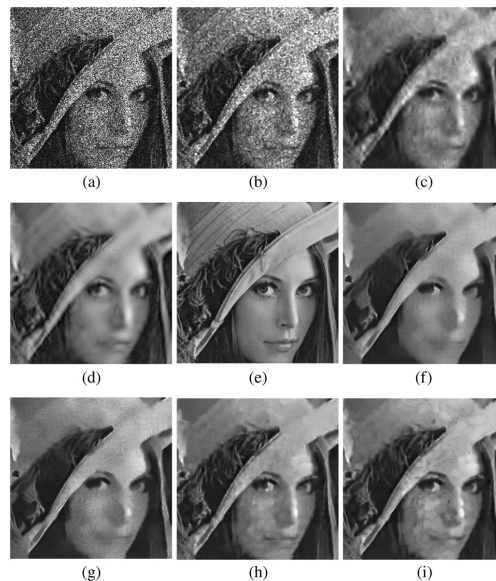


Figure 2: Zoom of filtered images for Lena corrupted by one-look speckle. (a) Noisy image. (b) Frost. (c) SA-WBMAE. (d) MAP-S. (e) Original image. (f) PPB. (g) H-PPB. (h) SAR-BM3-D. (i) H-BM3-D.

To de-speckle the polari metric and interferometry images, NL-SAR framework is proposed as an extension of Non-local filter for the images that suffer loss of resolution and feature loss. To mitigate these effects, pixels of the captured image are to be analyzed locally with estimation. NL-SAR does so by the weighting function on the similarities among patches. Thus, offering flexible mechanism for preservation of resolution and fine details. It builds non-local neighborhood for the PL, IN images defined on the similarity of pixel among the patches. After performing several estimations, non-locally best local patch is selected to form an image. By doing so, fluctuations and abnormalities can be handled but there exists room for improvement in resolution preservation estimation algorithms with reference to advancement in radar technology [20]. A transform acts as a function, takes an image as input and produces output image with varying characteristics as per function. Examples of image transforms are wavelet,

principle component analysis etc. Fabrizio Argenti in 2002 approached with translation by undecimated frequency wavelet breakdown for de-speckling of multiplicative noise. In which extended form of mean minimum square error has been applied on multiplicative noise with use of undecimated frequency wavelet. Advantage is of single independent noise by identifying the fact that undecimated frequency wavelet coefficient despeckling is equal to denoising of translation invariant [21]. In addition to the earlier work done in 2002, an enhanced mechanism was proposed including MAP (Maximum A Posteriori) criterion in application with un-decimated wavelet for de-noising of SAR images. The novelty of the work includes exact expression for estimation of Generalized Gaussian distribution without the need for any further assumptions and classification model of wavelet coefficients based on texture energy collected. Benefit of extended algorithm is smoothing of background, preservation of texture, refinement of parameters for estimation irrespective of underlying reflectivity [22].

For the preservation of synthetic aperture radar image edges linlin proposed bandelet transform to detect edge direction along with fuzzy clustering in TIBT (Translation invariant bandelet transform). By the use of Canny operator, edges are detected and removed. Afterwards, combined algorithm of fuzzy clustering and TIBT is applied to de-speckle the image with edge removed. At the end, edges are added back to the denoised image. Results of the experiment depict better visual quality and evaluation indexes outperforms the other methods with no edge preservation [23]. For edges refinement and detail preservation, zhang proposed despeckling algorithm using transforms for SAR images in which curvelet transform and particle swarm optimization techniques have been used that reduce speckle and refined edges details of which in figure 2 lena was subject to test. To non-linearly shrink and stretch the coefficients of curvelet a function for improved gain is put in place that combines speckle reduction with feature enhancement. Benchmark is been set to acquire optimal parameters in gain function. For searching among the finest de-speckle, SAR image enhanced PSO with better learning scheme and mutation operator were introduced. Experiments show that the proposed mechanism outperforms other transforms techniques such as bandelet, curvelet bandelet non adaptive SAR images de-speckling. The algorithm as a whole has high computational cost because of PSO iterative nature which can be reduced if PSO can run in parallel [24]. The filters were useful for noise reduction but were lacking to preserve the relevant features of image that is the textual information. In parallel to filtering techniques, transforms provided edge refinement. So, a lot more can be achieved by combing the two solutions for proposing a new hybrid solution. Sara Parrilli et.al in 2012 proposed a despeckling technique based on non-local filter wavelet transform that forms a basis for todays' hybrid solutions to despeckling. The proposed performed better in detail preservation and signal to noise ratio by doing the block matching for additive white gaussian noise with iteratively improving the estimator for minimum mean square error. Results of which are shown is figure 3 and 4. Limitations of the proposed solution are ; in actual SAR images with high resolution speckle statistics deviate at time from the solution and quality of SAR images degrade the output for which indicators and ad hoc simulated SAR images solution is been identified for future work [25].

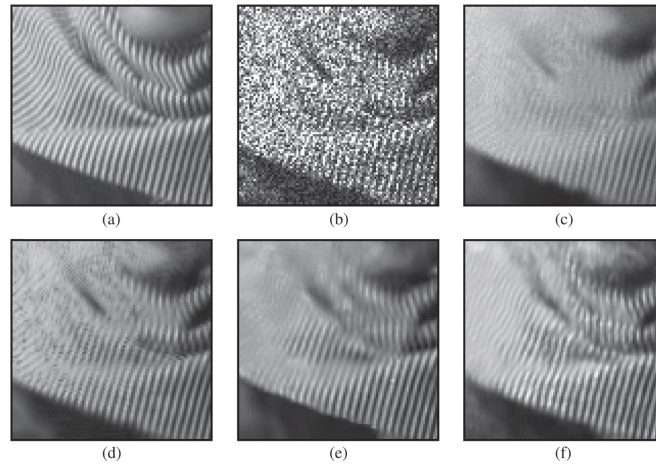


Figure 3: Barbara image zoomed and degraded by single-look speckle noise. (a) Clean image. (b) Noisy image. (c) PPB. (d) LPG-PCA. (e) SAR- BM3D. (f) Clustering based PCA

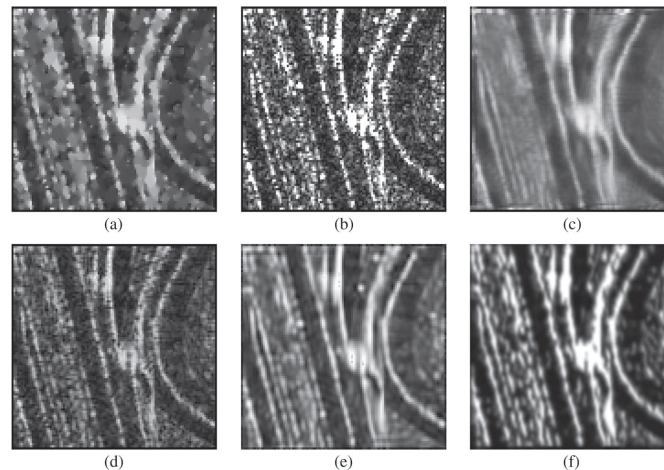


Figure 4: TerraSAR-X SSC image zoomed (126×116 pixels) of roads located at the SE of the Macdonald-Cartier Freeway/Allen Road interchange, Toronto, ON, Canada, with $L = 1$. (a) SRAD. (b) Original image. (c) PPB. (d) LPG-PCA. (e) SAR-BM3D. (f) Clustering based PCA

Following the work of non-local filters by combining with batch matching algorithm shortcomings are addressed by a fast adaptive non-local despeckling method in 2014 by Davide Cozzolino. Adding to the batch matching 3-d model time critical images are analyzed by an area of variable search size and exploiting the statistics of noise by probabilistic termination to lower computation time. After initial filtering of speckle computation, cost is further reduced by look up tables. Doing so increases performance with lower complexity [26]. With the combination of non-local filters and transforms, better results are obtained in terms of image resolution and detail preservation. The literature of despeckling SAR images is directed towards finding solution in which de-speckling details are preserved. Number of attempts have been put to apply filtering to the feature of image or it can be said more optimistically with the sole reason for

pixel classification in advance when diverse algorithms are applied such as in [30] area under consideration is put to heterogeneity test on variation coefficient. In transforms solutions, same can be adopted when wavelets are used with an appropriate of the textured energy for an AD-Hoc maximum a posteriori [31]. Lately non-local filters gained popularity but possessed same drawbacks of their predecessor that one offering detail preservation and other denoising of the image but each compromising on the other for which alternate adaptive mechanisms have been proposed to overcome the shortcoming [32]. Such as in [16], local selection is been adopted for selection of best parameter is incorporated for denoising. [26] is based on local image estimate in which controlling parameters are adjusted to match the image area under study. To cope with the difference among phase combination to address heterogeneous and homogenous areas stacked processing is done for improved classification [33]. At very small ratio of signal to noise in illustrating the captured image numerous estimators are used to lower the variance factor of the estimate. Lilin clustered the SAR image into disjoint local regions for de-speckling and each region is denoised by applying LMMSE and then the principal component analysis transform. K-mean clustering algorithm is applied by identifying principle components on length criteria. The filtered clusters are joined to form a denoised image [27].

Diego proposed that by doing the pixel wise image classification better results can be achieved that would allow to select an estimator based on the combined characteristics and will be more appropriate. SAR image is divided into regions having homogeneous and geometrical properties for clustering. Several alternate estimates are calculated of the same data and by use of soft classification and de-speckling tools, experiments yield improved results when performed on real-world high-resolution SAR images[28]. Denoising of the SAR images is challenging due to induced phenomena of speckle in which what earlier techniques did not considered was the physical phenomena of electromagnetic scattering in the images that degrades the image quality. Gerardo study of block matching technique for SAR images considers the scattering factor and proposed an extended solution to the original SARBM-3D synthetic aperture radar batch matching with SBSARBM-3D Scattering Based SARBM-3D for speckle reduction without compromising on the details and also reduces the presence of arbitrary effects in homogenous areas of image. The solution leaves room for improvement in the areas of urban and rural environment of peculiar scattering [29]. In a recent study of hybrid de-speckling techniques Alessio et.al proposes an algorithm that requires knowledge about local geography. The solution considers different features of image for de-speckling which are surface description by scattering model, error part acquired by local incidence angle of surface parameters, digital evaluation model and information about errors when digital evaluation model is applied. The need of prior information can be substituted by calculating the local incidence angle directly from SAR data [30-34].

3 Discussion

The goal of this study was to study Hybrid de-speckling techniques for SAR images and explore hybrid solution to improve the results. For which local, non-local, transforms and non-local with transforms are studied and results recorded. Drawback associated with these approaches is introduction of new artifacts to the denoised image. The despeckling techniques have been classified in Table 2 based on the category of solution.

Table 2: Classification and Features of despeckling techniques

	Despeckling Technique	Category
1.	“Sensitivity analysis of a scattering-based ...”	Non-Local with Transform
2.	“Scattering-Based SARBM3D ...”	Non-Local with Transform
3.	“An iterative SAR image filtering method...”	Non-Local Filters
4.	“SAR despeckling based on soft classification ...”	Non-Local with Transform
5.	“Fast adaptive nonlocal SAR despeckling ...”	Non-Local with Transform
6.	“SAR image denoising via clustering ...”	Non-Local with Transform
7.	“ A nonlocal SAR image...on LLMMSE...”	Non-Local with Transform
8.	“SAR image despeckling using Bayesian...”	Non-Local Filters
9.	“An adaptive method of speckle reduction...”	Transforms
10.	“SAR image despeckling using edge ...”	Transforms
11.	“Iterative weighted maximum likelihood denoising...”	Non-Local Filters
12.	“Segmentation-based MAP despeckling of SAR ...”	Transforms
13.	“A non-local algorithm for image denoising ...”	Non-Local Filters
14.	“Speckle removal from SAR images ...”	Transforms
15.	“NL-SAR: A unified nonlocal framework . . .”	Non-Local Filter

Non-local filtering approaches in transformed domain outperform local and non-local filters. Their advantages include better detail preservation while speckle suppression. Whereas these approaches are cost inefficient, responsible for inducing some new artifacts to the image being denoised like non-local filters and used lossy transforms which results in information loss. Non-local, Transforms and combined solution has been applied on standard images acquired from the image database of University of Southern California for each image peak signal to noise ratio is calculated and recorded in table 3. The results indicate the performance of the despeckling techniques improves when hybrid solution is applied for denoising.

Table 3: PSNR after applying algorithms on standard images

Image/algos	DCT	PCA	UDWT	SARBM3D
Lena	2.5289	5.8873	5.9175	13.137
Baboon	1.8686	5.3426	5.372	19.4795
House	2.9273	6.3662	6.3984	20.26941
Bridge	2.2273	5.6566	5.687	20.2694
Woman	1.2609	4.1324	4.0921	20.25105

4 Conclusions And Future Work

Remote Image acquisition is an expensive process by both human resources and financially that requires a lot of scientific minds to plan, build and put the carrier in space along with

the mechanism to acquire the images for various purposes in which a country invests its capital to benefit. When these expensive systems are put in place, results which convey precise information about the phenomena are expected so that efforts are justified. Satellites have been successfully revolving around our planet for a long time for information transmission, in past optical sensors were used but due to their limitation of working only in visible spectrum made them obsolete for exploring about phenomena beneath earths' surface and put a constraint on their availability duration. Recent advances in remote sensing systems have introduced a more efficient solution in which Radar functionality can be synthesized and better results can be obtained without the hindrance of the atmospheric factors with all day and night image acquisition. Now by synthesizing antenna having hundreds of long lengths, the active images acquired because of either constructive or destructive combination of received signals speckle occurs. De-speckling, solutions have been provided since 19th century but the ongoing research focuses on the hybrid techniques of those already used to further improve the results. This study focused on the various despeckling solutions presented to cope with the issue and classified the solutions and motivated towards the search for improving the hybrid solution available.

References

- [1] Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I. and Papathanassiou, K.P., 2013. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1), pp.6-43. A.Gupta "Selected Topics in Image Processing", *Multiplicative Noise Removal Techniques – A Study*,2015
- [2] Mamatha, K.R., Hariprasad, S.A., Saranya, P., Shahi, S., Poddar, S. and Srivastava, V., 2016, May. A non local approach to de-noise SAR images using compressive sensing method. In *Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, IEEE International Conference on (pp. 1603-1606). IEEE.
- [3] Panda, S. and Singh, P., Study of Image De-noising Techniques for Facilitating the Process Selection to Determine the Best Suitable Approach for any given image Type. *International Journal of Engineering and Innovative Technology IJEIT*,vol. 4, no.4, 2014
- [4] Ramponi, G., D'Alvise, R. and Moloney, C., 1999, June. Automatic estimation of the noise variance in SAR images for use in speckle filtering. In *NSIP* (pp. 835-838).
- [5] Touzi, R., 2002. A review of speckle filtering in the context of estimation theory. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11), pp.2392-2404.
- [6] Deledalle, C.A., Denis, L., Poggi, G., Tupin, F. and Verdoliva, L., 2014. Exploiting patch similarity for SAR image processing: the nonlocal paradigm. *IEEE Signal Processing Magazine*, 31(4), pp.69-78.
- [7] Gao, G., 2010. Statistical modeling of SAR images: A survey. *Sensors*, 10(1), pp.775-795.
- [8] Loris, I., 2012. *Wavelets: A Concise Guide*.
- [9] Argenti, F., Lapini, A., Bianchi, T. and Alparone, L., 2013. A tutorial on speckle reduction in synthetic aperture radar images. *IEEE Geoscience and remote sensing magazine*, 1(3), pp.6-35.

- [10] Lee, J.S., 1981. Speckle analysis and smoothing of synthetic aperture radar images. *Computer graphics and image processing*, 17(1), pp.24-32.
- [11] Kuan, D.A.R.W.I.N.T., Sawchuk, A.L.E.X.A.N.D.E.R.A., Strand, T.I.M.O.T.H.Y.C. and Chavel, P., 1987. Adaptive restoration of images with speckle. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3), pp.373-383.
- [12] Frost, V.S., Stiles, J.A., Shanmugan, K.S. and Holtzman, J.C., 1982. A model for radar images and its application to adaptive digital filtering of multiplicative noise. *IEEE Transactions on pattern analysis and machine intelligence*, (2), pp.157-166.
- [13] Buades, A., Coll, B. and Morel, J.M., 2005, June. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 2, pp. 60-65)*. IEEE.
- [14] Deledalle, C.A., Denis, L. and Tupin, F., 2009. Iterative weighted maximum likelihood denoising with probabilistic patch-based weights. *IEEE Transactions on Image Processing*, 18(12), pp.2661-2672
- [15] Lee, J.S., Wen, J.H., Ainsworth, T.L., Chen, K.S. and Chen, A.J., 2009. Improved sigma filter for speckle filtering of SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1), pp.202-213.
- [16] Lee, J.S., 1983. Digital image smoothing and the sigma filter. *Computer vision, graphics, and image processing*, 24(2), pp.255-269.
- [17] Zhong, H., Li, Y. and Jiao, L.C., 2011. SAR image despeckling using Bayesian nonlocal means filter with sigma preselection. *IEEE Geoscience and Remote Sensing Letters*, 8(4), pp.809-813.
- [18] Xu, B., Cui, Y., Li, Z. and Yang, J., 2015. An iterative SAR image filtering method using nonlocal sparse model. *IEEE geoscience and remote sensing letters*, 12(8), pp.1635-1639.
- [19] Deledalle, C.A., Denis, L., Tupin, F., Reigber, A. and Jäger, M., 2015. NL-SAR: A unified nonlocal framework for resolution-preserving (Pol)(In) SAR denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4), pp.2021-2038.
- [20] Argenti, F. and Alparone, L., 2002. Speckle removal from SAR images in the undecimated wavelet domain. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11), pp.2363-2374.
- [21] Bianchi, T., Argenti, F. and Alparone, L., 2008. Segmentation-based MAP despeckling of SAR images in the undecimated wavelet domain. *IEEE Transactions on Geoscience and Remote Sensing*, 46(9), pp.2728-2742.
- [22] Zhang, W., Liu, F., Jiao, L., Hou, B., Wang, S. and Shang, R., 2010. SAR image despeckling using edge detection and feature clustering in bandelet domain. *IEEE Geoscience and Remote Sensing Letters*, 7(1), pp.131-135.
- [23] Li, Y., Gong, H., Feng, D. and Zhang, Y., 2011. An adaptive method of speckle reduction and

- feature enhancement for SAR images based on curvelet transform and particle swarm optimization. *IEEE Transactions on Geoscience and Remote Sensing*, 49(8), pp.3105-3116.
- [24] Parrilli, S., Poderico, M., Angelino, C.V. and Verdoliva, L., 2012. A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2), pp.606-616.
- [25] Cozzolino, D., Parrilli, S., Scarpa, G., Poggi, G. and Verdoliva, L., 2014. Fast adaptive nonlocal SAR despeckling. *IEEE Geoscience and Remote Sensing Letters*, 11(2), pp.524-528.
- [26] Xu, L., Li, J., Shu, Y. and Peng, J., 2014. SAR image denoising via clustering-based principal component analysis. *IEEE transactions on geoscience and remote sensing*, 52(11), pp.6858-6869.
- [27] Gragnaniello, D., Poggi, G., Scarpa, G. and Verdoliva, L., 2015, July. SAR despeckling based on soft classification. In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International* (pp. 2378-2381). IEEE.
- [28] Di Martino, G., Di Simone, A., Iodice, A., Poggi, G., Riccio, D. and Verdoliva, L., 2016. Scattering-Based SARBM3D. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6), pp.2131-2144.
- [29] Di Simone, A., Di Martino, G., Iodice, A. and Riccio, D., 2017. Sensitivity analysis of a scattering-based nonlocal means Despeckling Algorithm. *European Journal of Remote Sensing*, 50(1), pp.87-97.
- [30] E. Nezry, A. Lopes, and R. Touzi, "Detection of structural and textural features for SAR images filtering," in *IEEE IGARSS, 1992*, pp. 2169-2172.
- [31] M.Malik, M.Haris, A.H.Dar, A.A.Safi, M.Ashraf, "A Review of SAR Hybrid De-Speckling Methods". in *Springer Communications in Computer and Information Science series. Book, Intelligent Computing Systems, Volume 820*, pp-136-146
- [32] T. Bianchi, F. Argenti, and L. Alparone, "Segmentation Based MAP Despeckling of SAR Images in the Undecimated Wavelet Domain," *IEEE Trans. Geosci. and Remote Sensing*, vol. 46, no. 9, pp. 2728-2742, 2008
- [33] G. Di Martino, M. Poderico, G. Poggi, D. Riccio, and L. Verdoliva, "Benchmarking framework for SAR despeckling," *IEEE Trans. Geosci. and Remote Sensing*, vol. 52, no. 3, pp.1596-1615, 2014
- [34] D. Gragnaniello, G. Poggi, and L. Verdoliva, "Classification based nonlocal SAR despeckling," in *Tyrrhenian Workshop on Advances in Radar and Remote Sensing, 2012*, pp. 121-125.

Achieving High Availability in Cloud through Live Migration

Abdul Wahab Khan¹

Shaikh Adnan Ahmed Usmani²

Abstract

Increasingly relying on the cloud for deployment and assessing critical applications and services for businesses makes its high availability as an extremely critical aspect. The paper evaluates virtualization based systems and techniques for the betterment of the overall resilience of a cloud environment. We have highlighted systems to perform monitoring, load balancing and dynamic allocation of resources, replication and live migration at backup sites and a number of pioneering approaches such as ghost VMs and Byzantine fault tolerance to ensure high availability. Moreover, hurdles and bottlenecks for the effectiveness and application of these systems are also identified. A real-world implementation of live migration is also presented with a concise discussion of the challenges faced during the setup and configuration phases.

Keywords: High Availability, Virtualization, Cloud Computing, Live Migration, Load Balancing

1 Introduction

Cloud computing, an emerging paradigm and committed to flourish accuracy & well-grounded delivery of computing and storage services is becoming an eminent choice for the enterprises. Over the years, organizations have shifted their core & critical business applications to the cloud and numerous users are becoming depended on online services. Considering modern scenario and increasingly growing traffic, it is crucial to maintain high availability of cloud services. The research to follow discusses the virtualization based techniques used to ensure high availability in the cloud and the limitations and problems associated with those.

Cloud Computing has flickered a mammoth amount of interest in the technology community. The on-demand nature of this paradigm makes it a perfect fit for today's deft business environment. With increasingly computing cost, cloud computing provides one excellent solution and a more obvious choice. Through improved consumption of resources and condensed administration and management costs, it offers substantial cost savings to enterprises. The entry of major industry players such as Microsoft, Google, and Amazon has accelerated the cloud's industry adoption rate as they introduce innovative features and drive costs down.

Virtualization is the core enabling technology behind this pattern. Multiplexing resources among applications and customers provide the elasticity to add new resources quickly and dynamically to a customer's resource pool. Greater efficiency is achieved by improving resource usages and noteworthy cost savings are realized by combining multiple servers into a single machine with numerous virtual machine instances.

Hypervisors, one of the most popular products of the research in virtualization technology and also called virtual machine managers, are programs that run directly on the bare-metal

¹PAF-Karachi Institute of Economics & Technology, Karachi, Pakistan / abdul.khan@pafkiet.edu.pk

²University of Karachi, Pakistan / adnan@szic.edu.pk

hardware allowing multiple operating systems to share a single physical machine. Hypervisors also cope up the resources for the guest operating systems and ensure that they run in complete isolation from each other. When coupled with management tools and application programming interface, a hypervisor is an inclusive cloud platform that makes it convenient for users to setup and manage their own cloud.

Some of the renowned cloud platforms available today include commercial offerings such as Xen Cloud Platform [1] and VMWare ESX [2] as well as open source systems such as Eucalyptus [3] that are targeted for academic use. These platforms allow sharing of a single physical server by multiple virtual machines, failure in one node can result in disturbance of service to multiple applications and clients. Therefore, fiascos in the cloud have to be handled transparently to safeguard the cloud and its availability at all times.

Both Xen and VMWare have built-in mechanisms to warrant high availability. XenCenter includes a set of tools called Xen Essentials that comprises of high availability components. These include Xen Motion for performing live migrations of VMs and the Workload Balancer. VMWare HA (High Availability) is a utility that provides upbeat monitoring of servers and virtual machines, automatic detection of failures and swift restart and optimal placement of VMs after server failures [5]. Availability is ensured through dynamic adjustment of resource allocations, rapid restart or migration of VMs between hosts. Eucalyptus lacks in terms of high availability as its focus is further towards learning and research rather than usage for real-time applications.

2 Motivation

By 2012, 20% of the businesses will own no IT assets. The paradigm from in-house data centres to the cloud will result in businesses depending on the cloud for their precarious applications and services. Availability will be the key focus and concern for the organizations making this transition.

Companies are trusting on the cloud to provide diversified services ranging from photo sharing, online storage and social networking to financial services for enterprise customers. These companies include both start-ups as well as large corporations with thousands of customers. Therefore any outage in the cloud will affect thousands of customers and impact would be magnificent. For many, the consequences will be devastating and might be realized in the form of financial losses, customer dissatisfaction, negative publicity and a number of other ways.

In order to safeguard service availability, generally customers sign Service Level Agreements (SLAs) with the cloud providers. Availability is one of the most critical elements pursued by the customers when selecting a service provider. The chief industry players are capitalizing in research and energies to develop new and upgraded techniques for growing availability of cloud services.

Despite numerous systems and techniques developed, many are still new and immature which in fact serve the paramount basis of this research. We analysed the various techniques

and methods for ensuring reliable delivery of services in the cloud and tried to identify the matters and restrictions. As part of this research, we have also setup a small cloud environment to get a hands-on experience of live migration and the other technologies involved.

3 Research Contribution

Our research focuses on analysing the various monitoring, fault-tolerance and load balancing techniques used in practice to achieve high availability in the cloud with a special emphasis on Live Migration. An attempt is made to identify the gap in the research by highlighting the limitations of each of these techniques.

Supporting work includes the creation of a test cloud using the Xen Cloud Platform and the development of a management console. Via this console, users can manage the servers and virtual machines in the cloud. It also provides the ability to handle excess load by creating new instances of running VMs, dynamically change resource allocations and perform live migrations.

4 Literature Review

A Ganglia

- Just a Monitoring System. It provides scalable monitoring for distributed High Performance Computing (HPC) systems [10].
- Scalable, that is nodes can be added easily in a cluster without manual configuration.
- Every node of a cluster share heartbeat messages to each other using listen/announce protocol.
- Monitoring sys has 2 daemons.
- gmond daemon: Runs on every node of a cluster.
- gmetad daemon: Runs on each cluster and then data aggregation of each cluster.
- Uses XML for data representation. Support both built-in and application specific metrics.

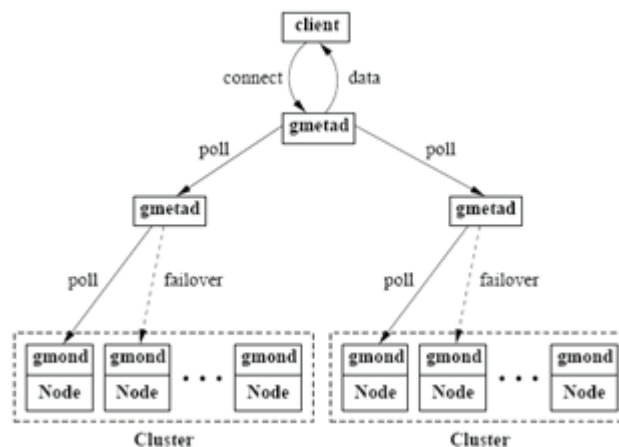


Figure1: Ganglia architecture

Figure 1 shows the architecture of Ganglia [10]. To federate multiple clusters, a tree of point to point connections is used. As each leaf node maintains monitoring information of its entire

cluster, it logically represents a distinct cluster whereas each non-leaf node represents a set of clusters. At each point in the tree, child nodes are polled at regular intervals to aggregate data. Multiple nodes in a cluster are specified to handle failures for a leaf node.

Ganglia has a low per node overhead. This is chiefly due to new nodes discovery without any manual configuration thus eliminating noticeable management overhead. Through its multicast approach, scalability is automatically addressed. It provides support for both built-in and application specific metrics. Usage of XML for data representation and XDR for data transport enhances its extensibility, allowing integration with other information services.

Limitations. Although, Ganglia has been deployed in a number of real world environments and has proven its effectiveness in most cases, a few issues and constraints may limit its widespread adoption.

- Excessive network traffic for large of nodes.
- Can't run on WAN if WAN doesn't support multicast.
- Cost increases as nodes increases.
- Planet Lab tried implementation but 19.15 GB WAN bandwidth consumption makes the server down in a week [1].

B Sandpiper

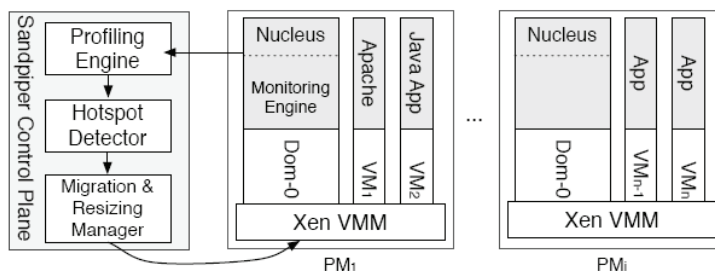


Figure 2: Sandpiper Architecture [12]

- Sandpiper goes a step further and in addition to monitoring the workloads, corrective actions are also possible [12].
- Nucleus runs on each physical host, for gathering network resource usage stats of the each host and then sends these stats to control plane.
- Control Plane uses automated resource management techniques and consists of:
 - Profiling Engine, which constructs resource usage profiles for each VM and host.
 - Hotspot Detector, which checks if the aggregate usage of processor, memory or network resource exceeds a threshold level.
 - Migration & Resizing Manager, which eliminates hotspots by
 - Assigning new resource shares to a VM
 - Or migrating it to another physical host.

- Two Monitoring modes
 - Blackboxes suitable for environments where detailed peeking inside a VM is not possible. So on the basis of basic stats resource allocation is increased by a constant amount Δm if a hotspot is detected.
 - Graybox can use detailed OS and application level statistics. So accurate prediction of resource needs is possible.

Limitations

- For a large resource hungry VM, constant amount Δm in the black-box approach may not be very efficient.
- On other side of the picture, this constant amount may be quite large and result in over-allocation of resources.
- Migrating it to another physical host leave a VM unresponsive for a small period of time.
- VM on a single instance for their users, a migration may result in denial of service.

C Remus

basic aim is to provide a general purpose high availability service that offers a high degree of fault tolerance without any modifications to software applications or the hardware

- It allows applications to switch on an alternate host within seconds of a failure [13].
- The Live Migration functionality is modified to replicate snapshots of running OS instances at frequencies as high as every 25ms. That is Remus is an extension of the ability of virtualization to migrate running VMs between hosts.

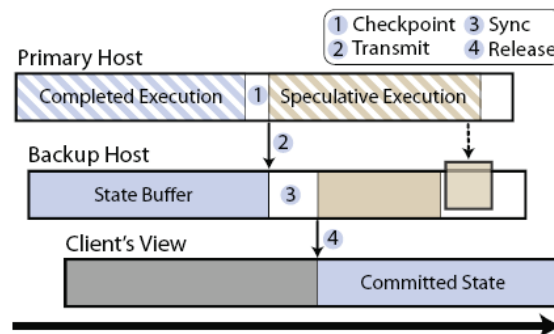


Figure 3: Asynchronous replication in Remus

- Asynchronous replication in Remus [15] shows the basic operation of Remus. At the checkpoint, the system state is transmitted to the backup host. And once the synchronization is complete, the output is released to the client.
- Speculative execution time allows the primary host to stay productive while its state is transferred asynchronously.
- Fast replicated performance is achieved by running the system tens of milliseconds in the past.

- A single physical machine can serve as a backup host for multiple physical hosts.
- In comparison to commercial High Availability products that respond to a failure by rebooting the VM on another host, Remus provides a greater degree of protection. It allows recovery from failures on near live migration time frames leaving the VM running with the network connections intact. The state visible to the clients stays consistent and the disks are not corrupted.
- Failure detection mechanism is also integrated in the checkpoint process. When backup host stops responding, primary host assumes that backup has crashed. Similarly, when backup host stops receiving new checkpoints from primary host, it assumes a failure and resumes execution from the most recent checkpoint.

Limitations. Remus aims to reduce the cost associated with ensuring high availability but additional hardware is still required in the form of backup hosts.

- The paper suggests that a single backup host can be used in N to 1 configuration to serve a number of active hosts but in case of large data centres, the number of dedicated backup hosts might be quite large resulting in significant additional expense.
- During tests, it has been noticed that Remus does introduce significant network delay, especially for applications that have poor locality in memory writes. Therefore, for applications requiring low latency, Remus may not be the right choice.
- When VM is transferred over the network to the backup host, it may result in excessive network traffic.

D *Live Migration across Wide Area Networks (WANs)*

The various techniques discussed above enable Live VM migration in a LAN environment making data centre management easier and non-disruptive. In a WAN environment, performing a live migration can yield similar benefits.

- It can be especially useful in scenarios where the data centre as a whole becomes unavailable due to catastrophic events or data centre wide maintenance operations. In such situations, migrating services to another data centre may be the only viable solution.
- Unfortunately, performing live migration across WAN involves a number of complexities and challenges. The cooperative context aware approach proposes a strategy where the server, network and storage cooperate and coordinate in such a way that they are aware of the service context to achieve seamless migration over WAN [14].
- The IP address mobility problem is addressed by using tunnelling technology to establish connectivity between data centres over WAN on a pre-emptive basis.
- Storage replication is performed in two steps. Initial transfer of data takes place in asynchronous mode. Afterwards, synchronous replication is used to ensure consistency and prevent data loss during migration process.

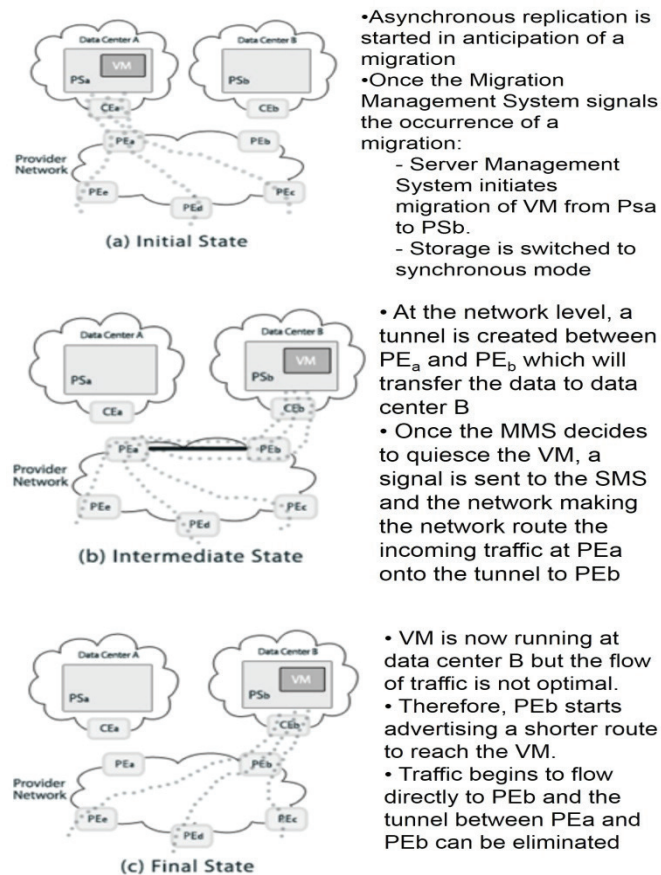


Figure 4: Live Server Migration across Wide Area Network [10]

Limitations

- Migrating VMs over WAN can be a very slow process. In comparison to live migrations over LAN, total migration time will be exponentially high. For VMs with large amount of state information, the technique may not be feasible. Especially in case of complete data centre outages requiring every single VM to be migrated to the backup site. Using this approach will require large amount of network bandwidth and an unreasonably long time frame.
- Using asynchronous replication on a continuous basis for transferring state information over WAN to the backup site might be a more feasible option. In case of an outage or failure, only the pending state changes will be transferred to the backup site decreasing the total time for migration. However, this approach will result in increased performance overheads under normal operation.

E Cloud Net and Cloud to Cloud Migration

- Uses Virtual Private Networks to connect multiple clouds forming Virtual Private Clouds [15].
- Through this technique, cloud to cloud migration can become as simple as performing a living migration over the local network.

- The architecture consists of two controllers: Cloud Manager and Network Manager that automate the management of resources between the two clouds.
- Cloud Manager handles the creation of new virtual machines. It is also responsible for managing the performance within the VPC. By partitioning physical routers into slices with independent control planes for each, virtual or logical routers are created and used to configure the Customer Edge routers in the cloud.
- The Network Manager creates VPNs and provides resource provisioning capabilities. These VPNs eliminate the need for endhost configurations and have lower overheads.
- CloudNet ties networks across WAN into a single LAN. As a result, VM migrations just take place as on a local area network.

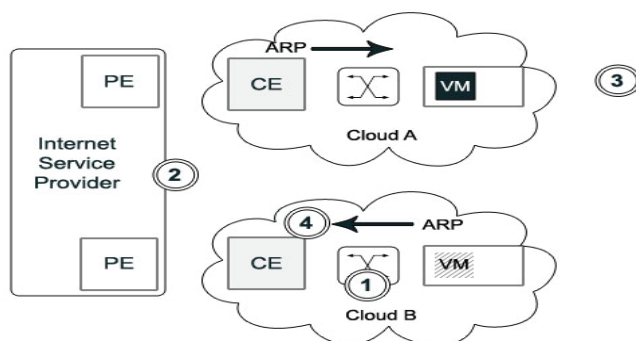


Figure 5: WAN migration using CloudNet [15]

- CloudNet first initializes a virtual LAN (VLAN) endpoint for the destination cloud. A VPN is created to link together the source and destination VLANs after which it is possible to migrate a VM between the two sites. Once the VM is transferred to the other cloud, it transmits an ARP message. This message is used by the local switch to map the VM's MAC address and its switch port. It is also forwarded to the VM's original site where the old switch will replace its MAC address mapping with the new entry. From this point onwards, the data will be forwarded to the VM's new location.

Limitations

- Network delays is quite high. Transferring the VM state over WAN can be a very slow process making it less viable for performing load balancing.

F Ghost Virtual Machines

The concept of Ghost VMs takes high availability to a new level by providing almost instantaneous resource allocations [16].

- Ghost VMs are pre-deployed spare VMs maintained on each physical server that remain in active state but do not service client requests as they are detached from the internet. These ghost VMs can be activated as and when required by making them visible to the content switch that handles client requests. These VMs consume minimal resources on their host machines in idle state.
- Although they remain hidden from the internet, the ghost VMs can communicate with each other using the host system's second network card or through the L2 functionality of the switch. When another VM is required for an application, a compatible ghost VM

can be activated by reconfiguring the switch. The process is extremely fast and has been shown to be completed within as little as 2 seconds.

- **Decision Manager Component:** Manages the virtual machines and performs the resource management function. This component performs its resource management function in three stages. It decreases the capacity for applications with over-allocation of resources. Next, it increases the resource allocations for applications that are under-allocated. Finally, it ensures that ghost VMs will be available to applications that are likely to see an increase in demand.
- **Ghost Manager Component:** Allocates ghost VMs to applications and ensures availability of spare ghost VMs

Limitations

- In a public cloud with a large number of heterogeneous applications running, it may not be possible to maintain a ghost VM for every different application.
- If application is overloaded, the ghost manager searches for compatible ghost VM. In case, if a compatible VM is not found, the ghost manager stops the running application and restarts the application. This is a highly undesirable scenario as the time taken to start the application will be much longer.

G *ZZ: Cheap Byzantine Fault Tolerance Using Virtualization*

- Fault tolerance is a highly desirable trait for systems requiring high availability. Byzantine Fault Tolerance (BFT) is a technique that uses replication to minimize the impact of faults. BFT is highly effective in dealing with faults but its high cost prevents its widespread adoption. It requires at least $2f+1$ execution replicas to tolerate f Byzantine faults.
- ZZ is a system that uses virtualization to decrease the cost associated with BFT to half [13]. In order to tolerate 2 faults, a typical BFT system will require $2f+1=5$ replicas. ZZ can decrease the number of replicas to $f+1=3$ resulting in significant cost savings and the usage of fewer resources to process non-faulty requests. The basic concept behind ZZ is to use additional replicas only in the event of a failure. Virtualization can enable these additional replicas to be activated on demand.
- In order to decrease the cost of implementing BFT, ZZ uses 3 mechanisms. First, it uses virtualization to enable fast replica start-ups and to multiplex a small pool of free servers across recovery replicas for a large number of applications. Second, it employs a recovery protocol that allows additional replicas to begin processing requests through a state transfer mechanism that fetches state on demand. Finally, ZZ performs incremental check-pointing by exploiting the snapshot feature of modern file systems without requiring modifications to the application source code.
- In a typical data centre, N independent applications will be running at a given time. During normal operation, $f+1$ replicas will be run as separate virtual machines. To run these replicas, a small pool of free servers can be multiplexed. The system makes the assumption that not all N applications will experience a fault simultaneously. Therefore we can use a smaller pool of $f+1$ replicas instead of the standard $2f+1$.

- A ratio “r” is used to determine the number of backup servers required for fault-mode execution and denotes the ratio of the time for ZZ to recover and replace a faulty replica to the mean time to failure for the application. Usually, this ratio is less than one as the recovery time is in seconds and the mean time to failure can be in days or more.
- Virtualization enables replicas to be started instantaneously in the event of a failure. The replicas can be maintained as VMs in paused state which requires no CPU consumption and negligible memory. On detecting a failure, the VMs can be un-paused within milliseconds.
- ZZ is based on the Xen platform and has been shown to recover a replica with 400MB of disk state in less than 4 seconds in test results whereas the existing approaches required 60 seconds. It incurs minimal overhead during normal operation.

Limitations

- In a public cloud with a large number of different applications running in parallel, the probability and the number of failures simultaneously occurring is difficult to determine. Therefore, the calculation of the ratio “r” may prove to quite difficult as it requires knowledge about the applications and their mean time to failures.

5 Development Of Live Migration Implementation

In order to support our research, we started off with a small project that aimed to setup a cloud computing environment. The purpose was to get hands on experience of live migration using one of the most popular cloud platforms called Xen. The research also includes the development of a small application that will enable users to manage their cloud by letting them perform basic operations as well as load balancing through live migrations and dynamic resource provisioning. Although due to a number of constraints and challenges, the load balancing part could not be completed.

A cloud environment was setup using Xen Cloud Platform. The purpose is:

- Hands on experience of setting up and configuring a cloud environment
- Perform Live migration and observe its efficacy with regards to ensuring high availability
- Provide high availability and service levels based on the defined QoS parameters

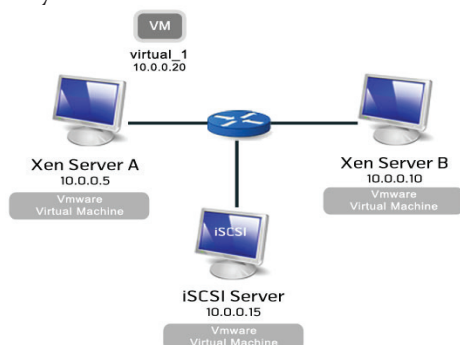


Figure 6: Test Cloud Architecture

Figure 6 shows the basic architecture of our test cloud. There are two physical hosts XenServerA and XenServerB. A separate iSCSI Server is used to provide the shared storage. The three servers are connected with each other through a router. In order to perform live migration, the virtual machine to be migrated has to be reside on a shared storage device so that it can be accessible from both hosts during the migration process. Due to this reason, virtual_1 uses iSCSI Server as its storage device. Note that due to the unavailability of dedicated hardware, we are running these servers as virtual machines on VMWare Workstation 6.5 using two physical hosts [18].

Table 1: Server Configurations

Machine hosting XenServer A (PM1)	Machine Hosting XenServer B (PM2)
AMD Athlon X2 @ 1.9GHz	Intel Core 2 Duo E6500 @ 2.66GHz
2.5GB of RAM	3GB of RAM
120 GB HDD	320GB HDD
VMWare Workstation 6.5	VMWare Workstation 6.5

Table 1 shows the configuration of the physical hosts PM1 and PM2. In addition to running XenServerA, PM1 also hosts the virtual machine for Open Filer iSCSI server [15].

A Management Interface

A software application has been developed to provide the management interface for the test cloud. The application makes use of XenAPI provided as part of the Xen Server Software Development Kit [19]. XenAPI exposes functions and parameters that allow remote configuration and management of a Xen Server and the associated guest VMs. The application enables users to connect to a specific server in the cloud and manage the virtual machines residing on that server. Detailed information of the virtual machines is displayed including their current state (started, paused, halted) and the number of processors assigned. Through the management interface, VMs can be started, paused, resumed, suspended or stopped. The user can start new instances of the same virtual machine by using the clone operation. Dynamic resource allocation can be performed by changing the amount of RAM allocated to each VM or migrating it to another physical host available in the pool.

B Migrating A VM

From the management interface, the user can perform live migration of a VM between two servers. Live migration is considered to be a non-disruptive operation that allows a running VM to be migrated to another host with unnoticeable downtime and all the network connections intact.

For live migration to take place, the VM should use a shared storage that is accessible to both source and destination servers. In my testing, I have installed the guest VM on XenServerA using iSCSI Server as the storage repository. The VM was pinged for network connectivity while it was being migrated to XenServerB. Packet loss was observed for a very brief moment approximating to 5 sec during which the final step of migration was being performed and the VM was active on XenServerB immediately afterwards. The total time taken by the migration process was around 4mins. Both the downtime and the overall migration time are quite high. Professional Xen implementations have reported downtimes in milliseconds and total

migration times less than 2 minutes. In my testing, I have used an extremely lightweight VM with small memory footprint. Therefore the performance is way below the standards. This poor performance can be easily attributed to the hardware used. Unlike commercial Xen implementations that use multi core multi-processor systems with high end SAN devices and Gigabit Ethernet, the test environment relied on notebook computers with limited system RAM and 100Mbps Ethernet. The performance was further degraded due to multiple layers of virtualization as the XenServers were also running as virtual machines.

C Challenges/Constraints

- Unavailability of Highly Specs Hardware: is a serious constraint. As a solution, VMWare workstation can be used.
- Installation of First Xen Server: is a difficult and time taking task. After a number of failed attempts and guidance from forums, we were able to get Xen running as a virtual machine.
- Guest VM installation: We need a guest VM with either Windows OS with HVM or Linux OS in paravirtualized mode. Windows OS with HVM was not available so we went with Linux OS and used Lightweight VM Xen SDK due to performance issue.
- Installation of Second Xen Server: is required once again and two Xen Servers cannot run on a single machine. As a solution, second physical host is needed.
- Unavailability of shared storage device: As a solution, we came across an open source storage management appliance Open Filer.
- Creation of a Resource Pool (Tightly coupled collection of servers): VMs can be migrated between servers belonging to the same resource pool. Live migration is supported between homogenous platforms, otherwise, we need to join resource pool forcefully.

6 Conclusion

This research explores the systems and techniques that can make the cloud more reliable and highly available. Our work focuses on the automated techniques that exploit the benefits of virtualization and technologies such as live migration to provide fast and efficient fault tolerance and failure prevention mechanisms. We have analysed monitoring systems such as Ganglia and Sandpiper that gather resource usage statistics to detect hotspots and perform load balancing. Sandpiper's graybox monitoring mode enables it to proactively adjust resource allocations and prevent hotspots from occurring. The scalable nature of these systems makes them ideal for usage in a cloud environment. Replication is one of the basic techniques for achieving high availability. Remus provides availability as a service at the virtualization layer by using replication in an innovative way. Live migration enables running VMs to be migrated with minimal downtimes between physical hosts in a LAN environment. Some other techniques and systems that we have discussed include Ghost VMs and ZZ. Ghost VMs provide the benefits of live migration without the performance overheads and with decreased downtimes. Based on Xen platform, ZZ is a system that uses virtualization to make the popular Byzantine Fault Tolerance (BFT) technique more cost effective and viable for usage in the cloud. Although, these techniques can prove to be extremely beneficial in improving the availability of a distributed

environment, a number of challenges and constraints limit their effectiveness. The research briefly discusses these challenges. In the end, we have presented the details of our live migration implementation. Using Xen Cloud platform, we have setup a cloud environment and configured it to support live migrations. Moreover, a management console has been developed to facilitate administration of the cloud. We have also provided details of the challenges faced during setup and configuration.

7 Recommendations For Future Work

There are a number of opportunities to further extend this research. Additional work can be performed in the areas of fault tolerance and replication. Multiple systems and techniques can be combined to take a holistic approach towards availability in the cloud. For example, the ghost VM approach can be incorporated in sandpiper to further optimize the load balancing component, reducing downtime and performance overheads at the same time. Increased network traffic and performance overhead is a major drawback of the load balancing and monitoring techniques discussed above. Further research can be focused on devising techniques to lower these overheads. A monitoring system such as Ganglia can be integrated in the system allowing cloud administrator to get real time access to usage statistics. The addition of automated load balancing techniques can also prove to be equally useful.

References

- [1] Xen Cloud Platform XAPI: Open source software to build private and public clouds
- [2] VMWare ESX Server
- [3] D. Nurmi, R. Wolski, C. Grzegorzczak, G. Obertelli, S. Soman, L. Youseff, D. Zagorodnov, "The Eucalyptus Open-source Cloud-computing System", in the Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid.
- [4] Siddharth Jain, Rakesh Kumar, Anamika, Sunil Kumar Jangir, "A Comparative Study for Cloud Computing Platform on Open Source Software". An International Journal of Engineering & Technology (AIJET) Vol. 1, No. 2 (December, 2014)
- [5] Makhdoom Muhammad Naeem, *Hidayatullah Mahar Furqan Memon, Muhammad Siddique, Abdul Rauf, "AN OVERVIEW OF VIRTUALIZATION & CLOUD COMPUTING". Sci. Int.(Lahore),28(4),3799-3803 ,2016 ISSN 1013-5316;CODEN: SINTE 8
- [6] Rakesh Kumar , Sakshi Gupta, "Open Source Infrastructure for Cloud Computing Platform Using Eucalyptus". Global Journal of Computers & Technology Vol. 1, No. 2, December 24, 2014 www.gpcpublishing.com ISSN: 2394-501X
- [7] CitrixXen Center
- [8] VMWare High Availability
- [9] Gartner, "Gartner Highlights Key Predictions for IT Organizations and Users in 2010 and Beyond", STAMFORD, Conn., January 13, 2010.

- [10] M. L. Massie, B. N. Chun, and D. E. Culler, "The Ganglia Distributed Monitoring System: Design, Implementation, and Experience", 15th June 2014.
- [11] Chandramouli Reddy , Suchithra R "Virtual Machine Migration in Cloud Data Centers for Resource management". International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 5 Issue 09 September 2016 Page No.18029-18034
- [12] T. Wood, "Improving Data Center Resource Management, Deployment, and Availability with Virtualization", University of Massachusetts Amherst, September 2011.
- [12] Mukil Kesavan, Irfan Ahmad, Orran Krieger, Ravi Soundararajan, Ada Gavrilovska, and Karsten Schwan(2013). "Practical Compute Capacity Management for Virtualized Data Centers". IEEE TRANSACTIONS ON CLOUD COMPUTING, VOL. 1, NO. 1, JANUARY-JUNE 2013
- [13] B. Cully, G. Lefebvre, D. Meyer, M. Feeley, N. Hutchinson, and A. Warfield, "Remus: High Availability via Asynchronous Virtual Machine Replication", Dept. of Computer Sciences, University of British Columbia.
- [14] K.K. Ramakrishnan, P. Shenoy, J. V. Merwe, "Live Data Center Migration across WANs: A Robust Cooperative Context Aware Approach", AT&T Labs – Research / University of Massachusetts.
- [15] T. Wood, P. Shenoy, A. Gerber, K. K. Ramakrishnan, J. V. Merwe, "The Case for Enterprise-Ready Virtual Private Clouds", AT&T Labs – Research / University of Massachusetts.
- [16] W. Zhang, H. Qian, C. E. Wills, M. Rabinovich, "Agile Resource Management in a Virtualized Data Center", Worcester Polytechnic Institute, Worcester.
- [17] T. Wood, R. Singh, A. Venkataramani, and P. Shenoy, "ZZ: Cheap Practical BFT using Virtualization", University of Massachusetts Amherst, 2009.
- [18] VMWare Workstation Workstation for windows
- [19] OpenFiler Storage Management Appliance
- [20] Xen Software Development Kit

Detection of Myocardial Infarction in ECG Base Leads using Deep Convolutional Neural Networks

Awais M. Lodhi¹ Adnan N. Qureshi² Usman Sharif³ ZahidAshiq⁴

Abstract

Myocardial infarction (MI), commonly known as a heart attack, occurs when blood flow decreases or stops to a part of the heart, causing irreversible damage to the heart muscle. It is a leading cause of mortality around the world according to the WHO reports and, therefore, it is critical to estimate the location & extent of the damaged tissue. Similarly, localization of MI is also significantly important to correctly manage the patient medically and/or surgically. In this paper we propose & implement a system in which the signals from 6 leads (I, II, III, aVR, aVL, aVF) of the ECG are used to detect the cases with MI in the lateral & Inferior walls of the heart. The use of Convolutional Neural Networks (CNN) & a novel voting scheme provides acceptably accurate estimates of MI. The proposed algorithm has been validated on MI & Normal Healthy Controls from the Physio Net dataset. This approach is robust & can be used in the clinical & research settings.

Keywords - Machine Learning, Bio Medical Signal Processing, Artificial Intelligence, ECG.

1 Introduction

Myocardial infarction (MI) or simply heart attack is a life threatening condition which occurs when there is a decrease or absent blood flow to the heart, resulting in damage to the cardiac muscles [1]. Heart receives oxygen rich blood through specialized arteries called coronary arteries. In certain conditions, there is a blockage of these coronary arteries, resulting in decrease or absent blood flow to the heart. The segment of the heart muscles, which is deprived of oxygen got damage resulting in MI. If timely treatment is not given, this damage can even be fatal [2].

Worldwide, Myocardial infarction (MI) is a leading cause of morbidity and mortality [3]. According to WHO, there are 32.4 million myocardial infarctions worldwide every year and about 17.7 million people die of cardiac diseases every year, 31% of all global deaths [4]. A lot of deaths caused by MI, can be prevented by an early treatment. There is a very narrow time frame, in which if proper treatment is given, the damage can be reversed to some extent. Therefore, in the management of MI, an early detection is crucial.

The basic, first line investigation, used to detect heart attack is ECG (Electrocardiogram). An electrocardiogram is a picture of the electrical conduction of the heart and is depicted in time vs. amplitude plot as shown in Fig I [6]. It is a non-invasive, economic mode of investigation which is performed by placing electrodes on the skin. It can measure heart rate and heart rhythm, as well as indirectly provide information about blood flow to the heart muscle. By comparing it to the normal signals obtained, clinicians can diagnose a number of heart diseases [5, 6].

^{1,2,4}University of Central Punjab, Pakistan

³Punjab University College of Information Technology, Pakistan

¹awais.lodhi@ucp.edu.pk | ²dr.qureshi@ucp.edu.pk | ³usman.sharif@pucit.edu.pk | ⁴zahidashiq@ucp.edu.pk

A standardized system of electrode placement for an ECG has been developed. There are a maximum of 12 leads. However, ECG can also be performed by using 6 leads. These six leads are called lead I, II, III, aVL, aVR and aVF.

The signal from these leads can be used to detect the anatomic site of myocardial infarction. Different studies have shown that ST segment elevation in leads I & aVL indicate infarction in the Lateral wall while signal in the leads II, III and aVF reflect the health of the Inferior wall [7, 8].

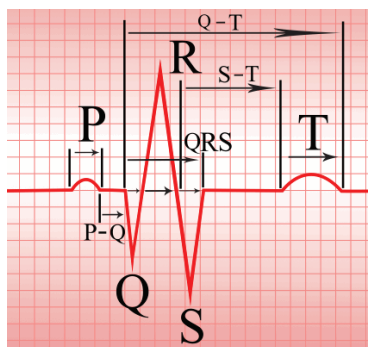


Figure 1: Components of the ECG Signal

Though ECG is a relatively simple test to perform, its interpretation requires significant amounts of training. Even after that, there is a huge room for error. As reported in [21] in which, 49% of the evaluators identified a normal ECG as ischemic and only 19% were able to correctly identify normal from abnormal scenarios. The same study [21] further reports that none of the evaluators were able to denote the correct leads, locations or amplitude of the ST-segment elevation. So a standardized, automatic method is required to interpret ECGs suggested in [20] especially in clinics where the experience, interpretation skills and/or expertise of the clinicians is below that of their counterparts in developed regions.

As seen in the [9, 10, 11], normally the ECG signal is de-noised and base line wandering is removed from the signal before extraction of features for further analysis using Artificially Intelligent systems, however as seen in [11], this process does not have a huge impact on the overall results of the system. Furthermore there is no consensus in the available literature regarding the features to be used for automatic MI detection from the ECG signals, therefore, we have followed the approach in [11] and have utilized deep learning approach for the task at hand. Also we propose that interpreting signals from multiple leads and then achieving a consensus (via voting) can achieve better results than [11]. This would reduce the misinterpretation(s) in case only signal from only one lead was evaluated.

Deep learning is a representation based learning which consists of an input layer, hidden layers, and an output layer [12]. This provides a network of systematic procedures which can be fed the raw data and the system automatically learns the necessary representations for classification. The term deep describes the multiple stages in the learning process of the network structure [12]. The deep learning neural network is trained using the backpropagation algorithm. The CNN is one of the most popular neural network techniques and provides quite desirable results [13].

A Convolutional Layers

The basic building block for any CNN is the convolutional layers which are responsible for most of the computational processing. These layers automatically extract the features from the input data.

B Rectifier Linear Unit Activation Function

Activation functions are an important part of the neural networks and are used to evaluate the excitement level of the neurons. Rectifier linear units (RELU) are interesting especially since they introduce nonlinearity in the data [17].

C Pooling Layers

Pooling function serve to down-sample and condense the features in the network thus reducing the overall computational complexity. Max-Pooling outputs only the maximum value from each kernel while sliding by a preset amount over the whole feature set. The sliding operation is called stride [17].

D Fully Connected Layer

Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular neural networks [17].

To prevent the misinterpretation in critical events such as MI, this paper proposes a system which can plausibly detect cases of MI in the lateral and inferior walls of the heart. The proposed method uses CNN and a voting scheme to demarcate MI from normal healthy controls more accurately. The rest of the paper is organized as: Section II – Dataset, Section III – Methodology, Section IV – Results, Section V – Discussion and Section VI – Conclusion.

Table 1: Characteristics of the Dataset

	MI	Healthy
Age Range	36 - 86	17 - 81
Average Age	60.37	43.4
Male	110	39
Female	38	13

2 Dataset

In this paper, we have used the Physikalisch-Technische Bundesanstalt diagnostic ECG database [14] available on PhysioNet[22]. This database contains 12 lead simultaneous signal data for 148 patients diagnosed with MI in different cardiac regions. Also data for 52 Healthy patients is available for control purposes. The signal has been sampled at 1000 Hz. The characteristics for the patients considered in this study are provided in the Table I.

3 Methodology

A Preprocessing

In this study we only considered patients with Acute MI involving the inferior or lateral wall of the cardiac muscle. First of all the R-peak detection was carried on the ECG signals of the selected patients using Pan Thompkins algorithm [15]. Then the signal was segmented using the detected R-peaks after omitting the first & last peak in each lead. Each segment consisted of 651 samples as described in [11] (250 samples before the R-peak and 400 samples after the peak ensuring the complete P-QRS-T wave is present in each segment. Z-Score normalization [19] was applied on each segment with the aim to resolve the amplitude scaling problem. Furthermore no noise removal or base line wandering correction was performed on any of the signals before or after the segmentation. The resultant segments were further divided in 2 different sets with ratio 80:20 for training & testing purposes respectively. After the preprocessing step we had 33,796 beats for the training set and 8,449 beats were set aside for testing purposes. The training set was now shuffled in order to ensure a randomized data is being fed to the neural network. Overall we had data for 42,245 individual beats out of which 31,671 were diagnosed as having MI in the Inferior or Lateral walls while 10,574 beats were extracted from ECG records of healthy patients for control purposes.

B Network Architecture

In this work we trained 6 individual Convolutional Neural Networks with similar structure as [11]. Each network consisted of 11 layers as shown in Figure II. The Input Layer (layer 0) takes 651 inputs which is then fed to a 1-D convolutional layer with kernel size 102. Next max pooling with step size 2 and stride of 2 is applied on each feature map in layer 2. This reduces the number of neuron from 550x3 to 275x3. In layer 3 the feature map is again convolved with kernel with filter size 24, following which max-pooling similar to layer 2 is once again applied in layer 4 reducing the neurons further to 126x10. Convolution with filter size 11 is once again applied in layer 5 followed by another max-pooling in layer 6 reducing the neurons further to 58x10. A final convolution with filter size 9 is applied to this feature map in layer 7 followed by another max-pooling in layer 8. We are now left with a feature map with 25x10 neurons. This feature map is flattened and fed to a fully connected layer with output size 30. Layer 10 is again fully connected with 10 units. Finally fully connected layer # 11 has the 2 output neurons.

Convolutional Neural Network (CNN) Architecture

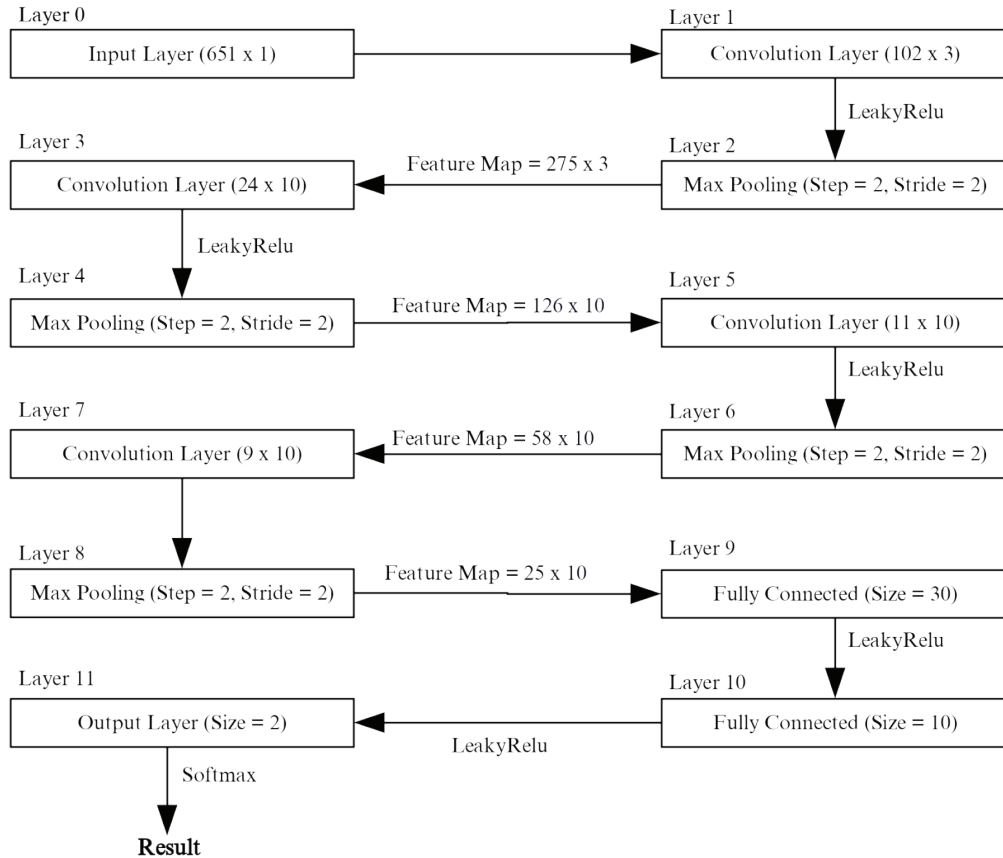


Figure 2: Convolution Neural Network (CNN) Architecture. All the 6 CNNs have same architecture for respective leads.

Multiple Lead MI Estimator

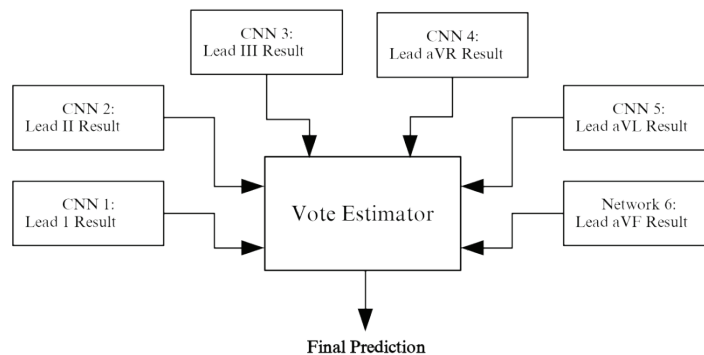


Figure 3: Vote Estimator. The results of respective CNNs are input to the Vote Estimator which generates final prediction based on consensus.

Table 2: MI Detection using Individual Leads

Lead	Train Accuracy	Test Accuracy	Sensitivity	Specificity	PPV	NPV
I	91.59%	92.20%	94.97%	83.86%	94.66%	84.71%
II	96.49%	96.39%	97.93%	91.74%	97.27%	93.65%
III	97.35%	97.47%	99.51%	91.31%	97.18%	98.41%
aVR	93.82%	92.46%	93.05%	90.70%	96.79%	81.25%
aVL	96.34%	97.46%	98.20%	95.21%	98.40%	94.62%
aVF	97.59%	97.82%	98.36%	96.20%	98.73%	95.12%

Table 3: MI Detection Using Multiple Leads

Votes	Accuracy	Sensitivity	Specificity	PPV	NPV
1	93.53%	100.00%	74.04%	92.06%	100.00%
2	96.84%	100.00%	87.33%	95.96%	100.00%
3	98.01%	99.95%	92.17%	97.46%	99.85%
4	98.98%	99.72%	96.77%	98.94%	99.12%
5	97.61%	97.18%	98.91%	99.63%	92.09%
6	88.83%	85.18%	99.81%	99.93%	69.11%

Leaky Rectifier Unit (LeakyRelu)[17] was used as activation functions in layers 1, 3, 5, 7, 9 and 10 and softmax was implemented in for the output layer.

C *Vote Estimator*

The output from the 6 individual CNN is generated as absence or presence of MI in the respective lead. These are input to the Vote Estimator (Fig III) which draws a consensus. If the presence of MI is indicated in at least 4 or more leads only then the system raises a true response.

D *Training*

The CNN was trained with Adam Stochastic Optimizer [16, 18] with batch size 500 for each lead. The learning rate, beta1, beta2 and decay parameters are set to 0.001, 0.9, 0.999 and 0 respectively for each network while minimizing the mean squared error. These parameters perform the following functionalities in the network [16, 18]

Learning Rate:	Helps with Data Convergence
Beta1:	The exponential decay rate for the first moment estimates
Beta2:	The exponential decay rate for the second-moment estimates
Decay:	The learning rate decay over each update

E *Implementation, Testing & Validation*

We ran each network for 25 epochs of training & testing rounds. After the completion of each epoch, the network was validated with the testing data for the corresponding lead.

We did not employ cross-validation in our current work. Finally after all the networks have been trained and tested with train & test beats data for their corresponding leads we once again tasked the trained networks for prediction on the test data for the respective leads. A final score for each test record was calculated based on the minimum number of networks voting the record as a healthy beat or indicating the patient had MI.

4 Results

In this study we implemented the above mentioned system using Keras® and Tensorflow® on Python®.

The algorithm was trained on a machine with Intel core i5 processor and 16 GB Ram. We also utilized nVidia GTX 950 GPU with 4 GB VRAM to aid in our training. It took approximately 200 seconds to completely train 25 epochs for each network. The training & testing accuracy along with sensitivity, specificity, PPV & NPV [19] for each lead is presented in Table II and visualized in Figure IV.

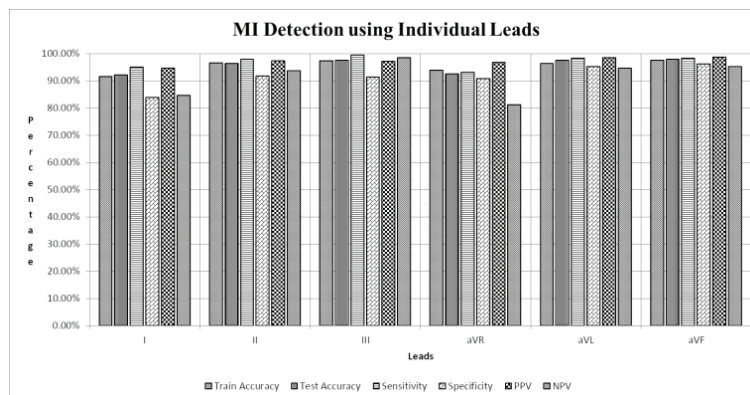


Figure 4: MI Detection using Individual Leads

It can be seen that individually each lead was able to accurately determine approximately 95% of the test data while the sensitivity & specificity figures also generally remained in high and low 90s respectively rarely rising above the 99% and 95% marks respectively.

Next step was the estimation of the test data using the novel voting scheme described above. We calculated the MI estimates from where only 1 lead marked the record as positive to where all 6 participating networks had to agree before classifying the record as a positive for the disease. The testing accuracy, sensitivity & specificity results for the voting are presented in Table III and have also been visualized in Figure V.

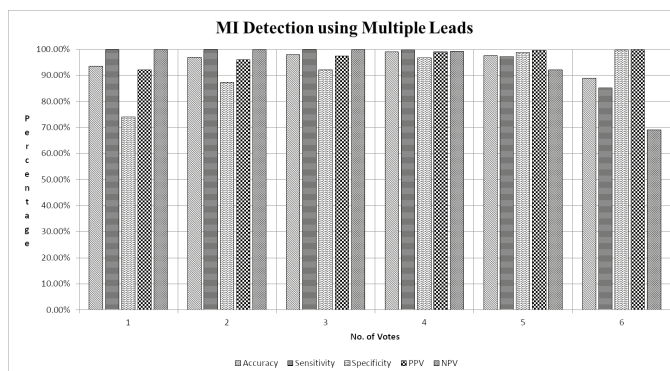


Figure 5: MI Detection using Multiple Leads

Here it can be observed that we can achieve 98 percent accuracy with the requirement of at least 3 leads agreeing on the outcome, while achieving the sensitivity at 99.9 percent mark. Requiring a simple majority i.e. at least 4 networks to agree on the outcome, we have approximately 99 percent accuracy while the sensitivity and specificity figures are also in the top half of 90 percent. However requiring all 6 leads to agree on the positive outcome, we see the changes in every category, with accuracy and sensitivity and specificity reduced to 88.83% and 85.18% respectively while increasing the specificity level to 99.81% as can be observed in Table III and visualization of the same is Fig V. The Confusion Matrix for 4 or more votes is also provided in Figure VI.

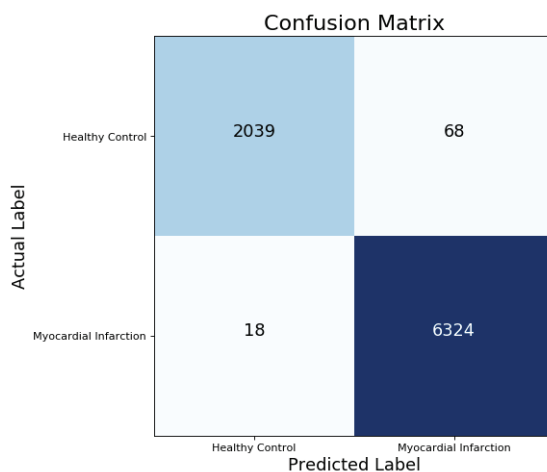


Figure 6: Confusion Matrix for 4 or more Leads

5 Discussion

The detection of MI using ECG as a non-invasive method is well established and it has been shown in literature that the response gathered through different leads of ECG can be useful in ascertaining MI s diagnosis. The experiments were designed on the above mentioned premise and the data was divided into healthy controls and MI patients and the ECG leads I, II, III, aVR, aVL and aVF were used in the algorithm.

When considering data from individual leads, it is observed that the test accuracy of more than 97% is achieved in leads III, aVL&aVF. This could be attributed to the fact that leads III and aVF represent changes due to MI in the inferior wall while the lead aVL shows changes due to MI of lateral wall. The data of the MI patients in the experiments includes samples from patients of inferior and lateral walls, as discussed in Section II. However this observation can be further exploited to localize the region of MI using the ECG base leads or all the 12 leads commonly used in emergency departments in hospitals.

To maximize the response of MI cases with higher sensitivity, a voting scheme has been adopted. The scheme considers that at least 'n' votes should favor the accuracy of final prediction and it is observed that combining the individual outputs of the respective CNN successively increases the accuracy of the proposed technique to more than 98% when $n \geq 4$ votes favor the prediction of MI case and the confusion matrix is provided in Fig V. The constraint of maximizing sensitivity and specificity is considered during evaluation. When the condition of $n \geq 5$ or $n \geq 6$ votes is imposed the results show decreasing accuracy and sensitivity which construes that not all leads had voted in tandem.

Hence the empirical optimization suggests that $n \geq 4$ votes should be considered in assessing the data for presence of MI will significantly improve results as compared to methods reported by other researchers.

6 Conclusion

Our work combines the concept of CNN combined with a voting scheme which uses prediction of MI from the 6 ECG leads. The results are better than the approach presented in [11] which focuses on detection of MI in the Lead II only and manages to achieve 93.53% test accuracy with 93.71% and 92.83% sensitivity and specificity respectively in the same PhysioNet dataset without baseline wander correction or applying any noise removal technique. We, on the other hand, are looking at data from the 6 base leads and the achieved accuracy, sensitivity and specificity are greater than 98%, 99% and 96% respectively when 4 or more leads vote together. Furthermore, we have limited our study to the patients with acute MI in the Inferior & Lateral walls which can be observed in the first 6 leads of the ECG signal.

Given the above 2 factors we have been able to get favorable results as compared to most recent literature. Our future approach will consider all 12 standard leads in the ECG signal and update the vote estimator algorithm to specify the probable walls where the acute MI is present in the patient and finally we also aim to validate our findings on other publically available datasets.

References

- [1]. Jeremias, Allen, and David Lloyd. Brown. "Diagnosis of Acute Myocardial Infarction." Cardiac Intensive Care. Philadelphia: Saunders Elsevier, 2010
- [2]. N. El-Sherif and C. Ramana Reddy, The Pathophysiology and Pharmacotherapy of Myocardial Infarction. Burlington: Elsevier Science, 2013.

- [3]. "Acute Myocardial Infarction", Clevelandclinicmeded.com, 2018. [Online]. Available: <http://www.clevelandclinicmeded.com/medicalpubs/diseasemanagement/cardiology/acute-myocardial-infarction/>. [Accessed: 05- Jan- 2018].
- [4]. American Heart Association. Heart Disease and Stroke Statistics 2017: At-a-Glance. Available from http://professional.heart.org/idc/groups/ahamah-public/@wcm/@sop/@smd/documents/downloadable/ucm_491265.pdf. [Accessed: 05- Jan- 2018].
- [5]. F. Benjamin Wedro, "What is an Electrocardiogram (ECG, EKG)?", eMedicineHealth, 2018. [Online]. Available: https://www.emedicinehealth.com/electrocardiogram_ecg/article_em.htm. [Accessed: 05- Jan- 2018].
- [6]. "Introduction to ECG", Healio.com, 2018. [Online]. Available: <https://www.healio.com/cardiology/learn-the-heart/ecg-review/ecg-interpretation-tutorial/introduction-to-the-ecg>. [Accessed: 05- Jan- 2018].
- [7]. R. Fuchs, S. Achuff, L. Grunwald, F. Yin and L. Griffith, "Electrocardiographic localization of coronary artery narrowings: studies during myocardial ischemia and infarction in patients with one- vessel disease", *Circulation*, vol. 66, no. 6, pp. 1168-1176, 1982.
- [8]. "ECG localization of myocardial infarction / ischemia and coronary artery occlusion (culprit) – ECG learning", ECG learning, 2018. [Online]. Available: <https://ecgwaves.com/localization-localize-myocardial-infarction-ischemia-coronary-artery-occlusion-culprit-stemi/>. [Accessed: 05- Jan- 2018].
- [9]. Acharya. U. R., Fujita. H., Sudarshan. V. K., Oh. S. L., Adam. M., Koh. J. E. W., Tan. J. H., Ghista. D. N., Martis. R. J., Chua. K. C., Chua. K. P., Tan. R. S. Automated Detection and Localization of Myocardial Infarction Using Electrocardiogram: A Comparative Study of Different Leads. *Knowledge-Based Systems* 99: 146-156, 2016.
- [10]. Arif. M., Malagore. I. A., Afsar. F. A. Detection and Localization of Myocardial Infarction Using K-nearest Neighbor Classifier. *Journal of Medical Systems* 36: 279-289, 2012.
- [11]. U. Acharya, H. Fujita, S. Oh, Y. Hagiwara, J. Tan and M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals", *Information Sciences*, vol. 415-416, pp. 190-198, 2017.
- [12]. LeCun Y, Bengio. Y., Hinton. G. Deep Learning. *Nature* 521: 436-444, 2015. 24.
- [13]. Greenspan. H., Summers. R. M., van Ginneken. B. Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging* 35(5): 1153-1159, 2016.
- [14]. R. Boussejot, D. Kreiseler, and A. Schnabel, "Nutzung der ekg-signaldatenbankcardiodat der ptüber das internet," *Biomedizinische Technik/Biomedical Engineering*, vol. 40, no. s1, pp. 317-318, 1995.

- [15]. Pan. J., Tompkins. W. J. A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering 32(3): 230-236, 1985.
- [16]. Bouvrie. J. Notes on Convolutional Neural Network, 2007.
- [17]. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems Conference 25, 2012.
- [18]. Kingma, Diederik P, and Jimmy Lei Ba. "ADAM: A Method For Stochastic Optimization."
- [19]. R. Mould and R. Mould, Introductory Medical Statistics. Bristol, England: A. Hilger, 1989
- [20]. J. Semmlow and B. Griffel, Biosignal and Medical Image Processing, Third Edition. Hoboken: CRC Press, 2014.
- [21]. Stephens, Kimberly E., et al. "Interpreting 12-Lead Electrocardiograms for Acute ST-Elevation Myocardial Infarction." The Journal of Cardiovascular Nursing, vol. 22, no. 3, 2007, pp. 186-193
- [22]. Goldberger. A. L., Amaral. L. A. N., Glass. L., Hausdorff. J. M., Ivanov. P. C. H., Mark. R. G., Mietus. J. E., Moody. G. B., Peng. C. K., Stanley. H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23): e215-e220, 2000.

Mining Diagnostic Investigation Process

Sohail Imran¹

Dr. Tariq Mahmood²

Abstract

Diagnostic investigation process of healthcare is complex, medical practitioners' goal is to find methods of standardizing their diagnostic investigation processes to reduce the time and cost and optimize the quality of healthcare. The technique that can be applied to mine valuable and useful knowledge of diagnostic investigation process of their interest from stored data is process mining. Process does not consider dynamic and causal dependencies in processes. This characteristic of process mining can be effectively applied in diagnostic investigation. This technique becomes more helpful and valuable where some treatment failed to provide favouring evidence. We used process mining in this paper to mine efficient diagnostic investigation process flow for hepatitis patients. There are several advantages of using process mining approach which can boost the effectiveness diagnostic investigation processes.

Keywords: Process mining, healthcare, diagnostic investigation process, process flow.

1 Introduction

The emerging discipline evolved from the data mining is process mining [1]. The objective of process mining approach is the reengineering of a given business process model and visualized a certain aspect. Information systems are the backbone of executing process mining techniques.

The vital element of healthcare system is the correct diagnosis of a patient's condition, and the most important part of this process is diagnostic investigation. Diagnostic investigations provide a critical component of the patient's visit within the clinical management [2]. The information these tests provide affect the healthcare decisions. As the complexity of diagnostic investigation are increasing, conventional approaches to diagnose patient's condition are becoming increasingly inadequate [3].

Currently, there is a need of bridging gap in translational healthcare research between the diagnostic investigations associated with a particular disease and proving that patients who are tested for these diagnostic investigations have better outcomes than those who are not. Studies of diagnostic investigation accuracy are not sufficient to justify clinical use [4], [5]. Implementation of basic process mining techniques applied to real-life data and business problems is well documented [6]-[8]. Process mining can be effectively used to different business domain including healthcare. To improve diagnostic, treatment decisions and unaided human inference, process mining could be considered [9], [10].

Several laboratory units can be the part of healthcare diagnostic investigation process of patients. Laboratory test units are operated using specialized software modules, it is difficult to get data in uniform format for a specific disease.

¹PAF-Karachi Institute of Economics & Technology, Karachi, Pakistan | sohail@pafkiet.edu.pk

²PAF-Karachi Institute of Economics & Technology, Karachi, Pakistan | mahmood@pafkiet.edu.pk

Event logs generated from these information systems are used to extract related knowledge [11]. There are several advantages of using process mining approach which can boost the effectiveness diagnostic investigation processes. In this paper, we have shown that it is possible to extract a useful and helpful picture of the real process without having in depth knowledge of the complex hospital process. We have presented different path maps. This can help not only medical practitioners but also forces health experts to unify diagnostic investigation process.

We have organized the remainder of this paper as follows: brief overview of process mining and the relationship between process mining and healthcare is presented in Section II. The description of the tool used to implement process mining in this paper is summarized in Section III. The process mining application is described in Section IV for healthcare diagnostic investigation using data, to obtain insights in an explorative manner. To get valuable insight the focus should not be on one aspect only. Therefore, we applied diagnostic investigations suggested to the patients on their visits and diagnostic investigation suggested by the medical practitioners. Finally, Section V concludes the paper and sketches a line for future research directions.

2 Related Work

A Process Mining

Process mining is the reverse engineering technique that uses event logs that are produced by information systems [7] [8]. Extraction of related knowledge from these recorded event logs is the basic idea of the application of process mining. Vast number of logs of events are generating against different activities of systems. These logs provided the basis for process mining. The aim of process mining is to extract valuable insights from the recorded event logs containing actual executed process instances and facts [12], [13].

Three process mining kinds are: i) Discovery, ii) Conformance analysis and iii) Extension. The first is no a priori model. It discovers previously unknown or undocumented processes from low-level events. It is performed when no model exists or when the quality of the existing documentation is poor. The event logs are analysed in order to discover the process. It generated documentation for the process. The Conformance analysis is an a priori model. It analyzes the deviation between the event log and the model. Its compares whether the current workflow conforms to the planned process. It generated discrepancies between the existing process and the model [7].

After locating discrepancies, processes are analyzed to suggest improvements in it. If these suggestions are beneficial to the process, the model is then reviewed to avoid these discrepancies using the information available at each node. Otherwise changes may then be made in the current process to conform to the model. The third kind of process mining is a prior model too. The goal of this kind of process mining is to find improvement in the existing model by extending the model. To find possible space of improvements in the existing model, the event logs are analyzed to get extension in the model or to get possible alternative paths in the workflow. Process models can be divided into two types: de jure process model and de

facto process model. The first type of process model is normative, its purpose is to design future processes by incorporating enhancements in the existing model [8]. A second type process model is descriptive. Its purpose is to capture reality by mapping current processes to create a baseline for process improvements.

The event log data can also be divided into two types: post and pre mortem. The first type event log data is about completed cases which can be used for process enhancements. It does not change or affect the referred cases. The second type of event log data is about uncompleted cases which are currently in progress.

B Process Mining in Healthcare

Process mining can be used and implemented in Healthcare management for diagnosis and treatment [9]. To improve treatment decisions and unaided human inference, this technique can be used. The process mining application is very effective in diagnosing and detecting disease. It is useful in situation where some treatment failed to provide favouring evidence.

Optimal allocation of available healthcare resources is another key area where process mining can play an important role. Apart from getting benefit of process mining in healthcare, data mining is also effective in discovering process flow of treatment of diseases. Process mining can save cost and time involved in conventional techniques by analysing event log files to recognize treatment flow. Process mining provides view of interaction among different diagnostic investigation of a disease. This view of interaction detects anomalies before they become problems and discover the actual problems instead of immediately visible symptoms. It discovers medical practitioners' similar prescription process based on verifiable data [14].

Process mining establishes baselines for the existing diagnostic investigations and uses them to determine whether specific changes are effective or not. It helps in understand the entire process. Since process flow causes of the disease can be identified and seen in the context of the prescriptions, medical practitioners understand where and why change is needed.

The identification of diagnostic investigation patterns and the eventual satisfaction they result in can be used to improve overall patient satisfaction. In many cases prediction of diagnostic investigation can aid in designing proactive initiatives to reduce overall cost and increase patient satisfaction. To optimize the execution of processes, one of the best practices follow today is the standardization. Standardization of diagnostic investigation process could make the usage of health resources more effective that optimizes quality and efficiency of patient care [15], [16]. The application of process mining has several advantages to boost the effectiveness of medical diagnostic investigation processes [17].

This can help not only medical practitioners but also forces health experts to unify diagnostic investigation process. Along with the improvement in the quality of services using this approach, variations in daily diagnostic investigation practices can also be avoided. Effective resource management of healthcare is another benefit of this approach. This result in improved foresees and account the costs of treatment of patients.

3. Tool For Analysis

To analyze our data set, we used Fluxicon Disco process mining tool from the aspect of process diagnostics. The Disco is based on the Fuzzy miner. The Fuzzy Miner is a mining algorithm used to introduce the map metaphor to process mining. It includes seamless process simplification and highlighting of frequent activities and paths.

The data set is a collection of events applied for process mining is referred as event log. In data mining, an individual record represents a complete process instance but in process mining, an event is just an individual row. For the application of process mining, event log is the starting point. The analysis of data stored in event logs from the aspect of a process is the basic aim of the process mining. Process mining maps the data to a process view.

To apply process mining, the data need to fulfil certain minimum requirements. The three key elements are the identification of cases, activities and timestamps. The scope of the process is determined by the case and the activity determines the depth of level for the process steps. Every event that was executed in the process refers to a process instance or a case. Each case is a collection of multiple connected events. Process mining compares several executions of the process to one another. Another important requirement is timestamps. To get the right order of the events for each case, at least one timestamp is required otherwise the events are correctly ordered.

The most important analysis output in Disco is the process map. It generates the actual execution of the process. In the event log data, the time stamping and ordering of the stored activities is of most importance because Disco discovered the process path flows automatically on the bases of these two. The major advantage of using this approach is that we can extract auseful and helpful picture of the real process without having in depth knowledge of the complex process. Disco visualized the discovered process in a simple way: the stat of the process map is represented by a triangle symbol and a square symbol represents the end of the process. Boxes illustrated the activities and arrow shows the process flow between two activities. There are two types of arrows used in the tool. One solid line, discussed in earlier, and the dashed arrows visualized activities occurred at the very beginning or at the very end of the process. The numbers shown within activities and the arcs represent the frequency of processes and it also illustrated visually by the arrow thickness and the activity color. To generate the process map for our diagnostic investigation of patients, we used absolute frequency metric, based on total frequencies.

4 Process Mining For Healthcare Diagnostic Investigation

The application of process mining for healthcare diagnostic investigation will be shown in this section. In healthcare diagnostic investigation processes for the treatment of patients, several laboratory units can be involved and these laboratory test units are operated through specialized software modules, it is difficult to get data in uniform format for a specific disease. To get track of diagnostic investigation of patients for a specific disease, hospitals need to have integrated software modules of all laboratory test units.

Our event log contains information on the activities related to laboratory test prescribed to the patient on their visits by their medical practitioners in a hospital. The log contains events related to a patients diagnosed with hepatitis. The raw data used for this paper, sample data in Table1, contains information of 575 doctors, 10 diagnostic investigation and 1,313,177 records of hepatitis patients treated in 2010 to 2012. For hepatitis patients, the diagnostic investigation process is supported by several laboratory test units.

For this paper, event logs are extracted from the information systems database of the hospital. Each event in the data set is referred as a laboratory test of a patient. The rest of this section describes the application of process mining, for healthcare diagnostic investigation of hepatitis patients using data, to obtain insights in an explorative manner. So, the discovery part of process mining is our focal point not the conformance and extension part.

Table 1. A part of the event log

Visit ID	Doctor	Entry Date	Test Description
49359193	GJAV	31-Oct-10	Hemoglobin Haematocrit
49359193	GJAV	31-Oct-10	White Blood Cell Count
49359193	GJAV	31-Oct-10	S. Sodium
49359193	GJAV	31-Oct-10	S. Potassium
50617620	AJAF	3-Mar-10	Glucose Random
57921561	SMUN	11-Oct-10	Prothrombin Time

Figure 1 and Figure 2 show the process model of a patient visit-based view on the diagnostic investigation process and a doctor-based view on the diagnostic investigation process from the event log. To get valuable insight the focus should not be on one aspect only. Therefore, we have applied process mining from two perspectives: diagnostic investigation to patients on their visit and diagnostic investigation suggested by doctors.

In Figure 1 and Figure 2, diagnostic investigation process may consist of following activities: Neonatal TSH (NTSH), Complete Blood Count (CBC), Prothrombin Time (PT), APTT, Memoglobin Hematocrit (MH), Blood Urea Nitrogen (BUN), Creatinine (Cr), Serum Electrolytes (SE), Urine Detail Report (UDR), S. Sodium (SS), S. Potassium (SP), Hemoglobin Haematocrit (HH), Magnesium (Mg), Platelets (PI), White Blood Cell Count (WBCC), SGPT, Glucose Random (GR). For a single instance of diagnostic investigation process, some of these activities might repeat or it is not necessary that all activities happen every time.

A On Patient Visit

In Figure 1, each case in the event log corresponds to a diagnostic investigation on patient visit. It can be seen that 30,974 different cases of the diagnostic investigation processes are there in the event log. There are two alternative process can be observed from the beginning. Since the beginning of the diagnostic investigation process, 2,868 cases ended after only one activity NTSH laboratory test. The other 28,106 cases perform activity CBC instead. It is clearly shown that 90.7% of the activities are started with the activity CBC.

Afterward, the process split into four alternative paths. Out of four activities, diagnostic investigation process terminated after the prescription of HH laboratory test without proceeding to next activity. Another pattern can be noticed for the other alternative that loop backed to CBC laboratory test activity after performing PT and APPT laboratory test activities. The third alternative, started from the activity BUN laboratory test, is the central path flow in this

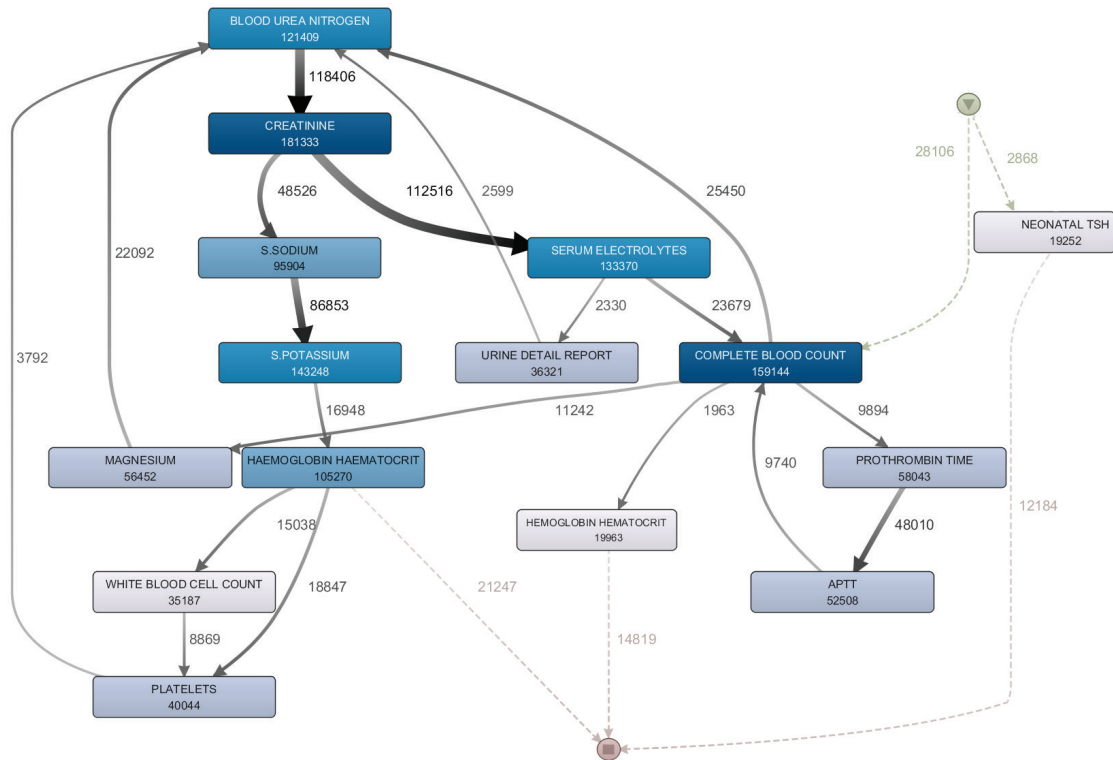


Figure 1: A patient visit-based view on the diagnostic investigation process

segment of diagnostic investigation process that can be visualized by the thickest arrow with a weight of 118,406 cases between BUN and Cr.

Overall, activity Cr laboratory test is most executed activity (in total 181,333 times). This activity further splits into two dominant loops. First, Cr-SE-CBC[UDR]-BUN-Cr loop and second Cr-SS-SP-HH-Pl[WBCC]-BUN-Cr loop. The fourth alternative is actually same process as third alternative with an additional activity Mg laboratory test, before joining it to the third alternative. 21,247 case of the second loop of third alternative completed and end with activity HH laboratory test.

The process map of Figure 1 discovered important results to diagnose hepatitis in the form of two important flows. The first flow CBC-BUN-Cr-SE-CBC[UDR]-BUN and the second CBC-BUN-Cr-SS-SP-HH-Pl[WBCC]-BUN identified most frequent laboratory test activities to diagnose the disease. Another analysis result was the detection of termination decision of disease treatment. The process terminated after NTSH and HH that showed activities helpful in decision making.

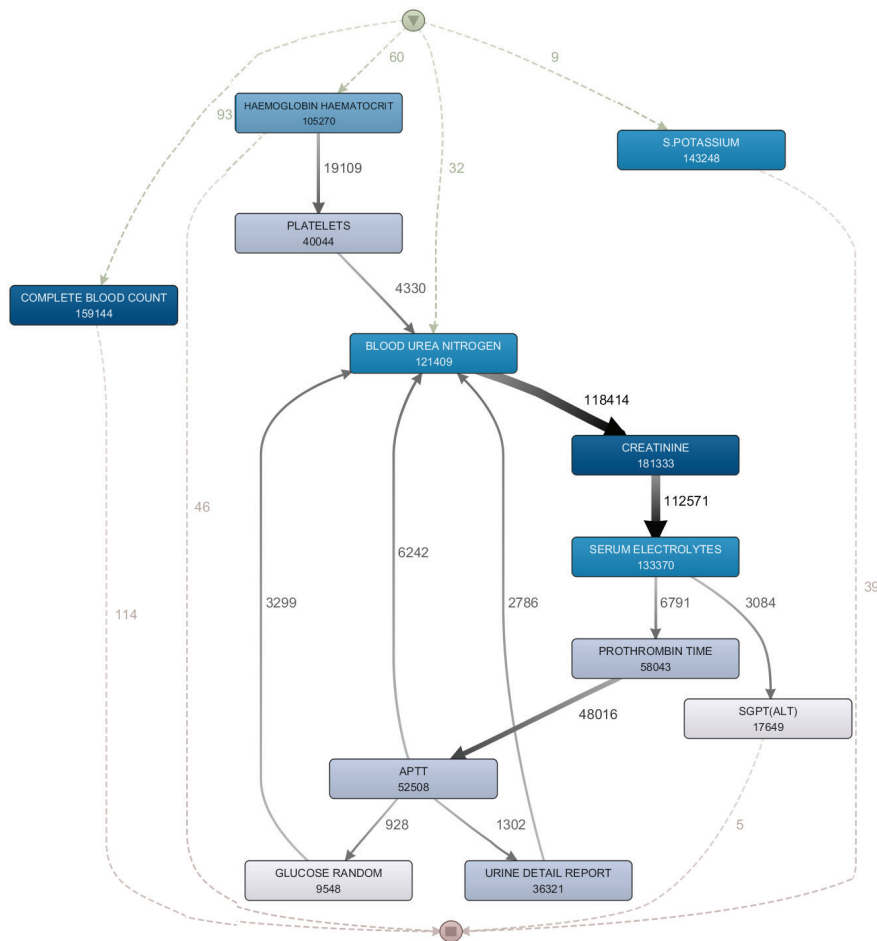


Figure 2: A doctor-based view on the diagnostic investigation process

B By Doctor

In Figure 2, each case in the event log corresponds to a diagnostic investigation by doctor. There are four alternative diagnostic investigation process can be observed from the beginning. Out of four, two ended after only one activity. First one ended after SP laboratory test, second after CBC.

The third alternative, started from HH laboratory test activity, further split into two fragments; first fragment ended in the similar fashion as of first and second alternatives. The other fragment merged with the fourth alternative path that started with BUN laboratory test, after performing an additional activity PI laboratory test. Afterward, the process continued Cr laboratory test activity. The second fragment of the third alternative, started from the activity HH laboratory test and merged with the activity BUN laboratory test after completing PI laboratory test activity, is the central path flow in this segment of the diagnostic investigation process.

A loop backed pattern to BUN laboratory test activity can be noticed after merging fourth alternative with the second fragment of third alternative. It is clearly shown that 99% of the

activities are started with the activity HH. Over all, activity Cr laboratory test is most executed activity (in total 118,414 times). This activity proceeds to SE laboratory test and then further splits into two. First one ended after performing SGPT laboratory test activity and the second one followed to APTT laboratory test activity. This activity further splits into two loops. First, APTT-UDR-BUN loop and second APTT-GR-BUN loop.

The process map of Figure 2 discovered important results to diagnose hepatitis in the form of an important flow. The flow BUN-Cr-SE-PT-APTT-[UDR|GR]-BUN is identified as most frequent test laboratory activities to diagnose the disease. Another analysis result was the detection of termination decision of disease treatment. The process terminated after SP, CBC, SG and HH that showed activities helpful in decision making.

5 Conclusions

This research paper showed that the effective application of process mining approach to the diagnostic investigation of healthcare domain data. We applied process mining from two perspectives to extract relevant knowledge: diagnostic investigation to patients on their visit and diagnostic investigation suggested prescribed by doctors. For these two aspects, we presented some initial results. We have shown that without having any existing diagnostic investigation process model and in depth knowledge of the complex daily diagnostic investigation processes, we can extract the objective picture of the real diagnostic investigation process using process mining approach.

In addition, to get valuable insight the focus should not be on one aspect only. Therefore, we applied diagnostic investigations suggested to the patients on their visits and diagnostic investigation suggested by the medical practitioners. We have presented different path maps. This can help not only medical practitioners but also forces health experts to unify diagnostic investigation process. Along with the improvement in the quality of services using this approach, variations in daily diagnostic investigation practices can also be avoided.

The process map discovered important results to diagnose hepatitis in the form of important flows. We identified most frequent test laboratory activities to diagnose the disease. Another analysis result was the detection of termination decision of disease treatment which is helpful in decision making. This paper recommended that a medical practitioners' group be established to discuss how best to develop a diagnostic investigation mechanism for the evaluation of new laboratory investigations.

This research has presented diagnostic investigation process for hepatitis patient. At the same way, these results evidence some facilities provided to the expert to guide their process to the knowledge about other diseases. Also we have data of one hospital only; using data of multiple hospitals can be used to compare diagnostic investigations of multiple hospitals. In future the application of process mining will be very effective to present generalized diagnostic investigation process using big data and NoSQL technologies.

References

- [1] W. M. P. van der Aalst, "Process Discovery: Capturing the Invisible," *IEEE Computational Intelligence Magazine*, vol. 5, no. 1, 2010, pp. 28–41.
- [2] P.J.F. Ramírez, R.T. Rodríguez, F.D. Olivera, V.M. Morejón, "Component to decision making in health. A social network analysis approach from process mining," *Revista Cubana de Informática Médica*, 8 (1), 2016, pp. 46-63.
- [3] C. Kulatunga-Moruzi, "Investigating the diagnostic process: The coordination of diagnostic rules and clinical experience," ETD Collection for McMaster University, 2007.
- [4] G. G. Raab, L.J. Logue, "Medicare Coverage of New Clinical Diagnostic Laboratory Tests: The Need for Coding and Payment Reforms," *Clinical Leadership & Management Review*, vol. 15, no. 6, 2001.
- [5] N. Foshay, C. Kuziemsky, "Towards an implementation framework for business intelligence in healthcare," *International Journal of Information Management*, 2014.
- [6] M. Dumas, W.M.P. van der Aalst, A.H. Hofstede, "Process Aware Information Systems: Bridging People and Software Through Process Technology," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 3, 2007, pp. 455–456.
- [7] W. M. P. van der Aalst, A. Andriansyah, A.K. Alves de Medeiros, F. Arcieri, T. Baier, T. Blicke, J.C. Bose, P. Van den Brand, R. Brandtjen, J. Buijs, "Process Mining Manifesto," in *Business Process Management 2011 Workshops Proceedings*, 2012, pp. 169–194.
- [8] R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, P.J.M. Bakker, "Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital," *Biomedical Engineering Systems and Technologies*, 2008, pp. 425–438.
- [9] H.C. Koh and G. Tan, "Data Mining applications in healthcare", *Journal of Healthcare Information Management*, vol. 19, no.2, 2005, pp. 64-72.
- [10] W.M.P. van der Aalst, B.F. van Dongen, J. Herbst, L. Maruster, G. Schimm, A.J.M.M. Weijters, "Workflow Mining: A Survey of Issues and Approaches," *Data and Knowledge Engineering*, vol. 47, no.2, 2003, pp. 237-267.
- [11] A.J.M.M. Weijters, W.M.P. van der Aalst, "Process Mining: Discovering Workflow Models from Event-Based Data," *13th Belgium-Netherlands Conference on Artificial Intelligence*, 2001, pp. 283–290.
- [12] W.M.P. van der Aalst, V. Rubin, B.F. Dongen, E. Kindler, C.W. Günther, "Process Mining: A Two-Step Approach to Balance Between Underfitting and Overfitting," *Software and Systems Modeling*, vol. 9, no. 1, 2010, pp. 87–111.
- [13] A. Rozinat, W.M.P. van der Aalst, "Conformance Checking of Processes Based on Monitoring Real Behavior," *Information Systems*, vol. 33, no. 1, 2008, pp. 64–95.

- [14] J. Tanawat, A. Poohridate, K. Kwanchai, I. Sarayut, P. Wichian, "Conformance analysis of outpatient data using process mining technique," ICT and Knowledge Engineering, 15th International Conference Proceedings, 2017.
- [15] N.R. Every, J. Hochman, R. Becker, S. Kopecky, C. P. Cannon, "Critical pathways. a review," National Library of Medicine and The National Institutes of Health, 2000, pp. 461-465.
- [16] M. Erraguntla, P. Tomasulo, K. Land, H. Kamel, M. Bravo, B. Whitaker, R. Mayer, S. Khaire, "Data Mining to Improve Safety of Blood Donation Process," System Sciences (HICSS), 2014 47th Hawaii International Conference, 2014, pp. 789-795.
- [17] H. Laihonon, "Knowledge structures of a health ecosystem," Journal of Health Organization and Management, 2012, pp. 542-558.

The Role of SEO Techniques to Enhanced Performance and Improved Rankingfor Intelli-Web Shop

Muhammad Noman khalid¹ Hira beenish² Muhammad Iqbal³ Kamran Rasheed⁴
Muhammad Talha⁵

Abstract

The use of internet is gaining popularity bercause of people freedom to connect each other and ultimately shrink the physical boundaries between different societies. This virtual world has a wide-ranging impact on communication since its rise and globalization. Additionally, it also brings new possibilities to individuals and companies who are mostly keeping in touch with web. Due to extensive use of internet, the web holds an immeasurable amount of data and search engines (se) are essential tools for finding, sorting, storing and ranking the value of that data on the web. The potential of SEO is very significant because search engine, such as google, bing, baidu, yahoo and their results routes end users drive a major portion of web traffic to specific website. Due to the vital role of SEs, search results have become more decisive for website owners to compete with other rivals. Search engine optimization (SEO) is a key process for getting better online visibility on search results from SEs. After employing SEO, website owners believe that their website position will appear before their rivals. Hence, there is inherent requirement for website developers to follow and apply SEO guidelines to address ranking issue. The objective of this study is to technically justify the importance of different search engine and SEO. In addition, we have outlined factors and improvement techniques that are helpful in both perspective (development & se). In order to evaluate result we have designed tool that is based on SEO methods (on, off page) which will be helpful for website testing. Results attained from our experimental work demonstrates the significance of key SEO factors and this study concluded that WordPress and CMS (content management system) platform is closer to the search engine. Furthermore, if a website is develop through these will achieve.

Keywords: Search Engine Optimization techniques; Website Performance testing; SEO in Ecommerce sites;

1 Introduction

Websites are the ultimate source of spreading business in different regions where physically presence is difficult. Since website provides a rapid way to transport content, ideas and business is a strong challenge which arises to the web performance. Moreover that it never disables or stops working [1]. Approximately more than 80% traffic of users on internet is handled by SE (search engines) [2]. If website is an optimal and follows the guidelines of SEO. As a result getting high ranking and producing better results among others webs. Web performance is measured according to the search engine optimization (SEO) techniques furthermore this is type of research to help in getting better result of web application on search engine [3]. Search engine provides guidance to the web developer to rank website and explain different ways to

¹³⁴⁵Bahria University, Karachi, Pakistan

²PAF Karachi Institute of Economics & Technology, Karachi, Pakistan

develop web application. Therefore, search engine is plays a vital role for web application to get high ranking on search engine [4, 5].

Web Application has been categorized into three ways including (1) web 1.0, (3) 2.0 and last 3.0. [6]. Web 1.0 was developed for reading and information purpose only. It's known as "static" website. These website does not contain any storage and only contains Hypertext Transfer Markup Language (HTML) file base structure. Web 2.0 are designed with the capability of "Storage" and known as "Dynamic" website. These websites are interactive in nature and provide user facility to interact with images and content. Web 3.0 is known as "Semantic web". These web are interactive plus intelligent in nature furthermore these versions are for converting content to a meaningful state [6, 7]. Table I shows a classification of website with the description of functionality and technology used.

Table 1: Categorization of web

Category	Functionality	TechnologyUsed
1.0	Information and Reading purpose	HTML, HTTP, URL
2.0	Database and Interacting web	AIAX, WIKIS AND BLOG
3.0	Sematic or Collaborative sites	3 DIMENSIONAL, XML

The objective of proposed study is to technically justify the importance of different search engines with SEO. Furthermore, we have outlined factors and improvement techniques are helpful in both perspective (Development & SE). In order to evaluate result of this study, we have designed a tool based on SEO methods (On, Off Page) that will be helpful for web developer to test a website.

The remainder of this research paper is organized as follows. Section 2 describes importance of SEO and related work. Section 3 describes the methodology of proposed method. Finally, the 4th section summarizes the concept of this paper.

2 Background And Litration Review

Researchers in various fields have studied search engines and search engine marketing have been studied on various aspects for example designing, behaviour of the client or user, aspects related to marketing, in addition to the influence of politics as well as social association. This part presents an overview of the reported papers on search engine ranking, it also outlines the lack of our knowledge and understanding in this domain. [4,6,8,9,12,13,14,18,20,21,23].

In order to verify the quality of website traffic is measured with the number of unique visitors. Search engine optimization was integral to this process [8] but, as the online market has matured, website switched from a provider of information to a valuable sales channel and the emphasis moved from web traffic to the conversion of such traffic into sales [9, 10]. The extant literature has followed this development with early work studying the technical requirements for website quality [11, 12] or focusing upon how customers are attracted to the website [13]. Generalizing, According to Nguyen the main focus of previously done researches

was on the improvement of client service by the use of tools of marketing, although, researches are needed to study the user behaviour and its control by the use of tools which are utilized for user services, also to control the rate of conversation and therefore its market [14]. As websites evolved into sales channels limited research has investigated the factors that positively influenced the intention to purchase. Examples of this type of work include website visits [15] browsing experience and navigational behaviour web design and language [16] and website aesthetics and length of exposure [17].

Examining conversion rates in the wider context, it has been well documented that traditional retail finds it hard to encourage people to purchase once they are in store, for example, [18] suggest that 54% of visitors leave a retail outlet without making a purchase. While this of concern to high-street retailers, the conversion rate of ecommerce websites (2–4%) is significantly lower [19, 20]. In high-street retail commonly accepted explanations are that many customers, visit several stores before buying or that shopping is simply a recreational activity [21]. These phenomena are particularly relevant in relation to e-commerce websites where the ease and speed with which users can move between websites is often seamless cutting the time and costs required to compare products encouraging such behaviour [22]. However, historically ecommerce also generates issues that relate to trust within the transaction process, with customers concerned about the safety of personal information, credit card data and product [23].

A *E-commerce websites*

E-Commerce refers to the business being generated digitally by mean of network. There is no specific definition of the ecommerce website but different scholars and student define this words according to their understanding. Some definitions are as follows.

According to Dr Allison E-Commerce could be a resource to swap the cost of conversation as well as computing automation. Electronic commerce or E-trade means to share business data, maintenance of trade relations, in addition to the conduction of business trade deals through telecommunication systems. Marketing and trading on World Wide Web is another way to define E-Commerce, it encompasses sales over internet and delivering of goods offline and online (that are digitized), for example a software program [24].

Almost every service and products has its pros and cons attached with it. Some of the advantages of these website are pointed as beneath Ecommerce Site reduced the cost of transaction and moving thing from one place to other however requires a strong Marketing strategy. These sites have 24*7 service availability that may cause security issues and User finds easy methods of selection of products but no guarantee is provided Everyone can develop E-Commerce sites which increase number of sites but on many occasions user get trap in fake business that lead to customer dissatisfaction Novice user also use these sites but developing a site full of simplicity become difficult It requires low cost for becoming operational but lack of development knowledge lead to business loss. It does not require any physical location for starting a business that's why user sometimes does not believe easily. User reaches these sites easily but sometimes user patience breakout due to delivery time and these advantages and disadvantages summarize in figure 1. [5, 23, 24].



Figure 1: Advantages & disadvantages of Ecommerce sites

B Search Engine Optimization SEO

SEO is the technique to increase the productivity of the website and successively high rank in the search engine. Moreover, these techniques guarantee for a high performance. With the help of these techniques website getting high rank [25].SEO is classified into ON page and OFF page SEO. On-page define as it is applied before and on the time of the development. These include heading style of the website, page title, Images format, responsiveness, amount of content and Meta tags. Off page related to the promotion and marketing of the website. These include Social media pages (Facebook, Twitter) sharing, Blogging, Forum and community as shown in Table 2 [25, 26, 27].

Table 2: Describes basic function of the on and off page SEO

On Page	Off-page
Lies inside boundary of the website	Lies outside of the boundary of the website
Needed to optimize at Development of the web	Needed to optimized or worked after development of the web
Images, Titles, Internal Links are the important factors	Important factors includes Social media, Blogging, Forum and communities

3 Methodology

The method proposed in this study is based on web marketing toolto find search engine ranking. This Research is involved a steps similar to waterfall model that heuristic evaluation of website conducted by this study. Our main steps includesgathering data of ecommerce website, Testing with SEO tool, generating results and designing an SEO based tool and finally Results & evaluations. Our proposed methodology is presented in figure 2.

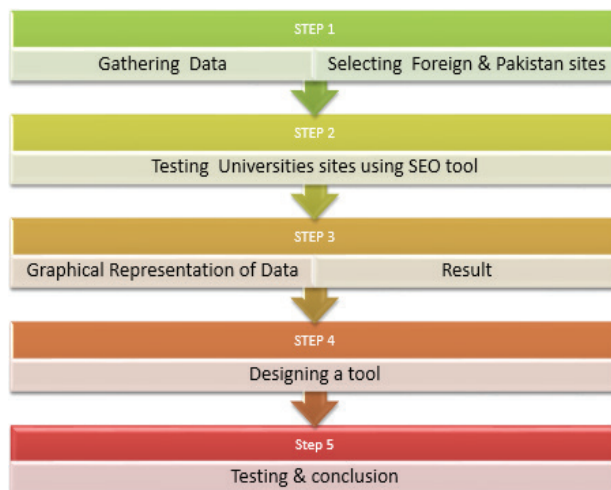


Figure 2: Proposed Method Methodology

There are many platforms for the website development. These start from HTML CSS, ASP, NET and then jump to the popular CMS like WordPress, Joomla, Drupal and Magento. For this presented work, we have selected three different well known platforms for web development such as WordPress, Blogger and HTML5 with CSS. Our proposed tool is an open source and it is simplest for any novice person to use. For easiness of user is to just enter web link on the browser and it also provides a comprehensive report after analysis of web applications. This tool is analysis of website based on seventeen different parameters such as code quality, social interest, internal link, page titles, heading, amount of content and some others as shown in Table 3. There are some limitations of this tool are as follows. user can only test 3 web links at a time and to test other website, they have to wait for some time to analyze other websites. Without signup users are not allowed to access tool.

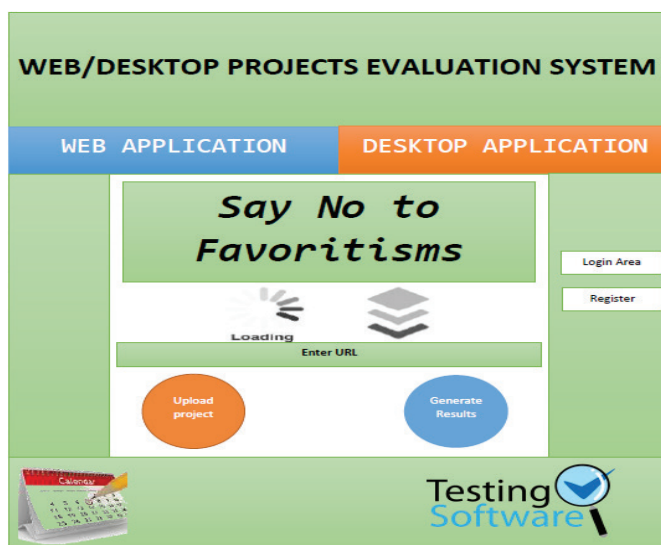


Figure 3: The Architecture of Proposed Tool

The architecture of our proposed method presented in Figure 3. However, analysis has only been confined to the feedback from HTTP which are consumed from the applications that are run verified web application. In order to keep the plan clear accessible, collective and compatible architecture has been used. The proposed approach contains the mentioned aspects: Gathering data of ecommerce site, testing with SEO tool and generating results and designing an SEO based tool and finally results & evaluations Likewise, to the reported strategies commonly found in other systems [6, 11, 14].

4 Experimental Result

For experimental result tool has been developed inPHP with the help of DBMS MySQL. For our studies, we use sample size of 6 website which was selected from renowned organization of national and international of ecommerce website. These website were obtained from Daraz (daraz.com.pk), Yayvo (yayvo.com.pk), HomeShopping (homeshopping.com.pk) amazon (amazon.com), alibaba (alibaba.com) and wallMart (walmart.com). website has been selected.

Table 3: Description of all criterions

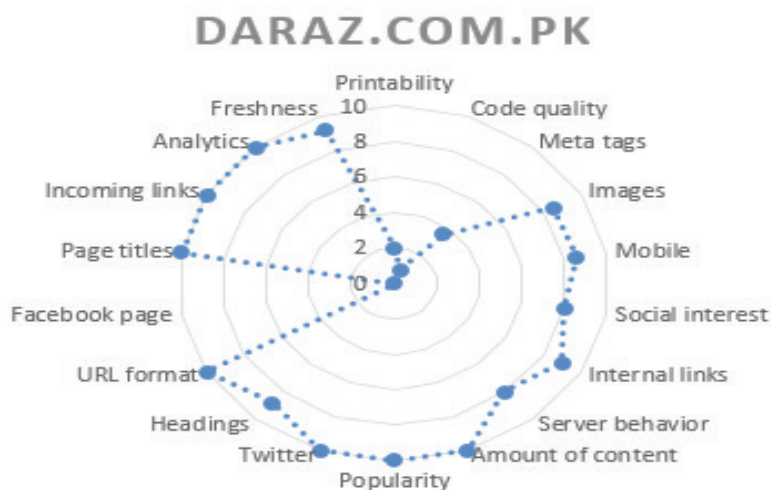
Criterion	Description
Code quality and ASP.	it has shown that in which platform wesite has been developed such as HTML with CSS. PHP
Meta tags	These tags are used to describe what content appear when a visitor finds on the search engine.
Mobile	This parameter describes website responsiveness.
Social interest	Its display, whether user are following the social media website (Facebook, twitter).
Images	what type of images are used in the development of the website such as JPEG, PSD, PNG OR GIF.
Internal links	it suggests website is connected internally and how pages are connected to each other?
Server behaviour	it describes what is the behaviour of the server and how it react when number of user increase.
Amount of content	This parameter describes amount of content is present in the web application.
Popularity	it describes, how many of people are following to access the website.
Twitter	This parameter describes twitter page connect with the website.
Heading	it displays that, what types of heading stylesare used. (H1 to h6).
URL format	it depicts how website is accessed by the visitors on the search engine.
Facebook page	it describes whether website is linked with Facebook page or not.
Page titles	it suggests that how the page title are assigned, what type of style, and font is used.
Incoming links	This parameter describes whether website has any focusing link or not.

Note: These all parameters are important for analyzing any website and these value are never static and change time to time whenever any change occurs to the content (Website, Blog) based on search engine representation, traffic, content and pagerank. The result of six web application presented in Table IV. In order to evaluate result for presented we performed testing on different website to find Search engine ranking. Many attempts have been made in order to aim to SEO Ranking to find accurate result. From the Table overall presented result of each web application Daraz have almost similar result as compared with others in case of code quality is 80%, in case of social interest is 80%, in case of Internal Link is 90%, in case of Page Titles is 100%, in case of Heading is 88%, in case of Amount of content is 100% and So on. In the same way, others sample website almost similar resultAs Daraz parameters result.

Table 4: Describes basic function of the on and off page SEO

Sample Website	C-Q	M-T	T	I-L	F	I	M	S-I	I-L	S-B	AOC	POP	H	URL	FB	P-T
Daraz	0.8	3.6	10	10	9.2	8.5	8.6	8.0	9.0	8.0	10	10	8.8	10	0	10
Yayvo	0.4	10	9.8	10	10	9.9	8.6	8	6.7	8.8	7.7	9.2	7	5.2	10	10
Homeshopping	0.4	10	9.6	10	5.5	9.6	8.6	6.9	9	8.8	10	10	9	7.6	10	10
Amazon	1.9	5.2	0	8.2	0	10	10	8	10	8.5	9.6	10	4	4.8	0	10
AliBaba	0.7	10	10	9.5	10	9.4	6.6	8	10	8.4	8.3	9.8	7	3.6	10	10
WallMart	4.3	5.2	9.3	7.7	10	10	8.6	8	10	8.7	7.7	6.4	5.5	8.8	10	8

AOC: Amount of Content
 P-T: Page Title
 C-Q: Code quality
 POP: Popularity
 M-T: Meta Tags
 T: Twitter
 I: Images
 H: Headings
 I-L: Internal Links
 S-I: Social Interest
 URL: URL Format
 FB: Facebook Page
 F: Freshness
 I-L: Incoming Links



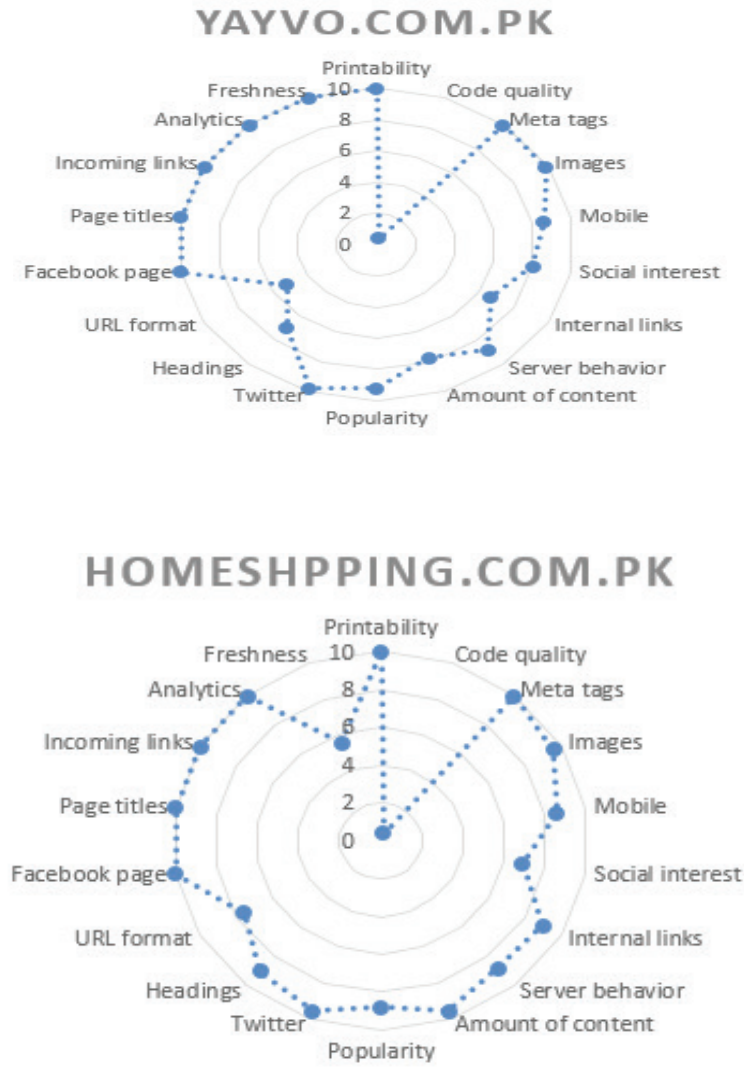


Figure 4: Result of Pakistani Ecommerce Website

Daraz, Yayvo and Homeshopping result of code quality, meta tags, images, mobiles, social interest, internal links, server behaviour, amount of content, popularity, twitter, heading, URL Format, Facebook Page, page titles, incoming links, Analytics and Freshness For Pakistani Website as shown in Figure 4.

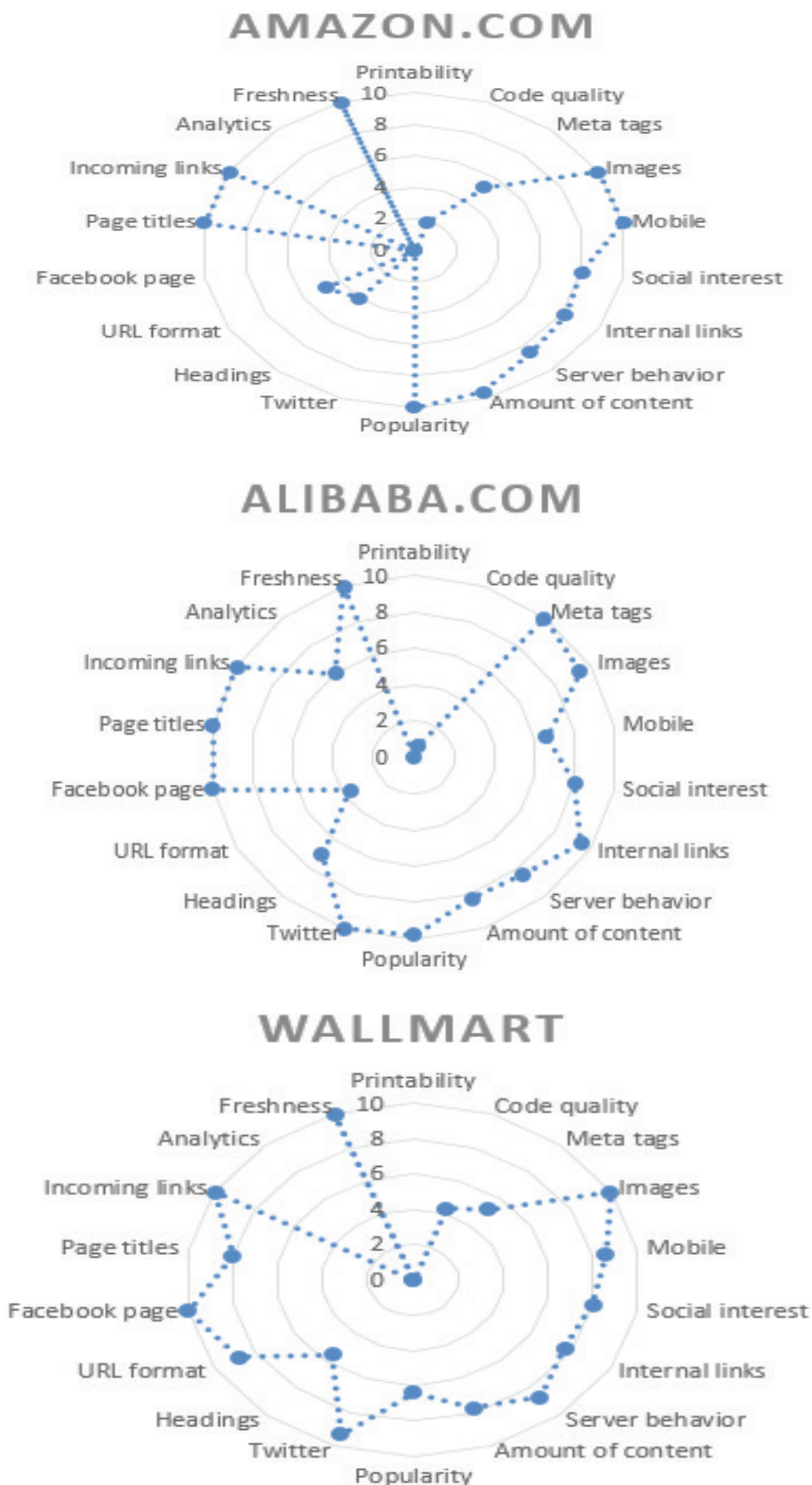


Figure 5: Results of International Ecommerce site

The result of amazon, alibaba and wallmart presented in figure 5. This presented study has found very interesting result to detect Search Engine representation SEO with respect to exiting methods.

5 Conclusion

Search engine optimization (SEO) is a type of research that helps in getting better result of ranking of Google's search engine. Moreover, these search engines are playing an important role for any access, facts on the internet. Furthermore, its importance is enhancing with the passage of time. In this study, we tried to enlighten the most common search engine parameters. Additionally, we have developed a new SEO tool to analyse website. To provide factual results, the experimental work is carried out on proposed tool on 6 renowned websites. There are certain areas where working can improve the performance of website. In order to find efficient result of website where results are less than 20%. Moreover, in result these websites are not effective for SEO in search engine. In the same way, if ecommerce websites meet criteria, more than 90% in all parameters that is effective for search engine. Finally, we conclude the average Pakistani website score is 82% and in the case of foreign website score is 72%. Furthermore, the purpose of this study is to analyse different parameters over website and we conclude that foreign ecommerce websites are less optimize than Pakistan website.

Future research will be based on the development of an upgraded version of tool to increase an accuracy. We will perform comparative study with existence solution. Furthermore, we are also setting up website for users to scan website and download scanner.

References

- [1] Al-Shammari, Eiman Tamah. "Towards Search Engine Optimization: Feedback Collation." *Procedia Computer Science* 62 (2015): 395-402.
- [2] Mondrus, V. L., and D. K. Sizov. "Application of Evolutionary Method of Search Engine Optimization to Calculate Absolutely Flexible Inextensible Thread." *Procedia Engineering* 153 (2016): 496-500.
- [3] Killoran, John B. "How to use search engine optimization techniques to increase website visibility." *IEEE Transactions on Professional communication* 56.1 (2013): 50-66.
- [4] Liu, Yong, Hongxiu Li, and Feng Hu. "Website attributes in urging online impulse purchase: An empirical investigation of consumer perceptions." *Decision Support Systems* 55.3 (2013): 829-837.
- [5] Egri, Gokhan, and Coskun Bayrak. "The Role of Search Engine Optimization on Keeping the User on the Site." *Procedia Computer Science* 36 (2014): 335-342.
- [6] Nath, K., Dhar, S., & Basishtha, S. (2014, February). Web 1.0 to Web 3.0-Evolution of the Web and its various challenges. In *Optimization, Reliability, and Information Technology (ICROIT)*, 2014 International Conference on (pp. 86-89). IEEE.

- [7] Newman, R., Chang, V., Walters, R. J., & Wills, G. B. (2016). Web 2.0 the past and the future. *International Journal of Information Management*, 36(4), 591-598.
- [8] Ayanso, A., & Yoogalingam, R. (2009). Profiling retail web site functionalities and conversion rates: A cluster analysis. *International Journal of Electronic Commerce*, 14(1), 79-114.
- [9] Drew, S. (2003). Strategic uses of e-commerce by SMEs in the east of England. *European Management Journal*, 21(1), 79-88.
- [10] Jelassi, T., Leenen, S., 2003. An e-commerce sales model for manufacturing companies: a conceptual framework and a European example. *Eur. Manag. J.* 21 (1), 38-47.
- [11] Yoo, B., Donthu, N., 2001. Developing and validating a multidimensional consumer-based brand equity scale. *J. Bus. Res.* 52 (1), 1-14.
- [12] Aladwani, A.M., Palvia, P.C., 2002. Developing and validating an instrument for measuring user perceived web quality. *Inf. Manag.* 39 (2), 467-476
- [13] Nguyen, D. H., Leeuw, S., & Dullaert, W. E. (2016). Consumer Behavior and Order Fulfilment in Online Retailing: A Systematic Review. *International Journal of Management Reviews*.
- [14] Moe, W.W., Fader, P.S., 2004. Dynamic conversion behavior at e-commerce sites. *Manag. Sci.* 50 (3), 326-335.
- [15] Sismeiro, C., Bucklin, R.R., 2004. Modeling purchase behavior at an e-commerce Web site: a task completion approach. *J. Mark. Res.* 41 (3), 306-323.
- [16] Hausman, A.V., Siekpe, J.S., 2009. The effect of web interface features on consumer online purchase intentions. *J. Bus. Res.* 62 (1), 5-13.
- [17] Lorenzo-Romero, C., Constantinides, E., & Alarcón-del-Amo, M. D. C. (2013). Web aesthetics effects on user decisions: impact of exposure length on website quality perceptions and buying intentions. *Journal of internet commerce*, 12(1), 76-105.
- [18] Söderlund, M., Berg, H., Ringbo, J., 2014. When the customer has left the store: an examination of the potential for satisfaction rub-off effects and purchase versus no purchase implications. *J. Retail. Consum. Serv.* 21 (4), 529-536.
- [19] Holzwarth, M., Janiszewski, C., Neumann, M.M., 2006. The influence of avatars on online consumer shopping behavior. *J. Mark.* 70 (4), 19-36.
- [20] Sohrabi, B., Mahmoudian, P., Raessi, I., 2012. A framework for improving e-commerce websites' usability using a hybrid genetic algorithm and neural network system. *Neural Comput. Appl.* 21 (5), 1017-1029.
- [21] Gensler, S., Neslin, S.A., Verhoef, P.C., 2017. The Showrooming Phenomenon: it's more than Just About Price. *J. Interact. Mark.* 38, 29-43.

- [22] Frost, D., Goode, S., Hart, D., 2010. Individualist and collectivist factors affecting online repurchase intentions. *Internet Res.* 20 (1), 6–28.
- [23] Kim, Y., Peterson, R.A., 2017. A meta-analysis of online trust relationships in e-commerce. *J. Interact. Mark.* 38 (2), 44–54.
- [24] Di Fatta, D., Patton, D., &Viglia, G. (2018).The determinants of conversion rates in SME e-commerce websites. *Journal of Retailing and Consumer Services*, 41, 161-168.
- [25] D. Viney. *Get to the Top on Google: Search Engine Optimization and Website Promotion Techniques to Get Your Site to the Top of the e Search Engine Rankings and Stay There*, pp. 115-117. Nicholas Brealey Publishing, 2007.
- [26] Knezevi, Boris, and MarijanaVidasBubanja. "Search engine marketing as key factor for generating quality online visitors." *The 33rd International Convention MIPRO*. 2010
- [27] Gupta, Swati, et al. "Search engine optimization: Success factors." *Parallel, Distributed and Grid Computing (PDGC)*, 2016 Fourth International Conference on.IEEE, 2016.

Optical Character Recognition Engine to extract Food-items and Prices from Grocery Receipt Images via Templating and Dictionary-Traversal Technique

Ali Sohani¹ Rafi Ullah² Faraz Ali³ Athaul Rai⁴ Richard Messier⁵

Abstract

This paper proposes a mix of some old and few novel techniques to nail down the fundamental problem of Food-Items and Prices recognition and eventual extraction of them from the Grocery Receipts. Considering in our research we didn't find any existing OCR engine that is up to that standard let alone specialized for this specific purpose. Since the target was to create a specialized OCR system, we began with an idea of creating the wrappers around basic OCR system to empower it with context of Grocery Receipt. For this, we've built pre-function and post-function wrappers over existing system called Tesseract-OCR. Our system follows specific work-flow to enhance basic OCR output. First it runs the provided image to image filters to make it most suitable for Section-level extraction. Our system then bifurcates the image into sections (like Price, Item-Names, Quantity are dealt separately from one another) according to given template layouts. Specific portion of images (sections) are then forwarded to Tesseract engine for basic OCR. Then text-extracted is forwarded to a contextual pattern matcher, to make sense of the text-extracted in a contextual manner. After testing system on particular grocery stores receipts, we successfully conclude that our techniques significantly improve on both the accuracy of overall context based text recognition and close-match detection when compared to an unassisted/ vanilla Tesseract OCR. Proposed system will empower Food-Kitchen Assistance Mobile Apps in the market.

Keywords: Accurate image to text converter, Receipt parsing using template matching, OCR using receipts template, Text retrieval from receipts images

1 Introduction

The objective of our work was grocery receipt's parsing i.e image to text conversion using open source Tesseract OCR [23][4][11]. As Tesseract OCR just retrieve text from images. We have proposed techniques for parsing receipts that is template matching. We stored templates or structures of the templates of different stores. We can easily retrieve items, quantity of each item and price of each item. We used the relative positions of items, quantities and prices in the receipts to find the item etc in new image of the same store. Thus OCR read only that required portion. That reduced the time and improved accuracy.

Before applying simple OCR technique, we first did some image pre-processing techniques. First of all image processing techniques; include image background removal. Background is actually non-textual area of image. Then we applied text deskewing [25] followed by image binarization [1][8]. We also applied resizing technique if image was smaller in size. Items, quantity and

¹²³⁴⁵ Data Science Department, Cubix, Pakistan

¹ali.sohani@cubix.co | ²rafiullah.khan@cubixlabs.com | ³faraz.ali@cubixlabs.com | ⁴athaul.raai@cubixlabs.com

⁵richard@cubix.com

price portions of image were calculated from the template then Tesseract OCR was applied to each portion in order to retrieve the text. Then context sensitive spelling was applied on result.

Rest of the paper includes Related work, Tesseract OCR open source API, Image Pre-processing techniques, Methodology, Image template, Context sensitive spelling correction, Results, conclusion and future work.

2 Related Work

In references [3] and [4] OCRing is done using pattern matching and advance heuristics. These methods are proven to be very successful for generic type of receipts parsing. They have used Regular Expressions to extract different texts. Heuristics in [4] are very helpful to discard garbage data, that are not necessary. Here OCRing is based on assumptions noticed in large number of receipts. These techniques are generic and work on every receipt.

[19] describes the ABBY cloud SDK for receipt recognition. As receipts are not always clear. These images may be noisy due to taken by movable mobile devices. So simple scanning may not give you an accurate results. This API is paid API.

[20] is an R&D about similar purpose. This R&D is basically for receipt parsing. This R&D also include similar steps like image binarization, text finding etc.

OCRDroid framework has been proposed in [8]. This uses image processing techniques like deskewing, binarization etc for better results. There is limitation on multiple images OCRing and Text detection from complex backgrounds.

Beside this a lot of work has been done and going on using OCR techniques, OCR improvement using image processing techniques, using advance state of the art techniques like Neural Networks etc.

3 Tesseract-Ocr

Tesseract is an open source Optical Character Recognition (OCR) Engine or API, available under the Apache 2.0 license. It can be used directly use or using an API to extract typed text, handwritten text or printed text from images of different formats. It supports a wide variety of languages (we have used python) and almost for all operating systems (have used Ubuntu 16.01) [23] [21].

For configuring pytesseract in Ubuntu, use the following commands:

```
sudo pip install pytesseract  
sudo get-apt install tesseract-ocr
```

After configuring it, you can select language, configuration according to your need. We have used 'eng' English as a language, "- psm 6" as a config parameter and Image object as a parameter.

4 Image Processing

Tesseract OCR is open source library sponsored by Google, It has accuracy issue. It is generic image to text converter. To clear the image in order to be read by OCR accurately same image processing steps have been applied as given in [3] and [4].

A Image Background Removal

As OCR process is little bit slower and there is accuracy issue in case of noisy background. Accuracy and speed issues have been improved. This will work only if you have image like as given below. When we applied OCR on below JPEG image having dimensions 3936 x 5248, it took 3.231 seconds. When we applied background removal, it took 1.725 seconds. And the result was also improved as mentioned in the table below:



Figure 1: Walmart receipts before and after image background removal

Table 1: Result Before And After Image Background Removal

, K"	\"
a mar m	a mar 9.4.
Save money. lee better.	Save money. me better.
(504) 522 ~ 4142	(504) 522 - 4142
MANAGER TODD JABBIA	MANAGER TODD JABBIA
1901 TCHOUPITOULAS ST	1901 TCHOUPITOULAS ST
NEW ORLEANS LA 70150	NEW ORLEANS LA 70130
5T2 5022 OP# 00005251 TB# 65 TR# 09552	ST# 5022 OP# 00005251 TE# 83 TR# 09552
15X12 PAS WC 084705715066 7.97 X	15x12 PAS wc 084703715088 7.97 x
HAND TOWEL 066572107026 2.97 X	HAND TOWEL 066572107028 2.97 x
GATORADE 005200055582 F 2.00 X	GATORADE 005200033582 F 2.00 x
GATORRDE 005200055652 F 2.00 X	GATORADE 005200033832 F 2.00 x
OXICLEAN VSR 075705751525 7.52 X	OXICLEAN VSR 075703751523 7.52 x
MTG CAR FL5G 009474657442 9.97 X	MTC CAR FLAG 009474657442 9.97 x
T-SHIRT 088529968450 16.88 X	T-SHIRT 088329968450 16.88 x
PUSH PINS 002775501514 1.24 X	PUSH PINS 002775501314 1.24 x
ULTRATECH 076526451591 5.97 X	ULTRATECH 076326431391 5.97 x
REESE MINI 005400044660 F 2.66 R	REESE MINI 003400044860 F 2.88 R
16 02 CUP 064541654511 0.67 X	16 oz CUP 064541654511 0.87 x
COPY PAPER 005650010265 4.22 X	COPY PAPER 003650010285 4.22 x
SRAGRAHS LQ 006700070070 25.47 T	SEAGRAMS LQ 008700070070 23.47 T
SUBTOTAL 87.96	SUBTOTAL 87.96
TAX 1 9.000 5 7.66	TAX 1 9.000 % 7.66
TAX 2 4.500 % 0.15	TAX 2 4.500 % 0.13
TOTAL 95.75	TOTAL 95.75
AHEX TEND 95.75	AHEX TEND 95.75
ACCOUNT # **** * 289 8	ACCOUNT # **** * 289 3
APPROVAL # 925741	APPROVAL # 923741
REF # 420200465440	REF # 420200485440
Beg Bai Tran Amt End 831	Beg Bal Tran Amt End Bal
CREDIT 254.54 95.75 156.79	CREDIT 234.54 95.75 138.79
TERMINAL # 26005954	TERMINAL # 26003934
07/21/14 10:45:26	07/21/14 10:45:28
CHANGE DUE 0.00	CHANGE DUE 0.00
. # ITEMS SOLD 13	# ITEMS SOLD 13
TC# 7757 9454 7517 6470 6445	TC# 7737 9454 7317 6470 8445
our gnuIrnntcmnd Lew: Pricngs	Our Guaranteed Low Prices
Are Unbeatable with Ad Hatch!	Are Unbeatable with Ad Match!
07/21/14 10:45:24	07/21/14 10:45:28
UUSTOHHR COPY	***CUSTOMER COPY***

B *Image Binarization*

Image binarization is process of converting colored image to black and white image [1] [8]. This is used to clean dirty images i.e images having noisy backgrounds [13]. Tesseract OCR by default use Otsu's Binarization [23]. But we have used this is an extra layer to make results more accurate. And the fact is that we have used images taken by mobile camera which has great chance to be noisy.



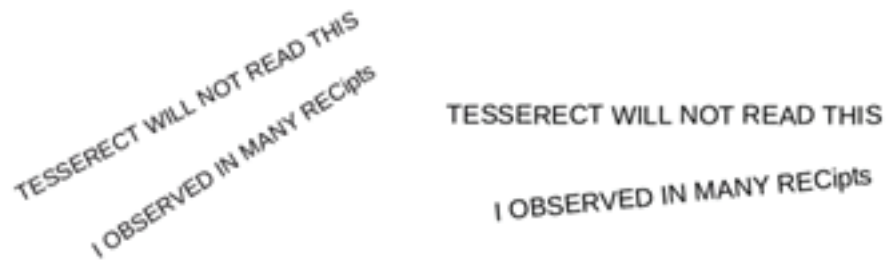
Figure 2: Trader’s joe receipt before and after image binarization

Table 2: Result Before And After Image Binarization

<p>gunman 401- '3 Mmmmm Chicago IL 60611 Store #696 — (312) 951-6369 OPEN 8:00AM TO 10:00PM DAILY OLIVE OIL POTATO CHIPS.. 1.99 HUMMUS GARLIC ROASTED EC 1.99 OHEDDAR NEH ZEALAND SHARP 3.71 PITA NHOLE NHEAT 5" 1.69 OLIVES MANZANILLA 2.29 CREAMY SALTED PEANUT BUTTER 2.49 SUBTOTAL \$14.16 STATE TAX 1 \$0.32 ITAL m1m ITEMS 6 v, Karl 05-31-2015 03:11PM 0696 06 1173 0559 THANK YOU FOR SHOPPING AT TRADER JOE'S www.t.raIJgr'Oe\$Im,,</p>	<p>1111311053 JOE'S I 44 East Ontario Street Chicago IL 50511 Store #596 ' (312) 951-5359 OPEN 8:00AM 10 10:00PM DAILY OLIVE OIL POTATO CHIPS.. 1.99 HUMMUS GARLIC ROASTED EC 1.99 CHEDDAR NEW ZEALAND SHARP 3.71 PITA WHOLE WHEAT 5" 1.89 OLIVES MANZANILLA 2.29 CREAMY SALTED PEANUT BUTTER 2.49 SUBTOTAL \$14.15 STATE TAX 1 \$0.32 TOTAL \$14.48 ITEMS 6 v, Karl 05'31-2015 03:11PM 0695 06 1173 0559 THANK YOU FOR SHOPPING A1 TRADER JOE'S www.trader'oes.com</p>
--	---

C Image and Text Deskewing

Sometimes text in receipts are skewed may be in any direction. In that case, OCR doesn't provide correct results or sometimes doesn't. To avoid such situation we used an additional filter for text deskewing. We have used technique mentioned in [25] for text deskewing. Results have been improved. Following figure (on left) shows the skewed text found in receipt, we deskewed it first and then applied OCR. Accuracy has been improved. Comparison has been shown in table given below.

Figure 3: Skewed text in image**Table 3: Results Before And After Text Deskewing**

<pre><ee~e~eeec< «£va View “As 6 \0%9<NS0 WM“ «awe</pre>	<pre>TESSERECT WILL NOT READ THIS ECipIS IOBSERVED IN MANY R</pre>
---	--

D *Image Resizing*

This technique has been used to speed up the processing. As OCR is a game of playing on pixels so, higher the resolution of image, more will be the processing time and vice versa. So we reduced the size but not too much to effect the OCR accuracy. For example High Definition image of 4000 x 6000 will have the same result as 2000 x 4000 dimension image and processing of image discussed later will be faster than the image discussed earlier. By reducing size, the OCR performed very poor because of information loss in image. Image having DPI (Dots per Inch) greater than 300 has been observed to have good results.

E *Image Stitching*

For long receipts you have two options either you will take photo from far distance or you will take multiple snapshots. For earlier case the quality of image can be disturbed and in the later case, many images should be stitched together first and then OCRed. For this we used image stitching algorithms discussed in [9] [10] [11] and the result was fine. You have also an option to skip image stitching. OCRed multiple snapshots and then clean the result. But this can be very difficult to clean data.

**Figure 4: First part of the image (Upper portion)**



Figure 5: Second part of the receipt (middle portion)

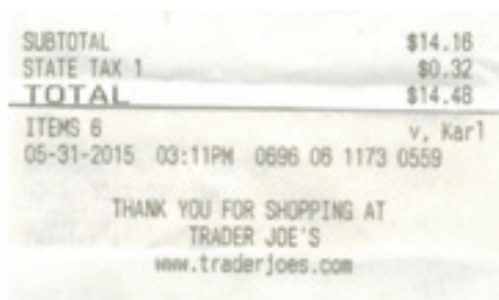


Figure 6: last part of the receipt (Lower portion)

There is no limitation on images. But you must tell us about the base image. So that other images are simply stitched to the base image.

After stitching the result is given below



Figure 7: Image after stitching all three parts

Now this image is simply used as an input to Tesseract OCR and text is retrieved from image. In case of high resolution, this process is very slow. To solve this problem, we first resized all the images to low resolution then stitched them together. But resizing to low resolution will affect the OCR accuracy so, that should be done carefully. We resized then to low resolution first, then stitched and then again resized to high resolution.

5 Image Template

We have used the stored templates of stores in database.

And while testing the image, we retrieved that specific store template. Store template have the (x, y) coordinate points, width and height information of

Footer: Xf, Yf, Wf, Hf

Item: Xi, Yi, Wi, Hi

Quantity: Xq, Yq, Wq, Hq

Price: Xp, Yp, Wp, Hp

Logo: Xl, Yl, Wl, Hl

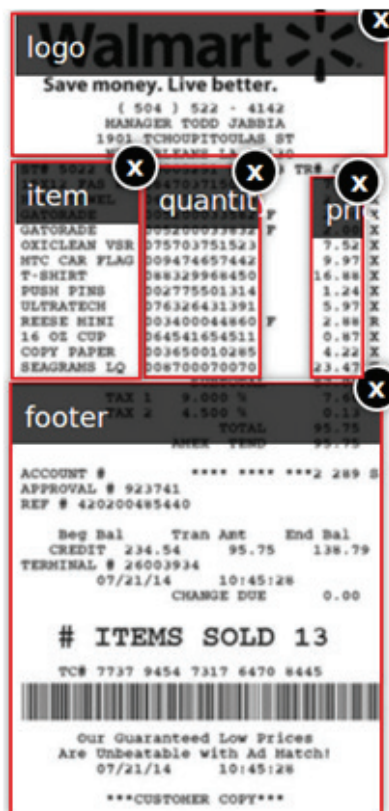


Figure 8: Different regions of image (Image Template)

We have templates of all the receipts in our database as shown in above Figure 8.

It is clear from figure the item portion is between logo and footer. So we used this is heuristic for other images to be tested. We retrieved items portions mathematically.

item x point = item x point

item y point = item y point

item width = item width

item height = footer y point (because item is up to the start of footer)

6 Methodology

Complete methodology is given in the below flow diagram. In case of multiple images, image stitching is applied first then background is removed. Background contains all the non text area of image. Template may be smaller than or larger than the image that is processed. We had the location of items, prices and quantity in the template. We used this knowledge to retrieve image's specific portion. For example we had the image portion having (x, y, w, h) in template.

Where "x", "y" is position in image and "w" is the width of image and "h" is the height of image. We found the percentage of items portion in template image.

Following things are known from template,

- Template Picture Width = W_t
- Template Picture Height = H_t
- Template Picture "x" = X_t
- Template Picture "y" = Y_t
- Template Picture items portion width = W_i
- Template Picture items portion height = H_i

Percentage of items portion width = PW_i

Percentage of items portion height = PH_i

Percentage of items portion X = PX_i

Percentage of items portion Y = PY_i

$$PW_i = (W_i / W_t) * 100 \quad (1)$$

$$PH_i = (H_i / H_t) * 100 \quad (2)$$

$$PX_i = (X_t / W_i) * 100 \quad (3)$$

$$PY_i = (Y_t / H_i) * 100 \quad (4)$$

We have found that what percent of the image is items in the template. Used relative calculation to find items portion in new image.

Set the current picture width and height with respect to template image percentage of width and height

Current image is the image being processed

Current image width = W_c

Current image Height = H_c

Item portion width in new image = IPW

Item portion height in new image = IPH

Item portion X in new image = IPX

Item portion Y in new image = IPY

These values can be calculated by

$$IPW = (W_c / 100) * PW_i \quad (5)$$

$$IPH = (H_c / 100) * PH_i \quad (6)$$

$$IPX = (IP_w / 100) * P_{X_i} \quad (7)$$

$$IPY = (IPH / 100) * P_{Y_i} \quad (8)$$

Using above technique we retrieved the image portions/sections using template information independent of requested image size, whether greater or lesser than template image. This gave us better results.

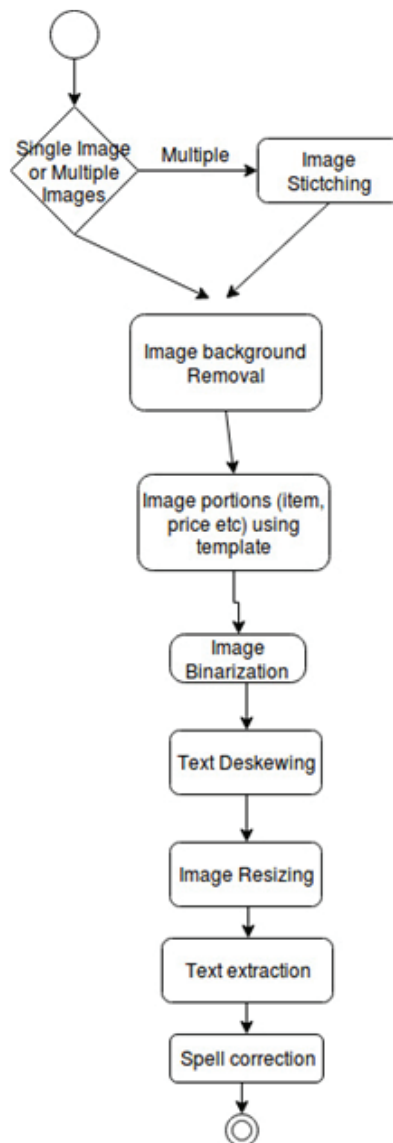


Figure 9: Flow chart of the proposed technique

The next challenge was to read the item and quantity portion if the items in tested image is lesser or greater than items in template image. For example in template image we had 10 items and in tested image we had 2 items. Now portion covered by 10 items will not be the same as portion in image covered by 2 items. If OCR read same portion, result will not be acceptable. To tackle this issue, we used some heuristics, It is clear from the receipts (Figure 7 and Figure 8) that portion between logo and footer is always items portion in receipts.

7 Context Sensitive Spell Correction

To make Tesseract OCR results more accurate, we used context sensitive spelling correction. Context sensitive spelling correction is a technique of correcting OCR results by matching them with dictionaries of stores. For example we scanned walmart Receipt, OCR returned result, we matched the result with walmart dictionary i.e words/terms used in that store. Each stores name terms/words as well as matching score is stored. For the next read, result was automatically corrected. After scanning a lot of receipts, at last this method perfectly worked. Beside this we have used two levels of dictionaries for spell correction, store specific and grocery related dictionary.

Beside this we have used a corpus of text that are part of receipt but not our required text. Words such as tax, total, subtotal, discount etc are included in that corpus. These words are excluded at the very first stage from OCR result.

8 Results

We have tested our algorithm on various receipts template of various sizes and resolution. We observed it is performing best in many cases. We have stored template of walmart having 397 width and 804 height. Tested using walmart image having width 2043, height 4128 and noisy background.



Figure 10: Test Image (Walmart Receipt)

Table 4. Result of Walmart Receipt In Figure 10

Items Name	Prices
15x12 fas wc	7.97
hand towel	2.97
gatorade	2.00
gatorade	2.00
oxiclean vsr	7.52
mtc car flag	9.97
t-shirt	16.88
push pins	1.24
ultratech	5.97
reese mini	2.88
16 oz cup	0.87
copy paper	4.22
seagrams lq	23.47

Given below is the receipt of Trader's joe

**Figure 11: Test Image (Trader's Joes Receipt)**

Table 5: Result Of Trader's Joe Receipt In Fig. 11

Items Name	Price
olive oil potato chips	1.9
hummus garlic roasted ec	1.9
cheddar new zealand sharp	3.71
pita whole wheat 5\	1.8
olives manzanilla	2.2
creamy salted peanut butter	2.4

For both of the above receipts there is no quantity portion, so quantities have been returned empty. If there were quantities in template images, they will be returned.

We have observed that, OCR accuracy has been improved up to great extent using image processing techniques, that was pre-processing step in our proposed technique and then the OCR result has been adjusted using our post-processing technique that is spell correction of OCRed Text.

9 Conclusion

A In this paper we worked on template based matching of receipt and retrieving text from receipt images. For the new image (receipt) first we retrieved the structure of receipt and then calculated the different portions of images i.e items portion, price portion and then parsed the new receipt accordingly. Retrieved text is then cleaned by filtering (context sensitive spelling correction). We have tested system for 10 different stores receipts and our proposed template based matching and parsing gave good results. It also worked best in case of noisy images.

10 Future Work

We have shown this idea seems good in case of noisy and complex receipts. Future work is to do generic receipt parsing and making template based matching efficient. Because receipts may vary in content from time to time. On some occasions receipts may have discount portion while in normal situation it may be simple containing only items, prices and quantities. Generic receipt parsing works without template. Our proposed system cannot parse receipts whose templates are not available in our database. We will also do receipt recognition using Machine Learning techniques and then parse it.

Acknowledgments

I am a strong believer of the fact, that a man needs a solid team in surrounding to achieve great results. Thanks to my team, who has worked with me to follow the vision and lead, also for them to bear my many requests of making different suggested updates in algorithm in pursuit of accuracy and performance. I am very thankful for their respect and hard-work. Also thanks to the organization Cubix, that gave us an opportunity to work on most pragmatic-level industry-grade problems in the field of our passion (Data-Science).

References

- [1] Chaki, Nabendu, Soharab Hossain Shaikh, and Khalid Saeed. "A comprehensive survey on image binarization techniques." In *Exploring Image Binarization Techniques*, pp. 5-15. Springer India, 2014.
- [2] Troller, Milan. "Practical OCR system based on state of art neural networks." (2017).
- [3] Rafi, Ali, Faraz, Athaul "OCR Engine to extract Food-items and Prices from Receipt images via pattern matching and heuristics approach" In *International Conference of Computing and Related Technologies*, December 2017, SMIU, Karachi, Pakistan
- [4] Rafi, Ali, Faraz, Athaul "OCR Engine to Extract Food-items, Prices, Quantity, Units from Receipt Images, Heuristics Rules Based Approach" in *IJSER Volume 9, Issue 2, February 2018* (accepted)
- [5] Stadermann, Jan, Denis Jager, and Uri Zernik. "Hierarchical Information Extraction Using Document Segmentation and Optical Character Recognition Correction." U.S. Patent Application 15/620,733, filed September 28, 2017.
- [6] Modi, Hiral, and M. C. Parikh. "A review on optical character recognition techniques." *Int J Comput Appl* 160, no. 6 (2017): 20-24.
- [7] Oudah, Nabeel, Maher Faik Esmail, and Estabraq Abdulredaa. "Optical Character Recognition Using Active Contour Segmentation." *Journal of Engineering* 24, no. 1 (2018): 146-158.
- [8] Zhang, Mi, Anand Joshi, Ritesh Kadmwala, Karthik Dantu, Sameera Poduri, and Gaurav S. Sukhatme. "OCRdroid: A Framework to Digitize Text Using Mobile Phones." In *MobiCASE*, pp. 273-292. 2009.
- [9] Kumar, Asit, and Sumit Gupta. "Detection and recognition of text from image using contrast and edge enhanced mser segmentation and ocr." *IJOSCIENCE (INTERNATIONAL JOURNAL ONLINE OF SCIENCE)* Impact Factor 3, no. 3 (2017): 3.
- [10] Farahmand, Atena, Hossein Sarrafzadeh, and Jamshid Shanbehzadeh. "Noise removal and binarization of scanned document images using clustering of features." (2017).
- [11] Wang, Fu-Bin, Paul Tu, Chen Wu, Lei Chen, and Ding Feng. "Multi-image mosaic with SIFT and vision measurement for microscale structures processed by femtosecond laser." *Optics and Lasers in Engineering* 100 (2018): 124-130.
- [12] Zhang, Jing, Guangxue Chen, and Zhaoyang Jia. "An image stitching algorithm based on histogram matching and SIFT algorithm." *International Journal of Pattern Recognition and Artificial Intelligence* 31, no. 04 (2017): 1754006.
- [13] ZHAO, Yan, Yue CHEN, and Shi-gang WANG. "Corrected fast SIFT image stitching method by combining projection error." *Optics and Precision Engineering* 6 (2017): 029.

- [14] Sharma, Manoj, Anupama Ray, Santanu Chaudhury, and Brejesh Lall. "A Noise-Resilient Super-Resolution framework to boost OCR performance." In Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on, vol. 1, pp. 466-471. IEEE, 2017.
- [15] Brisinello, Matteo, Ratko Grbić, Matija Pul, and Tihomir Anđelić. "Improving Optical Character Recognition Performance for Low Quality Images." In 59th International Symposium ELMAR-2017. 2017.
- [16] Patel, Amit, Burra Sukumar, and Chakravarthy Bhagvati. "SVM with Inverse Fringe as Feature for Improving Accuracy of Telugu OCR Systems." In Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, pp. 253-263. Springer, Singapore, 2018.
- [17] GOCR - A Free Optical Character Recognition Program. <http://jocr.sourceforge.net/>.
- [18] OCR resources (OCROPUS). <http://sites.google.com/site/ocropus/ocr-resources>
- [19] OCRAD - The GNU OCR. <http://www.gnu.org/software/ocrad/>.
- [20] Simple OCR - Optical Character Recognition. <http://www.simpleocr.com/>.
- [21] <https://ocrsdk.com/documentation/quick-start/receipt-recognition/>
- [22] <http://rnd.azoft.com/applying-ocr-technology-receipt-recognition/>
- [23] Tesseract OCR Engine. <http://code.google.com/p/tesseract-ocr/>
- [24] <http://opencv-python-tutorials> last visited 10-Oct-2017
- [25] <https://github.com/tesseract-ocr> last visited 6-Oct-2017
- [26] <http://pyimagesearch.com> last visited 9-Oct-2017
- [27] All images are from <http://google.com>

Ontology Based System for Expert Searching in Academia using SWRL and SPARQL

Furqan Hussain Essani¹ Quratulain Rajput²

Abstract

Searching an expert with relevant experience and expertise is an important and a challenging task in academics. A lot of work has been carried out in this regard, however the semantic web technologies for modelling the information to search an expert is not being explored extensively. This paper proposed an ontology based system to search an academic expert of a particular field of study. The system comprised of the ontology which consists of an academic contribution of an individual. Additionally, SWRL (Semantic Web Rule Language) rules were created based on the academic contribution as publications made by an individual to infer their field of expertise. Finally, the SPARQL queries were performed to search an expert. This research developed a tool to experiment the proposed system and used IEEE explore Digital Library to retrieve academic contribution of an individual. The ontology based system of an expert searching is foundan efficient in terms of reducing the data modelling cost and making the system easily extendable and reusable for other applications.

Keywords: Expert searching, Ontology, SPARQL, SWRL.

1 Introduction

Expert finding is a challenging problem that has practical applications in many fields. Academia is one such domain where the question of finding an expert on a particular topic frequently arises. The expertise of an individual in academia is primarily judged by the individual's contribution in the form of his/her published work. Due the explosive growth of information over web made it extremely difficult to search an expert from unstructured and heterogeneous web sources. Therefore, several platforms have already been introduced to provide experts information such platforms are DBLP³, CiteSeer⁴, ACM digital Library⁵, IEEE explore⁶, Google Scholar⁷, and so on. In contrast to traditional keyword based searching from unstructured sources these platforms provide structured representation of information of academic contribution of an individual. However, still, none of these platforms used semantic technologies to represent semantic information for searching experts.

In the recent years, the semantic web technologies have emerged as a much needed platform that provide structured information with semantically rich data model. Moreover, semantic technologies overcome the problem of data integration as well as provide easily scalable and reusable data modelling technique[1]. Ontologies as data model is one of the key

^{1,2} *Institute of Business Administration. Karachi, Pakistan*

¹*fessani@iba.edu.pk* | ²*qrajput@iba.edu.pk*

³*www.dblp.com*

⁴*www.citeseer.com*

⁵*www.acm.org*

⁶*www.ieee.org*

⁷*Scholar.google.com*

components of semantic technologies. RDF, RDFS and OWL are Ontology languages from low high semantic expressivity respectively. In ontology, information is represented as statement that consists of a subject, object and predicate. The collection of these statements makes data model that has potential to directly apply logical inference in the data model to infer new knowledge without any pre-processing task.

The task of expert finding gained significant attention and a lot of work has been suggested in this regard after the inclusion of expert search task in the TREC Enterprise track[2]. The objective of expert finding task is, given a query, to find out a ranked list of experts. Some of the closely related works include fine-grained expert search model by Bao[3], topic-based and language hybrid model by Deng[4] and ranking workgroup members using citation analysis by Bogers[5]. Similarly social networks have also been used for expert searching for which notable work includes ArnetMiner[6].

All the above mentioned work is focused on the use of probabilistic approach and revolves around the traditional text mining techniques. However, not much attention is given to the semantics of the information. This paper proposed ontology based system for searching experts that can be integrated and reused across different applications built using semantic technologies.

The rest of the paper is organized as follows. Section II provides an overview of the related work while Section III describes the proposed methodology of ontology based system for searching experts and section IV describes the application to understand the potential of the presented work. Finally Section V, concludes the paper and provides future research directions.

2 Related Work

The problem of expert searching in academics has been addressed with different perspectives. Most of the work carried out in this regard use traditional probabilistic based approach. Deng[4] in his work suggested three different models for achieving the task. One of these models assigns a prior probability to a document to signify its importance and impact. The document prior probability is used along with the query-topic relevancy to rank a particular expert. In his other model he considered the fact that an expert may have expertise in various fields and hence ranking of expert is based on the aggregate expertise on various topics. His last model uses a hybrid approach by combining the two models.

Another probabilistic based work, carried out by Bao[3] proposes a more specific evidence oriented model. The model extracts evidence when a topic and a person with a specific relation are found in a given document. The evidence is then evaluated on different measures and each collected measure is then used to score and rank an expert.

The work carried out by Mangaravite in [7] to find expert in academia suggests modelling the document-person association as an estimate rather than a Boolean variable. The suggested method introduces probability of a document being informative of the expertise of the author.

A seminal work that uses semantic approach is presented by Nazimuddin[8]. In his work he has suggested indexing of academic information from ontological perspective. The approach

combines academic ontology with the academic social network and finds the score of an expert based on his/her academic contribution and relationships with other experts.

However this model, like others, uses feature vector representation of an academic topic and cosine similarity measure to calculate similarity between the given query and the topic. In this suggested work, this task is simplified by representing the topic with a set of keywords and writing Semantic Web Rule Language (SWRL) rules to identify the similarity [6]. This suggested model is applied to find the experts in different fields of Computer Science.

The work presented in this paper is similar to the Nazim Uddin's work [5] in the use of ontology, however, in contrast to calculate similarity between topics, the novelty of our approach lies in the publications retrieval mechanisms from IEEE explore. Further the set of keywords in each publication were used to create SWRL rules to infer the experts of the field.

3 Ontology Based System For Experts Searching

In this section we have discussed the research that provides ontology based system for expert's searching. The system utilized ontology that incorporated academic information as well as SWRL rules to infer experts from academic information. The main components of the presented system as shown in Fig. 1 comprised of the following three major steps:

- A. Ontology construction for academic information
- B. Ontology reasoning
- C. Ontology Querying

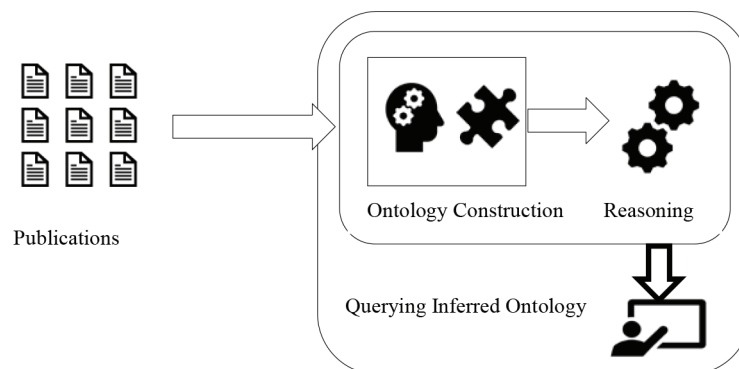


Figure 1: Ontology based System for Expert Searching

A *Ontology Construction for academic information*

In this component we semantically stored all the academic information of an individual. The research publications of an individual could be utilized to categorize the person's expertise into different fields. For this purpose, we suggested to retrieve the set of publications given fields as keyword on the available online digital libraries. Then each of the retrieved publication associated with that field in the ontology and the author of the publication will be considered an expert of the field. This ontology model consists of several classes, object properties, data type properties, and domain range restrictions as discussed below:

Classes

In order to semantically store the academic information from publication's details, there was a need to specify several classes in ontology those are commonly used in the publication.

Publication: This class includes the set of published articles in the different subject areas of computer science.

Journal: A journal is a type of publication. Therefore, it is a subclass of Publication.

Conference Proceeding: A conference proceeding is a type of a publication. It is a subclass of Publication.

Book: A book is a type of publication. It is a subclass of Publication.

Researcher: This class includes the set of people who have made contribution in the subject areas through their published work.

Field: This class includes set of different subject areas in computer science domain.

Computer Vision: Computer Vision is one of the fields of Computer Science.

NLP: Natural Language Processing is one of the fields of Computer Science.

Image Processing: Image Processing is one of the fields of Computer Science.

Similarly, other subject areas with their sub-files can be added to extend the ontology knowledge base.

Object properties

The relations (object properties) in the ontology describe how the classes and their individual members are related to each other. In this case we have defined five different relations as explained below and also shown in Table 1.

hasFirstAuthor: A publication (p) has a first author whose contribution is more significant than the other contributing authors

isFirstAuthorOf: A researcher who has contributed as a first author to a publication (p). It is an inverse relation of hasFirstAuthor

isCoAuthoredBy: A publication (p) may have contributing authors besides the first author

hasCoAuthored: A researcher who has contributed as a co-author to a publication (p). It is an inverse relation of is Co Authored By

Include: This relation is used to identify all the publications relevant to a particular subject area.

Table 1: Ontology Properties

	Property Name	Domain	Range
Object Properties	hasFirstAuthor	Publication	Researcher
	isFirstAuthorOf	Researcher	Publication
	isCoAuthoredBy	Publication	Researcher
	hasCoAuthored	Researcher	Publication
	include	Field	Publication

Data type properties

Data type properties describe more semantic information as literal that is associated with the particular class in the ontology. Following are some data type properties described in ontology. Table 2 lists the data type properties with domain range restrictions.

hasTitle: Each publication has a title, so it is stored in the ontology. The title is stored as a string.

hasKeyword: Each publication is associated with set of keywords. Each keyword is stored as a string.

hasLastName: Every researcher has a last name. The last name is stored as a string.

hasAffiliation: A researcher has an affiliation with any academic or professional organization. This data property is stored as a string.

hasEmail: A researcher has an email address which is used for correspondence over internet. This data property is stored as a string.

Table 2: Data type Properties

	Property Name	Domain	Range
Data Properties	hasTitle	Publication	String
	hasKeyword	Publication	String
	hasLastName	Researcher	String
	hasAffiliation	Researcher	String
	hasEmail	Researcher	String

In the presented work we considered three fields of computer science that include Computer Vision, Natural Language Processing (NLP) and Image Processing. Instead of manual tagging of publication with the field we suggested retrieving the publications from IEEE explore based on the provided field. For example, we searched IEEE explore for the list of publications by providing NLP keyword. Thus, it is assumed that this list of publications associated with the field NLP.

Further the association of field with publication also introduced the association of set of keywords in publication with the field. This would help to describe the set of keywords related to particular field.

The ontology used to create the data store is shown in Fig. 2. In this figure the circles represent the classes in the ontology, the arcs connecting those circles represent the relations between these classes and the rectangles represent the literal values of the classes.

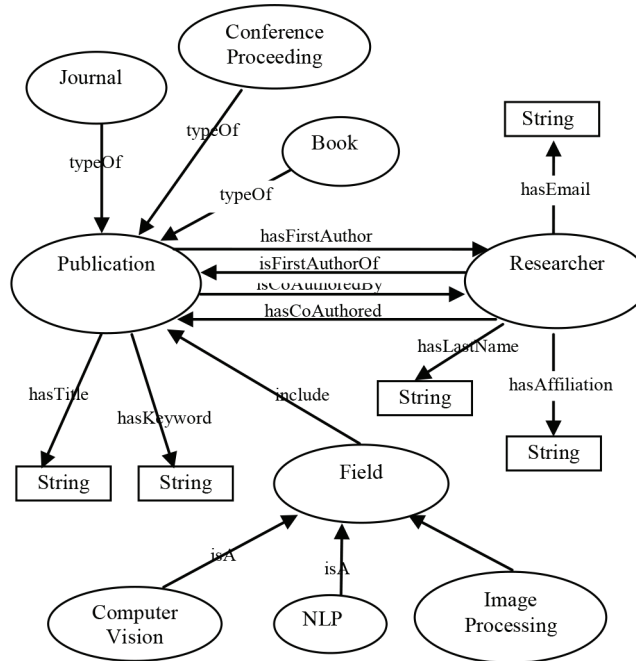


Figure 2: Ontology Diagram

In this approach there is a possibility that one keyword may be represented in more than one field of Computer Science. For example, as shown in Table 2, the keyword, “feature extraction” is being used to represent all the three fields under our consideration.

The set of keywords for different fields will form an overlapping set as shown in Fig. 3. There will be some keywords that will represent only a single field, some keywords that will be common to any two fields, while some will be common to all three fields.

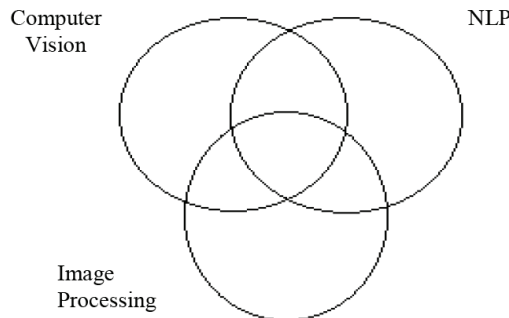


Figure 3: Overlapping set of keywords associated with fields

Table 3: Selected Keywords representing fields in Computer Science

	Field of Computer Vision	Field of NLP	Field of Image Processing
Keywords	Feature extraction	Feature extraction	Feature extraction
	Computer vision	Sentence completion	Accuracy
	Image block representation	Lexical disambiguation	Biomedical imaging
	Pattern recognition	Ngram	Breast cancer
	Vision development	Opinion mining	Image segmentation
	Vision analysis	Recommendation systems	SVM
	HCI	User comments	Fuzzy logic
	Image analysis	Machines learning	Signal processing

B Rules Construction and Ontology Reasoning

Rules help in deducing knowledge that is present in the data set but cannot be expressed through the ontology language. The Semantic Web Rule Language (SWRL) is used to construct the rules which are built on top of the ontology. Using the SWRL rules, a publication is assigned to belong to a particular field of Computer Science based on the keywords. Depending on the scope of the publication, one publication may belong to different fields at the same time.

For example, a publication with the keyword “Feature extraction” or “HCI” or “Pattern recognition” or “Vision analysis” would belong to the field of Computer Vision in our data store. Any publication with keyword “Ngram” or “Opinion Mining” or “Machine learning” would belong to the field of Natural Language Processing. And a publication with the keyword “Accuracy” or “SVM” or “Fuzzy logic” would belong to the field of Image Processing.

The set of rules for these assignments is given in table 3. The corresponding rules in SWRL rules are mentioned in table 4. Similar rules can be written for other keywords forming the set representing different fields.

Table 4: Rules to assign publication and it's keyword to a field

Rule No.	Rule Description
1	A publication belongs to the field of Computer Vision if it has a keyword of “Feature extraction”
2	A publication belongs to the field of Computer Vision if it has a keyword of “HCI”
3	A publication belongs to the field of Computer Vision if it has a keyword of “Computer vision”
4	A publication belongs to the field of Computer Vision if it has a keyword of “Vision analysis”
5	A publication belongs to the field of NLP if it has a keyword of “Feature extraction”
6	A publication belongs to the field of NLP if it has a keyword of “Opinion mining”
7	A publication belongs to the field of NLP if it has a keyword of “Machine learning”
8	A publication belongs to the field of Image Processing if it has a keyword of “Accuracy”
9	A publication belongs to the field of Image Processing if it has a keyword of “SVM”
10	A publication belongs to the field of Image Processing if it has a keyword of “Fuzzy logic”

C *Ontology Querying*

The query component is used to search the expert from the ontology data store. In this component we have used SPARQL (Semantic Protocol and RDF Query Language) queries to retrieve the data from the data store [7]. Discussed in detail in the experiment section.

4 Experiment

A *Data Collection and instance population in ontology*

There are various academic databases available that store articles which are published in journals and other different repositories. In this paper, we have restricted ourselves to only those academic publications that belong to the area of computer science. Some well-known academic data bases in the area of computer science include the ACM Digital Library from ACM⁸, the CiteSeerX provided by the Pennsylvania State University⁹, the DBLP from University of Trier¹⁰ and the IEEE explore Digital Library by IEEE¹¹.

For proof of concept, we developed a tool to create the data store. We have used publications retrieved from the IEEE explore Digital Library. IEEE explore Digital library is one of the largest resources in the computer science domain with nearly 3.7 million entries as of now. The information extracted from each publication includes names of authors, affiliation, emails and the set of IEEE keywords associated with the publication.

As seen in Figure 4 we have added instances of different publications few of them listed here, namely p1, p2, and p9. The instance p1 has first author a1 and has two keywords of “computer vision” and “feature extraction”

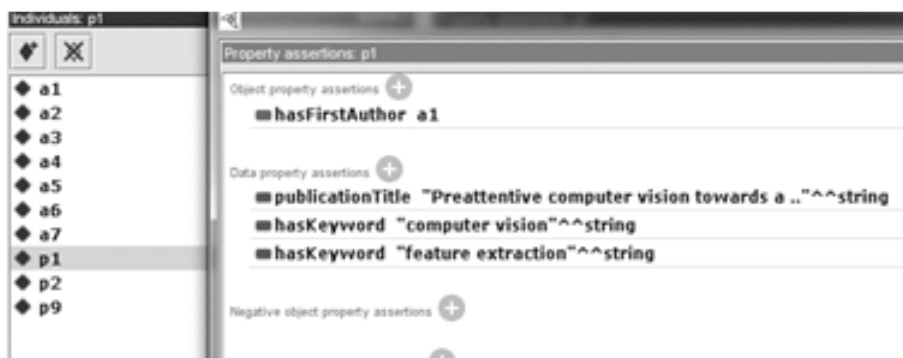


Figure 4: Instances of Publications

B *Reasoning with SWRL rules*

After populating the instances in the ontology with various publications details, the SWRL rules given in Table 4 were used to infer their membership to different fields. Since the set of

⁸<http://dl.acm.org/>

⁹<http://citeseerx.ist.psu.edu>

¹⁰<http://dblp.uni-trier.de>

¹¹<http://ieeexplore.ieee.org>

keywords describing the fields were overlapping, one publication may belong to more than one field at one time.

As seen in Figure 5 and in Figure 6 the publication p1 was assigned to the fields of Computer Vision and NLP based on rules 3 and rule 6 given in table 4.

Table 4: SWRL Rules to assign publication to a field

Rule No.	Rule Description
1	Publication(?p), hasKeyword(?p, "Feature extraction")→ComputerVision(?p)
2	Publication(?p), hasKeyword(?p, "HCI")→ComputerVision(k2?p)
3	Publication(?p), hasKeyword(?p, "Computer vision")→ComputerVision(?p)
4	Publication(?p), hasKeyword(?p, "Vision analysis")→ComputerVision(?p)
5	Publication(?p), hasKeyword(?p, "Feature extraction")→ NLP(?p)
6	Publication(?p), hasKeyword(?p, "Opinion mining")→ NLP(?p)
7	Publication(?p), hasKeyword(?p, "Machine learning")→NLP(?p)
8	Publication(?p), hasKeyword(?p, "Accuracy")→ImageProcessing(?p)
9	Publication(?p), hasKeyword(?p, "SVM")→ImageProcessing(?p)
10	Publication(?p), hasKeyword(?p, "Fuzzy logic")→ImageProcessing(?p)

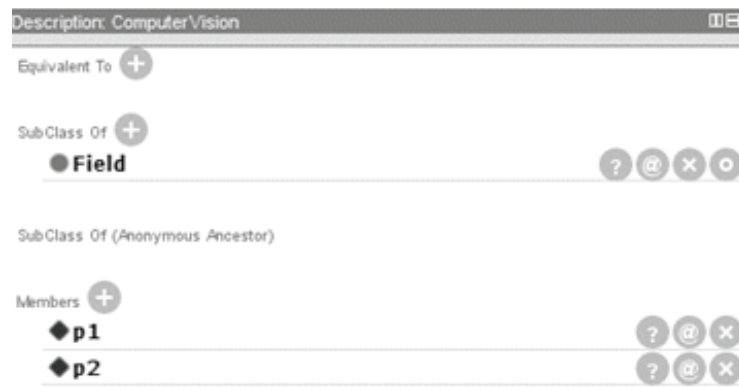


Figure 5: Publication p1 inferred to belong to field of Computer Vision

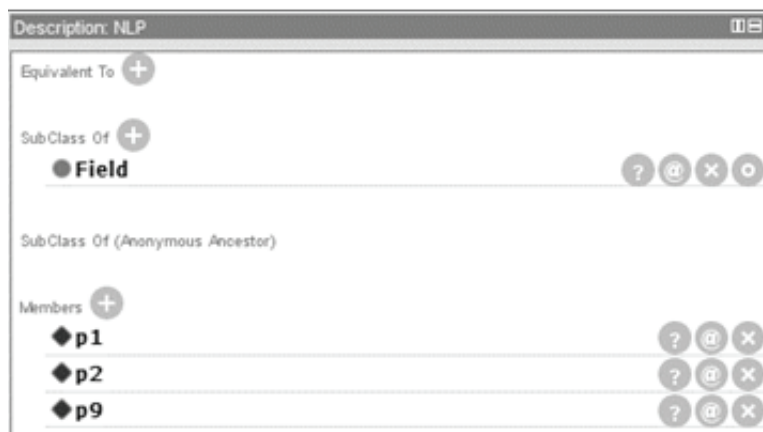
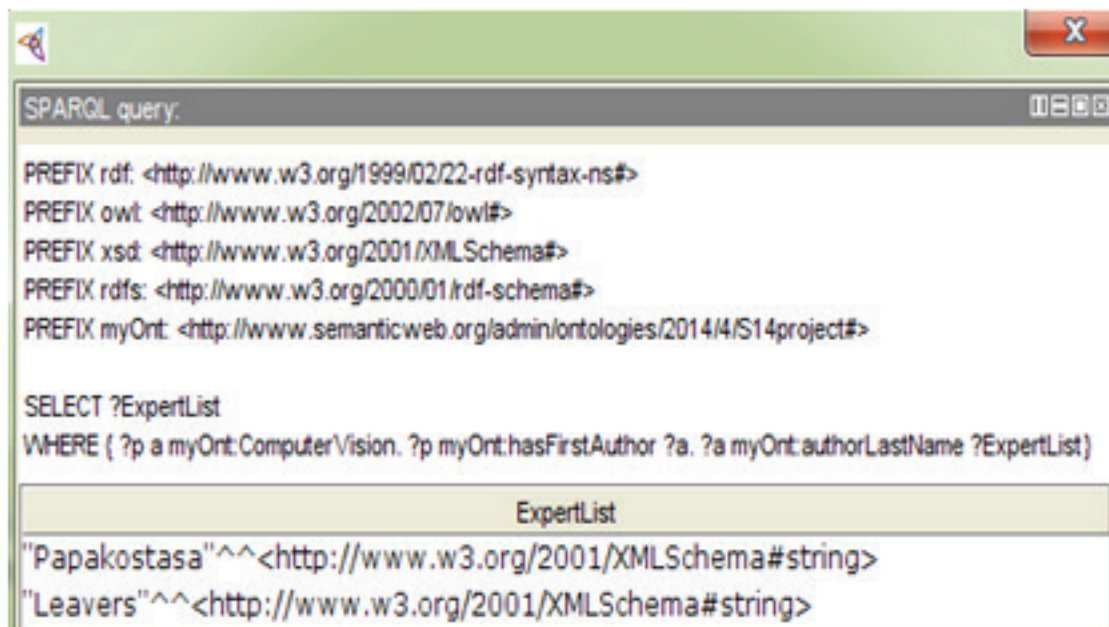


Figure 6: Publication p1 inferred to belong to the field of NLP

C Querying for Expert Searching

The experts were retrieved according to their contributions as a first author of the publication in each field of computer science. For retrieval purpose, SPARQL query was used which first retrieves all the publications depending on the user given query field and consider their first author as experts in that field.

In Figure 7 we have made query to get the list of experts in the field of Computer Vision. The query first evaluates all the publications for the given field and then lists down the name of first author for each publication. Along with the author name, other details like the affiliation and the email address can also be retrieved by extending the same query.



```

SPARQL query:
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX myOnt: <http://www.semanticweb.org/admin/ontologies/2014/4/S14project#>

SELECT ?ExpertList
WHERE { ?p a myOnt:ComputerVision. ?p myOnt:hasFirstAuthor ?a. ?a myOnt:authorLastName ?ExpertList}

ExpertList
"Papakostas"^^<http://www.w3.org/2001/XMLSchema#string>
"Leavers"^^<http://www.w3.org/2001/XMLSchema#string>

```

Figure7: Query to list experts in the field of Computer Vision

5 Conclusion

The expert searching model proposed in this paper used ontology based data model that is well defined, structured and semantically rich academic information. Publications data retrieved from IEEE explore Digital Library for the experiment potential of proposed approach. The different fields in the area of computer science are represented using a set of keywords and the publications were assigned to an individual field and used SWRL using publication to infer expert of the field. The expert in a field is considered to be a researcher who has a contributed work as a first author. Based on the user query a list of experts was generated using the publications that belong to the field that has been queried. This research, further extends in future by ranking the authors in a particular field.

References

- [1] N. Shadbolt, T. Berners-Lee, and W. Hall, "The Semantic Web Revisited," *IEEE Intell. Syst.*, vol. 21, no. 3, pp. 96–101, Jan. 2006.
- [2] "Text REtrieval Conference (TREC) Home Page." [Online]. Available: <http://trec.nist.gov/>. [Accessed: 25-Dec-2017].
- [3] S. Bao, H. Duan, Q. Zhou, M. Xiong, Y. Cao, and Y. Yu, "A Probabilistic Model for Fine-Grained Expert Search.," in *ACL*, 2008, pp. 914–922.
- [4] H. Deng, I. King, and M. R. Lyu, "Formal models for expert finding on DBLP bibliography data," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, 2008, pp. 163–172.
- [5] T. Bogers, K. Kox, and A. van den Bosch, "Using citation analysis for finding experts in workgroups," in *Proc. DIR*, 2008, pp. 21–28.
- [6] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 990–998.
- [7] Vitor Mangaravite and Rodrygo L. T. Santos, "On Information-Theoretic Document-Person Associations for Expert Search in Academia," presented at the *SIGIR '16*, Pisa, Italy, 2016.
- [8] T. H. D. Mohammed Nazim Uddin, "EXPERTS SEARCH AND RANK WITH SOCIAL NETWORK: AN ONTOLOGY-BASED APPROACH," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 23, no. 01, 2013.



PAF-KIET 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>



KARACHI INSTITUTE OF ECONOMICS AND TECHNOLOGY

PUBLISHED BY:

College of Computing and Information
Sciences, PAF-KIET.

www.cocis.pafkiet.edu.pk

www.kjcis.pafkiet.edu.pk

kjcis@pafkiet.edu.pk

MAIN CAMPUS

PAF BASE Korangi Creek,
Karachi-75190.

Ph: (9221) 35091114-7

Fax: (9221) 35091118

CITY CAMPUS

28-D, Block 6, PE.C.H.S,
Karachi-75400.

Ph: (9221) 34543280

Fax: (9221) 34383819

NORTH NAZIMABAD CAMPUS

F-98, Block B, (Near KDA roundabout)
North Nazimabad, Karachi-74700.

Ph: (9221) 36628351 or 36679314

Cell: 0336-2444191-92