Graduate Student Day Workshop

National Conference on Communications (NCC) 2019

21 February 2019
Indian Institute Of Science
Bangalore, India

Contents

Deep CNN-based Semantic Multinomial Representation for Scene Recognition

Shikha Gupta, Ph.D. Scholar (IIT Mandi)

***Abstract*—Describing visual semantic content of images is an effective and direct way to improve scene image recognition. Semantic multinomial (SMN) representation is one such representation that captures semantic information using posterior probabilities of concepts. The core part of obtaining SMN representation is building the concept models. For building the concept model, it is necessary to have ground truth (true) concept labels for every concept present in an image. Manual labeling of concepts is practically not feasible due to a large number of images in the dataset. In this work, we propose an approach for selecting pseudo-concepts in the absence of true concepts labels. We propose to generate a novel deep CNN-based SMN representation using weakly supervised pseudo-concept modeling. In this approach, activation maps (filter responses) from deeper convolutional layers are considered as the cue for pseudo-concepts. We propose to use subspace analysis of data of pseudo-concept classes to group the similar pseudo-concepts. The effectiveness of the proposed approach is studied for scene recognition tasks on standard datasets like MIT67 and SUN397.**

I. Introduction and Motivation

Scene recognition is one of the onerous problem in visual recognition tasks as scene is a human-scaled view of real world environment comprising of multiple entities called concepts (like car, sky, tree, etc.), that are arranged in a spatially correlated manner. Successful recognition methods need to generate a powerful descriptive and discriminative visual representation of images to deal with issues such as, high intra-class and low inter-class variability, complex semantic structures, large overlapping of concepts among classes, and varying size, shape, aspect ratio of same semantic concept across dataset.

Over the past few years, many of the research communities have addressed these issues and proposed a variety of image representations such as hand-crafted local image representations (SIFT and HOG) and trained CNN-based representations. However, these representations fail to describe semantic concepts present in the image effectively. Other than these feature representations, there exists semantic content based semantic multinomial (SMN) representation which can be obtained by mapping the image I onto a point in a c dimensional probability simplex, here c corresponds to total number of concepts present is database images. In SMN representation each dimension represents a meaningful visual concept and its elements correspond to the posterior probabilities $(\pi_1, \pi_2, \pi_3, \ldots, \pi_c)$ of the semantic concepts. An important issue in generating SMN representation is the availability of true concept labels for database images. In the absence of true concept labels, we propose to obtain the semantic scene representation using pseudo-concepts that are obtained using (a) hand crafted local feature vectors [1] (b) last conv layer filter responses [2], of all database images. In our prior work [1], pseudo-concepts models are build using dynamic kernel (IFK, CIGMMIMK, GMMSPMK) based SVMs. Disadvantages of the SMN representation proposed in [1] are (a) hand-crafted features used for concept model building are the local descriptors and do not capture semantic concepts as whole (b) concept models are built using local features of complete image instead of concept specific features, as a single image comprises of multiple concepts. In order to overcome from these disadvantages, we further propose weakly supervised pseudo-concept modeling for generation of deep SMN representation using last conv layer activation maps responses of deep CNNs in the absence of true concept labels.

II. Proposed framework

In this work, we propose to use conv layer filters as concepts detector, but ground truth information of filters ($i.e.$, which filter is learning what concept) is not known during training process of deep CNNs. Hence, we can not infer true-concept identity of particular filter from its activation. However, we can only visualize the activation maps responses of images using different visualization techniques. In Fig. 1, we have shown visualization results of the MIT67 dataset images for few filters of Places365-AlexNet [3] conv5 using deep-vis toolbox [4]. From visualization result, we notice that the activation responses are visually similar in structure and semantically consistent. Proposed framework of building the concept model is free from any image label, concept/region annotation or segmentation. A set of concept rich images with corresponding activation maps are sufficient for pseudo-concept modeling. Since we are not using the true concepts labels of images, ($i.e.$, marked bounding box or pixel-wise concept label) but only the pseudo-concepts identity generated from conv layer filters of CNNs, the process is weakly supervised. The proposed framework includes:

1) Selection of pseudo-concepts and corresponding data generation from the conv layer filters responses in the absence of true concepts label data [2].
2) Subspace analysis of pseudo-concepts data for pseudo-concepts grouping.
3) Weakly supervised pseudo-concept modeling using LK-based SVM when images are pass in fixed size to CNN and using DSPMK-based SVM to handle varying size conv layer activation maps when images are pass in their true $i.e.$, varying size to CNN.

Fig. 1. Visualization of few pseudo-concept filters for MIT67 Indoor scene dataset using AlexNet architecture. Here, 3, 9, 12 are few of the prominent filter numbers. Rectangular mask indicate pseudo-concept captured by those filters.

4) Deep CNN-based novel SMN representation for scene images classification.

The detailed explanation of proposed work is under review which is extension of a preliminary work published in [2].

III. Experimental studies

In this section, we study the strength of the proposed deep CNN-based SMN representation for scene recognition task using χ^2 kernel-based SVM classifiers. In our studies, we use VGGNet-16 as pre-trained CNN architecture [3] for pseudo-concept selection and generation of SMN representation. VGGNet-16 is employed without its fully-connected (fc) layers as concept information and spatial structure of concepts are preserved in activation maps of convolutional layer. We evaluated our proposed approach on four widely used scene image classification datasets: MIT-8 scene, Vogel-Schiele (VS), MIT-67 and SUN-397. Details about datasets is given in our previously published work [1], [5]. The number of filters f in last conv layer of VGGNet-16 is 512. Number of selected pseudo-concepts c from f after subspace analysis of pseudo-concept data for VS, MIT-8, MIT-67 and SUN-397 are 160, 105, 385 and 498 respectively. Lower bound on the minimum number of images required in any pseudo-concept class is empirically chosen as 30. Default value of trade-off parameter $C = 1$ is considered in all the experiments. The effectiveness of the proposed SMN representation is studied by building the χ^2 kernel-based SVM classifier with one-vs-rest approach of LIBSVM tool kit. In Table I results for scene recognition using proposed framework are shown and compared with classification using fc7 and pool5 features of pre-trained CNN. It is seen that proposed SMN representation for scene images classify the scenes better. It is observed that for small datasets like MIT-8 scene and VS less number of pseudo-concepts are enough but for complex scene dataset such as MIT67 and SUN397 large number of pseudo-concepts are needed. We observe that pseudo-concept modeling is further improved by passing images to CNN in their true size, as resizing results in loss of some concept related information. However this results in varying size activation maps. For handling the same, we build the pseudo-concept model using DSPMK-based SVMs.

TABLE I
Classification accuracy (CA) (in %) over multiple folds on scene-centric databases using Deep-CNN-based SMN representation and the SVM-based classifier. Base features for building SMN are extracted from VGG16 Net which is pretrained on Places-205 and Places-365 dataset respectively.

Method	VS	MIT-8	MIT-67	SUN-397
SIFT+BoVW	67.49	79.13	45.86	24.82
Places-Alex-fc7[3]	76.02	88.30	68.24	54.32
Places-VGG16-pool5[3]	79.12	89.10	74.32	58.21
SMN using last conv layer of Places205-VGG with pseudo-concept modeling using LK	85.01	93.23	78.87	60.78
SMN using last conv layer of Places365-VGG with pseudo-concept modeling using LK	84.67	93.88	73.86	63.81
SMN using last conv layer of Places205-VGG with pseudo-concept modeling using DSPMK	87.65	95.87	82.45	62.45
SMN using last conv layer of Places365-VGG with pseudo-concept modeling using DSPMK	85.98	96.08	80.65	63.92

IV. Conclusion

In this work, we have proposed a novel deep CNN-based SMN representation for classification of complex scene image datasets. Weakly supervised pseudo-concept modeling using deeper conv layer filters responses is proposed in the absence of true concept labeled images. Effective number of pseudo-concepts present in dataset are obtained using proposed pseudo-concept selection procedure and subspace analysis of pseudo-concept class data. The proposed SMN representation captures the information of diverse concept present in the image and hence results in state-of-the art classification accuracy.

References

[1] S. Gupta, A. D. Dileep, and V. Thenkanidiyoor, "The semantic multinomial representation of images obtained using dynamic kernel based pseudo-concept SVMs," in *Proceedings of National Conference on Communication (NCC 2017)*, Chennai, India, March 2017, pp. 1–6.

[2] S. Gupta, A. D. Dileep, and V.Thenkanidiyoor, "Deep CNN based pseudo-concept selection and modeling for generation of semantic multinomial representation of scene images," in *Proceedings of the ACM Conference on Data Science and Management of Data (CoDS-COMAD 2018)*, Goa, India, 2018, pp. 336–339.

[3] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[4] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," in *Proceedings of the Deep Learning Workshop in International Conference on Machine Learning (ICML 2015)*, 2015.

[5] S. Gupta, D. K. Pradhan, A. Dileep, and V. Thenkanidiyoor, "Deep spatial pyramid match kernel for scene classification," in *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, 2018, pp. 141–148.

V. List of publications

A. Conference publications

- **S. Gupta**, A.D. Dileep and V. Thenkanidiyoor, "Segment-Level Pyramid Match Kernels for the classification of varying length patterns of speech using SVMs", in *Proceedings of 24th European Signal Processing Conference (EUSIPCO 2016)*, Budapest, August 2016, pp. 2030-2034.
- **S. Gupta**, V. Thenkanidiyoor and A.D. Dileep, "Segment-Level Probabilistic Sequence Kernel based Support Vector Machines for Classification of Varying Length Patterns of Speech", in *Proceedings of 23rd International Conference on Neural Information Processing (ICONIP 2016)*, Kyoto, Japan, October 2016, Part IV, LNCS 9950, pp. 321-328.
- **S. Gupta**, A.D. Dileep and V. Thenkanidiyoor, "The Semantic Multinomial Representation of Images Obtained Using Dynamic Kernel Based Pseudo-concept SVMs", in *Proceedings of 23rd National Conference on Communications (NCC 2017)*, Chennai, India, March 2017, pp. 1-6.
- **S. Gupta**, D. Kumar, A.D. Dileep and V. Thenkanidiyoor, "Deep Spatial Pyramid Match Kernel for Scene Image Classification", in *Proceedings of 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM'18)*, Portugal, pp. 141-148.
- D. Kumar, **S. Gupta**, V. Thenkanidiyoor and A.D. Dileep, " Semantic Multinomial Representation for Scene Images using CNN-based Pseudo-concepts and Concept Neural Network", in *Proceedings of 6th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG'17)*, Mandi H.P, India.
- K. Sharma, **S. Gupta**, A.D. Dileep , R. Rameshan, "Scene Image Classification Through Reduced Virtual Feature Representation in Sparse Framework", in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP-2018)*, Calgary, Canada, pp. 2701-2705.
- **S. Gupta**, A.D. Dileep and V. Thenkanidiyoor, "Deep CNN based Pseudo-concept Selection and Modeling for Generation of Semantic Multinomial Representation of Scene Images", in *Proceedings of Young Researchers Symposium (CODS-COMAD 2018)*, Goa, India.
- **S. Gupta**, K. De, A.D. Dileep and V. Thenkanidiyoor, "Emotion recognition from varying length patterns of speech using CNN-based segment-Level pyramid match kernel based SVMs", accepted in *25rd National Conference on Communications (NCC 2019)*, Bangalore, India.

B. Extended abstracts for poster presentation

- **S. Gupta**, V. Thenkanidiyoor, A.D. Dileep, "Spatial Probabilistic Sequence Kernel for Scene Classification using Support Vector Machines", *CVPR 2016 workshop on Women in Computer Vision 2016 (WiCV 2016)*, Las Vegas, NV, USA.
- **S. Gupta**, S. Jain, V. Thenkanidiyoor, A.D. Dileep , "Semantic Multinomial Representation for Scene Images using Dynamic Kernel based SVMs", *Scene Understanding Workshop (SUNw 2016), (CVPR 2016)*, Las Vegas, NV, USA.
- **S. Gupta**, A.D. Dileep and V. Thenkanidiyoor, "Varying Size Scene Image Classification using CNN based DSPMK" *Grace Hopper Celebration India (GHCI 2018)*, Bangalore, India.
- **S. Gupta**, K. De, A.D. Dileep and V. Thenkanidiyoor"CNN-Based Segment-level Pyramid Match Kernel for the Classification of Varying Length Patterns of Speech", *Young Female Researchers in Speech Science and Technology Workshop (YFRSW), INTERSPEECH 2018*, Hyderabad India.

C. Book Chapter

- **S. Gupta**, M. Mangal, A. Mathew, A.D. Dileep, A. Bhavsar, V. Thenkanidiyoor. "CNN-based Deep Spatial Pyramid Match Kernel for Classification of Varying Size Images", in Pattern Recognition Applications and Methods, 2018, (Revised selected papers from ICPRAM 2018).

D. Journal

- **S. Gupta**, A. Karanath, K. Mahrifa, A.D. Dileep, V. Thenkanidiyoor, "Segment-level probabilistic sequence kernel and segment-level pyramid match kernel based extreme learning machine for classification of varying length patterns of speech", in *International Journal on Speech Technology (IJST)*, December 2018.
- **S. Gupta**, K. Sharma,, A.D. Dileep, V. Thenkanidiyoor, "Deep Semantic Multinomial Representation for Scene Recognition using Weakly Supervised Pseudo-concept Modeling", communicated to *IEEE Transactions on Multimedia*, December 2018 (under review).

Spontaneous Facial Expression Recognition in Wild.

1st Nazil Perveen
Department of CSE, IIT Hyderabad
Hyderabad, India
cs14resch11006@iith.ac.in

2nd C Krishna Mohan
Department of CSE, IIT Hyderabad
Hyderabad, India
ckm@iith.ac.in

ABSTRACT

Motivation- Facial expressions are the vital signaling system that reflects human emotions through facial muscle movements. There is multiple facial expressions formation depending on the combination of movement of the facial muscles. In the prior research works, various facial muscle movements are categorized into seven different categories of universal facial expressions, namely, angry, disgust, fear, happy, sadness, and surprise. Around the decades, multiple approaches try to address the automatic recognition of facial expressions for various applications. The issue in most of the approaches are, they are application dependant, highly constrained, and consist of posed (acted) expressions. The main objective of this research work is designing and development of the model, which recognize spontaneous (not acted) facial expressions in the unconstrained environment irrespective of views and subjects for real world applications. We also explore the application of facial expression to detect facial paralysis and its degree of effectiveness.

Related Work- Approaches for automatic facial expression recognition (AFER) can be achieved in two ways, i.e. parameter based AFER and vision based AFER. Parameter based FER system is more subjective where the facial muscles are basically considered to recognize human expression. This is further categorize into descriptive parameter based FER and judgemental based FER. Descriptive parameter based recognition usually captures the facial parts movements, like eyebrow, eye, cheek, lip movements. These movements are first analyzed by Carl and later elaborated by Ekman as facial action units (FAUs). These FAUs are then combined using certain rule based system known as facial action coding system (FACS). In general, FACS are used by psychologist for recognizing facial expressions. Similarly, subjective conclusions are made for judgemental parameter based expression recognition. Judgemental based FER considered to capture the latent emotions, i.e. emotion which are generated by questionnaire or by watching emotional activity etc. Another approach to automatic FER is vision based system which has pipelined version of recognizing facial expression with each stage has research area in itself. In the vision-based recognition system, representation of facial expressions plays an important role.

Based on the previous literature there are two approaches for representing facial expression dynamics, explicit modeling and implicit modeling of facial expression dynamics. Explicit methods basically consist of low level or hand coded feature representation like histogram of oriented gradients (HOG), scale invariant feature transform (SIFT), Gaussian, local phase quantization (LPQ), etc. Explicit modeling features are generally meant for image based feature representation. Recently, many temporal descriptors along with spatial representation are used to provide discriminative information like local binary pattern three orthogonal planes (LBP-TOP). On the other hand implicit feature representation are high-level feature representation, which learns the spatial and temporal dynamics of facial expression implicitly. Implicit modelling are meant for visual based feature representation. In this research work, we explore implicit based feature representation for capturing spatial and temporal dynamics like convolution neural network (CNN) and Gaussian mixture modelling (GMM).

Also, as an extension for facial expression recognition, quantitative and subjective assessment for facial paralysis patients are analysed for providing aid to clinicals. For detecting facial paralysis and measuring the degree of paralysis effect on the patients multiple approaches like multi-resolution LBP (M-LBP), thermal imaging, active appearance modeling (AAM), and active shape model (ASM) are used in literature. Issues in most of the approaches are, only frontal faces are considered for the facial paralysis assessment, only a few facial expressions, which involves early and active facial movements like raising of eyebrows, closure of eye tightly, wrinkle nose, and toothy movement are considered for the analysis of facial paralysis, and only 3 score grading scales are considered for evaluation i.e. score-0 (high level facial paralysis), score-1 (middle-level of facial paralysis), and score-2 (least level facial paralysis). To overcome these limitations, we introduce the quantitative assessment model which is subject and view independent with detailed 5-score and 3-score grading scale analysis. The proposed work also models all types of expressions to provide a better assessment to clinical.

Proposed Approach- A part-based approach for spontaneous expression recognition using audio-visual feature and deep convolution neural network (DCNN) is proposed. The ability of convolution neural network to handle variations in translation and scale is exploited for extracting visual features. Initially, the face is detected and landmarks are used to divide the facial region into two most expressive facial parts i.e. eye and mouth regions. The sub-regions, namely, eye and mouth

parts extracted from the video faces are given as an input to the deep CNN (DCNN) inorder to extract convnet features. The audio features, namely, voice-report, voice intensity, and other prosodic features are used to obtain complementary information useful for classification. The confidence scores of the classifier trained on different facial parts and audio information are combined using different fusion rules for recognizing expressions. The effectiveness of the proposed approach is demonstrated on acted facial expression in wild (AFEW) dataset.

Another approach is to recognize spontaneous facial expression in the wild using the configural representation of facial action units. Each facial muscle movement form certain facial action units (FAUs) like raising of eyebrow, eyes, mouth open etc. To recognize facial expressions using these FAUs we propose an approach for spontaneous expression recognition in the wild using configural representation of facial action units. Since all configural features do not contribute to the formation of facial expressions, we consider configural features from only those facial regions where significant movement is observed. These chosen configural features are used to identify the relevant facial action units which are combined to recognize facial expressions. Such combinational rules are also known as coding system. However, the existing coding systems incur significant overlap among facial action units across expressions, we propose to use a coding system based on subjective interpretation of the expressions to reduce the overlap between facial action units which leads to better recognition performance while recognizing expressions.

The most challenging issue in recognizing spontaneous facial expressions are availability of the database and the manual annotations of the database by the experts. Therefore next, we develop the unsupervised framework for learning the discriminative features from the videos to recognize facial expressions without any annotations. Initially, spatial and temporal features like histogram of oriented gradients (HOF) and motion boundary histogram (MBH) are extracted from the detected facial region. The large universal Gaussian mixture model or universal attribute model (UAM) is trained to adapt the multiple facial muscle movements implicitly whose combinations form various facial expressions. The maximum posterior adaptation of UAM means are then used to form the super expression vector (SEV), which is huge in dimension and contain redundant attributes from each expressions. Therefore, we use factor analysis for low-dimensional effective representation of the SEV vector, also known as expression vector. The proposed methodology does not require any class label to classify the expression using distinct expression vector for each expressions.

The proposed methodology of expression vector are also applied for quantitative assessment of facial paralysis analysis. Facial paralysis is the growing disease with difficulty in the facial muscle movement on either one side or both side of the face. The subjective assessments are widely used techniques to determine the measure of degree with which the patient is affected. However, the subjective assessments are highly dependant on the experts view and a few set of grading rules. In this paper, the quantitative assessment to measure the degree of facial paralysis is proposed. A video database of 85 facially paralyzed patients of different age groups and gender is collected under expert supervision. Using Yanagihara grading scale, experts assign subjective scores to the patients based on the degree of facial paralysis, which is used as a ground truth. The database collected consists of seven different views and multiple subjects with ten different expressions suggested by the experts. The proposed model measures different degree of facial paralysis for all types of expressions better than existing approaches for quantitative assessment.

Empirical Results- We demonstrate the efficacy of the proposed approach in the publicly available benchmark databases, namely, (i) constrained database like Maja and Michael Initiative (MMI) facial expression database, (ii) unconstrained database like automatic facial expression in the wild (AFEW), and (iii) spontaneous database like Bosphorous 4d facial expression (BP4D) database. The experimental results shows the proposed methodology outperform the state of the art techniques with the increase in the rate of classification performance of 3%in MMI, 12% in AFEW, and 10% in BP4d database, respectively. Also, the proposed database construction for facial paralysis assessment outperform the current trends in subjective and quantitative assessment analysis.

List of Publications

- **International Journal**
  1) Nazil Perveen, Debaditya Roy, Chalavadi Krishna Mohan, "Spontaneous Expression Recognition Using Universal Attribute Model", IEEE Transactions on Image Processing, vol. 27, no. 11, pp. 5575-5584, 2018. **(ACCEPTED)**
- **International Conference**
  1) Nazil Perveen, Dinesh Singh and C. Krishna Mohan, "Spontaneous Facial Expression Recognition: A Part Based Approach", IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, California, USA, Dec. 18-20, 2016. **(ACCEPTED)**
  2) Nazil Perveen, C. Krishna Mohan, Naoki Matsushiro, Masataka Seo, Yutaro Iwamoto, and Yen-Wei Chen,"Quantitative assessment for facial paralysis using GMM and factor analysis", Submitted to IEEE Winter Conference on Applications of Computer Vision (WACV2019). **(Under Review)**
  3) Nazil Perveen and C. Krishna Mohan, "Configural representation of facial action units for spontaneous facial expression recognition in the wild", Submitted to IEEE Winter Conference on Applications of Computer Vision (WACV2019). **(Under Review)**

Comparison of Ocular Parameters for Distraction Detection of Drivers in Cars

Gowdham Prabhakar
Center for Product Design and Manufacturing
Indian Institute of Science
Bangalore, India
gowdhamp@iisc.ac.in

Pradipta Biswas
Center for Product Design and Manufacturing
Indian Institute of Science
Bangalore, India
pradipta@iisc.ac.in

I. INTRODUCTION

In automotive domain, distraction of drivers increases with increasing complexity of the user interaction with secondary tasks like infotainment, lights, adjusting windows, side mirrors, and so on. Though there are different modalities like voice and gesture recognition systems coming up to facilitate undertaking secondary tasks, reducing the cognitive load of drivers is challenging. This dissertation aims at decreasing the cognitive load of drivers and detect the distraction of drivers while operating secondary tasks. The first phase of dissertation explores the different modalities of operating secondary tasks. The second phase will explore the detection of distraction due to operating secondary tasks.

In the early stage of research, we explored different modalities for operating a GUI display by virtual touch. We evaluated finger movement-based trackers and laser tracker. We found Laser tracker outperforming other devices and chose to adapt it in automotive environment. We proposed two new interaction devices using a Laser pointer with hardware switch as well as Laser pointer with eye gaze switch which does not require drivers to physically touch a display. We compared their performances with a touchscreen in the automotive environment by measuring the pointing and selection time for a secondary task, deviation from the lane, average speed, variation in steering angle, cognitive load and system usability. Our results found that the Laser pointer tracking system with eye gaze switch did not significantly degrade driving or pointing performance compared to the touchscreen in standard ISO 26022 lane changing task. We have filed a provisional patent for Laser tracker with eye gaze switch as well as hand tap switch.

We have also evaluated eye gaze-controlled interaction for automotive as well as aviation. We have built a projected transparent screen for projecting the dashboard display. This projected display can be mounted anywhere on the windshield and does not occlude road vision. We have evaluated eye tracker-based interaction with this projected display. Our user studies involving driving and flight simulators have found that the proposed projected display can improve driving and flying performance and significantly reduce pointing and selection times for secondary mission control tasks compared to existing interaction systems.

We have investigated ocular parameters like pupil dilation and saccadic intrusion to estimate cognitive load of drivers. We have correlated these ocular parameters with EEG and head movement in a dual task study involving a driving simulator. We found that our proposed algorithm significantly classified drivers' cognitive state from normal to distracted conditions. We undertook another study inside vehicles driven by professional drivers. We recorded ocular parameters and analysed their temporal association with upcoming road hazards. We presented detailed analysis on comparing different algorithms analysing pupil dilation and saccadic intrusion with respect to oncoming road hazards and reported F1 scores on temporal association of ocular parameters and hazards. Our proposed algorithm using pupil dilation was able to detect 87% of road hazards in comparison to saccadic intrusion which detected 44% of road hazards.

II. USER STUDY

A. Procedure

There was no traffic on the road in the driving simulator. The assembly of the setup is illustrated in Figure 1. The first trial was to record reference data by letting the participant take a free drive without doing any secondary task. This was taken as reference case (C1). In the second trial, the participant had to drive as well as follow the Lane changing instructions. This trial corresponds to case 2 (C2). In the third trial, the participant had to drive with lane changing instructions as well as perform a secondary task of selecting buttons on the dashboard display in response to an auditory cue. This trial corresponds to case 3 (C3). The dashboard display is mimicked from one of the existing dashboard displays of Jaguar Land Rover. The dashboard display was displayed to the left of the driving simulator (for right hand driving in India).

B. Sum of Magnitude of Single-sided Spectrum (SMSS)

An FFT (Fast Fourier Transform) was performed over the raw data of pupil dilation, head yaw and EEG (T7). The sum of magnitude of single-sided spectrum (SMSS) was calculated for every 1-sec length of the signal. The SMSS of FFT for each participant was compared if the SMSS in C3> SMSS in C2> SMSS in C1.

C. Measure of Cognition due to Distraction (MCD)

The raw data of pupil dilation, head yaw and EEG were processed for coefficients of time-frequency components using DWT (Discrete Wavelet Transform) as well as CWT (Continuous Wavelet Transform). If the value was higher than the threshold, it was assigned a binary value 1. If the value was less than the threshold, it was assigned a binary value 0. The total number of such thresholded peaks were counted for each case. We refer this number as MCD (Measure of Cognition due to Distraction). MCD values were compared between cases if MCD of C3> MCD of C2> MCD of C1 for each participant. Each set of data (pupil dilation, head yaw, and EEG) was calculated for MCD values from DWT and CWT for the full-length raw signal as well as a 1-second window (Marshall 2007) of the raw signal for real-time implementation. The zeros in data were

removed as it was due to either the inability of the tracker to capture the eyes or the participant blinked his/her eyes. A set of 200 values were trimmed out from the beginning of the data after the calibration of data was done. Amor wavelet function was used for CWT and db8 (Daubechies 8) wavelet function was used for DWT.

D. *Results*

Initially, we took the full-length signal of head yaw, pupil dilation and EEG (T7) and performed FFT, DWT and CWT. After we found significant difference between three cases of driving for all sensor data, we performed same analyses for 1-second window running over the full-length signal of each sensor data. All the signal processing techniques were carried out using the MATLAB inbuilt functions. The mean SMSS of pupil dilation of left eye of all participants is plotted for each case of driving as shown in Fig. 1.

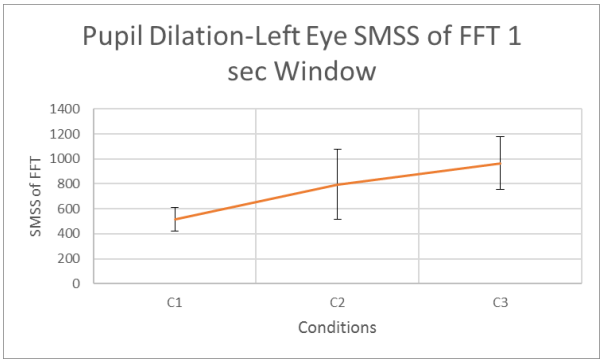


Fig. 1. Mean SMSS (in 1 sec windows) of pupil dilation of left eye for three cases

A Kruskal-Wallis test found a significant difference ($H=19.16$, $p < 0.01$) between mean ranks of at least one pair of groups. Signed rank tests were carried out for three pairs of groups. There was also an evidence ($p<0.01$) of a difference between pairs C1vsC2 and C1vsC3. There was no significant difference between the pair C2vsC3. A Kruskal-Wallis test did not find any significant difference between the groups for the SMSS in windows of 1 second for pupil dilation of the right eye, head yaw, and EEG. We performed similar analysis for SMSS as well as MCD for pupil dilation of left and right eyes, head yaw and EEG data in 1 second window as well as for full-length signal. Most results using MCD found significant difference between mean ranks of at least one pair of groups.

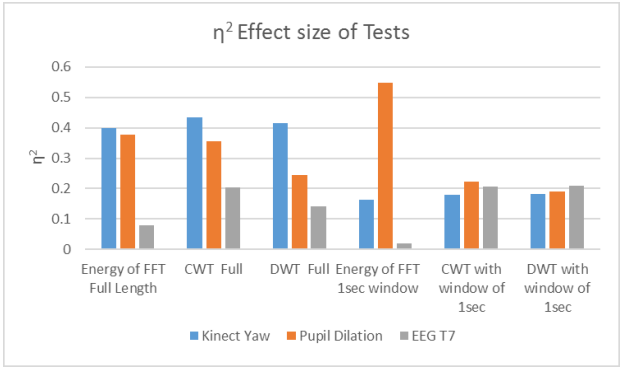


Fig. 2. Effect size of each test to find best performance for real-time implementation

We have plotted the effect size ($\eta 2$) of coefficients of head yaw, pupil dilation and EEG corresponding to different tests we analyzed in Fig. 2. From the graph, we can see the effect size is higher for different sensors for same test methods. For real-time (1 second window) implementation of the distraction detection system, DWT (MCD) can be chosen for head yaw and EEG whereas FFT (SMSS) can be chosen for pupil dilation. For implementing detection methods in real-time, we thresholded SMSS values per second and MCD values per second. If the value is greater than threshold, it will interpret as a detection and is represented by value 1 and a value 0 for no detection. If the system detects a distraction (value 1), it can alert the driver. Detections happened in a sliding window of 1 second. We have plotted the mean of maximum F1 score versus the different boundary of hazard-timestamp for SMSS of pupil dilation and SI velocity of all drivers in Fig. 3.

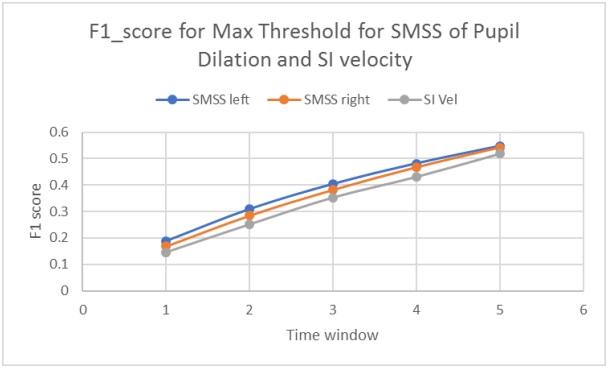


Fig. 3. Accuracy for different boundary of hazard-timestamp for SMSS of pupil dilation and SI velocity

III. Conclusion

We have initially conducted a user study in which the participants undertook driving task along with the operation of secondary tasks in a driving simulator. Our results showed that the pupil dilation, head yaw, and EEG can detect the increase in cognitive load due to operation of secondary task within a time buffer of 1 second which gave the confidence to investigate this method in cars. We have also found that an FFT shows better performance in detection of rise in cognitive load than DWT and CWT for pupil dilation. Though the EEG did not give expected results, the synergy of head movement (yaw) and pupil dilation makes the system robust to noise while detecting the increase in cognitive load of the driver. We conducted another user study in which eye metric data and scene video from professional drivers were recorded. We calculated the optimal threshold for maximum accuracy of detection for each driver. Our proposed SMSS algorithm was able to detect 87% of the road hazards in comparison with Saccadic velocity algorithm which detected 44% of road hazards. Though we were not able to find a universal threshold for all drivers, we are planning to improve the accuracy of detection by not limiting our ground truth of detections to road hazards. We are planning to investigate the detections which were not associated with road hazards. We also planned to improve the validity of road hazards by collecting tags from more than one individual. The working of distraction detection in cars with professional drivers shows an external validity of our proposed system outside simulation environment. The usage of wearable glass-based eye tracker

helps in detection of cognitive activity from pupil dilation without the need of limiting the head movement by a chin-rest. In the future, we are planning to implement our proposed expert system to detect distraction and integrate with the Alert system for real cars as well as aircraft cockpits.

IV. LIST OF PUBLICATIONS

[1] Prabhakar, G. and Biswas, P. (2018). Eye Gaze Controlled Projected Display in Automotive and Military Aviation Environments. Multimodal Technologies and Interaction, 2(1)

[2] Prabhakar, G., Madhu, N. & Biswas, P. (2018). Comparing Pupil Dilation, Head Movement, and EEG for Distraction Detection of Drivers, Proceedings of the 32nd British Human Computer Interaction Conference 2018 (British HCI 18)

[3] Prabhakar, G., & Biswas, P. (2017, July). Evaluation of laser pointer as a pointing device in automotive. In Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2017 International Conference on (pp. 364-371). IEEE. [Best Paper Award]

[4] Prabhakar, G., Rajesh, J., & Biswas, P. (2016, December). Comparison of three hand movement tracking sensors as cursor controllers. In Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2016 International Conference on (pp. 358-364). IEEE. [Best Paper Award]

[5] Prabhakar, G. and Biswas, P. (2017, September). Interaction Design and Distraction Detection in Automotive UI, ACM Automotive UI Doctoral Consortium (ACM Automotive UI 2017) [ACM Travel Award]

[6] Biswas, P., & Prabhakar, G. (2018). Detecting drivers' cognitive load from saccadic intrusion. Transportation research part F: traffic psychology and behaviour, 54, 63-78.

[7] Biswas, P., Prabhakar, G., Rajesh, J., Pandit, K., & Halder, A. (2017, July). Improving eye gaze controlled car dashboard using simulated annealing. In Proceedings of the 31st British Computer Society Human Computer Interaction Conference(p. 39).

[8] Biswas, P., Roy, S., Prabhakar, G., Rajesh, J., Arjun, S., Arora, M., & Chakrabarti, A. (2017, July). Interactive sensor visualization for smart manufacturing system. In Proceedings of the 31st British Computer Society Human Computer Interaction Conference (p. 99).

[9] Prabhakar, G., & Biswas, P. Wearable Laser Pointer for Automotive User Interface, Patent Application No.: PCT/IB2018/057680.

V. REFERENCES

[1] Afzal, S., & Robinson, P. (2009, September). Natural affect data—Collection & annotation in a learning context. In Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on (pp. 1-7). IEEE.

[2] Basir, O., Bhavnani, J. P., Karray, F., & Desrochers, K. (2004). Drowsiness detection system, US 6822573 B2.

[3] Biswas, P., & Prabhakar, G. (2018). Detecting drivers' cognitive load from saccadic intrusion. Transportation research part F: traffic psychology and behaviour, 54, 63-78.

[4] Boril, H., Sadjadi, S. O., & Hansen, J. H. L. (2011). UTDrive: Emotion and cognitive load classification in-vehicle scenarios. In Proceeding of the 5th biennial workshop on DSP for in-vehicle systems.

[5] Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., & Giannopoulos, I. (2018, April). The Index of Pupillary Activity: Measuring Cognitive Load vis-à-vis Task Difficulty with Pupil Oscillation. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (p. 282). ACM.

[6] Gavas, R., Chatterjee, D., and Sinha, A. (2017), "Estimation of cognitive load based on the pupil size dilation," 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, 2017, pp. 1499-1504.

[7] Healey, J. A., & Picard, R. W. (2011). Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems, 6(2), 156–166.

[8] Hess, E. H. (1975). The tell-tale eye. Van Nostrand Reinhold Company.

[9] Lee, Y. C., Lee, J. D., & Ng Boyle, L. (2007). Visual attention in driving: the effects of cognitive load and visual disruption. Human Factors, 49(4), 721-733.

[10] Liang, Y., & Lee, J. D. (2014). A hybrid bayesian network approach to detect driver cognitive ditraction. Transportation Research Part C, 38, 146–155.

[11] Marshall, S. (2002). The index of cognitive activity: Measuring cognitive workload. In Proc. 7th conference on human factors and power plants (pp. 7–5).

[12] Marshall, S. (2007). Identifying cognitive state from eye metrics. Aviation, Space, and Environmental Medicine, 78(Suppl. 1), B165–B175.

[13] Palinko, O., Kun, A. L., Shyrokov, A., Heeman, P., (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In Proceedings of the 2010 symposium on eye-tracking research & applications (pp. 141–144).

[14] Prabhakar, G., & Biswas, P. (2018). Eye Gaze Controlled Projected Display in Automotive and Military Aviation Environments. Multimodal Technologies and Interaction, 2(1), 1.

[15] Ranney, T. A., Baldwin, G. H., Smith, L. A., Martin, J., & Mazzae, E. N. (2013). Driver behavior during visual-manual secondary task performance: occlusion method versus simulated driving (No. DOT HS 811 726).

[16] Redlich, E. (1908). Ueber ein eigenartiges Pupillenphänomen; zugleich ein Beitrag zur Frage der hysterischen Pupillenstarre. Deutsche medizinischeWochenschrift, 34, 313–315.

[17] Sezgin, T. M., & Robinson, P. (2007, September). Affective video data collection using an automobile simulator. In International Conference on Affective Computing and Intelligent Interaction (pp. 770-771). Springer, Berlin, Heidelberg.

[18] Tokuda, S., Obinata, G., Palmer, E., & Chaparo, A. (2011). Estimation of mental workload using saccadic eye movements in a free-viewing task. In 23rd international conference of the IEEE EMBS (pp. 4523–4529).

[19] Westphal, A. (1907). Ueber ein im katatonischen stupor beobachtetes Pupillenphänomen sowie Bemerkungen über die Pupillenstarre bei Hysterie. Deutsche medizinische Wochenschrift, 33, 1080–1084.

[20] Yoshida, Y., Ohwada, H., Mizoguchi, F., & Iwasaki, H. (2014). Classifying cognitive load and driving situation with machine learning. International Journal of Machine Learning and Computing, 4(3), 210.

[21] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang (2009), "A Survey of Affect Recognition Methods: Audio Visual & Spontaneous Expressions", IEEE Trans. PAMI, vol. 31, no. 1, pp. 39-58.

Efficient Compressed Sensing Magnetic Resonance Image Reconstruction for Clinical Applications

Sumit Datta and Bhabesh Deka Senior Member, IEEE
Department of Electronics and Communication Engineering
Tezpur University, Tezpur-784028, Assam, India
e-mail: sumit89@tezu.ernet.in, bdeka@tezu.ernet.in.

Magnetic resonance imaging (MRI) is one of the most rapidly growing medical imaging techniques as it does not use any ionizing radiation; able to provide high contrast for soft tissues, like, brain, abdomen, etc. However, traditionally it suffers from a serious limitation owing to its slow acquisition speed. The compressed sensing (CS), introduced by Donoho [1] and Candes *et al.* [2], is a technique of efficiently acquiring a signal at sub-Nyquist rate and reconstructing the original signal back from the few measured values with a decent accuracy. The two key requirements for the successful application of the CS are- (1) signal or image should be either sparse or compressible in nature or they become so under some transformation, and (2) the aliasing artifacts due to undersampling must be incoherent to the transform domain.

MRI is naturally satisfies both requirements, namely, the sparsity and incoherence. This is so, because MR images are in general compressible in a transform domain, like, wavelets. Further, MRI data are acquired in k-space which is equivalent to the Fourier transform of the MR image and sufficient incoherence exists between the k-space and the wavelet transform. It means significant reduction in MRI acquisition time may be achieved at reduced cost and most importantly, improves the patient comfort substantially [3]. Mathematically, the CS-MRI reconstruction problem can be represented as-

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & ||\mathbf{\Psi x}||_1 + \lambda\, ||\mathbf{x}||_{\text{TV}} \\ \text{subject to} \quad & ||\mathbf{F}_u\mathbf{x} - \mathbf{y}||_2 \leq \varepsilon \end{aligned} \tag{1}$$

where $\mathbf{x}$ is the MR image to be reconstructed, λ is a regularization parameter, $\mathbf{y}$ is the measured k-space data and $\mathbf{F}_u$ is the undersampling Fourier operator which mimics the data acquisition in k-space domain. The total variation (TV)- norm may be defined as-

$$||\mathbf{x}||_{\text{TV}} = \sum_{ij} \sqrt{\left(\mathbf{D}_x\, x_{ij}\right)^2 + \left(\mathbf{D}_y\, x_{ij}\right)^2}$$

where $\mathbf{D}_x$ and $\mathbf{D}_y$ denote the finite difference operators on the x and y -axis respectively.

A number of algorithms have been proposed to solve the CS-MRI reconstruction problem. Among them some of the well known reconstruction techniques are- the fast composite splitting algorithm (FCSA) [4], the wavelet tree sparsity MRI (WaTMRI) [5], the block sparsity and iterative support detection (SDBS) [6], and the lattice split augmented Lagrangian (LaSAL) [7]. Still it is an open challenge to reconstruct MR images with clinically acceptable quality using a minimum number of measurements. Depending on its application, our works can be divided into three categorizes as given below.

1) 2D MRI reconstruction:: During iterative ℓ_1-norm minimization, coefficients having larger magnitudes are heavily penalized than the smaller ones. To address this imbalance, Candes *et al.* [8] proposed the weighted ℓ_1-norm minimization instead of ℓ_1-norm, which tries to penalize uniformly all nonzero coefficients. Again, MR images are generally piecewise smooth, so, their gradients are also sparse. So, we use the concept of weighted minimization for both ℓ_1 and TV-norms in a composite minimization model. Besides, better reconstruction accuracy, it also improves the speed of convergence. Further, we propose an efficient adaptive weighting scheme to enhance the sparsity of MR images in both the wavelet as well as the spatial domain. The weights are updated in the current iteration depending on the results obtained in the previous iteration. The computational cost of the weight calculation is only $O(n)$. Although, computational cost per iteration increases slightly, the total number of iterations required for convergence reduces sufficiently.

Due to k-space undersampling artifacts are generated and appear as white noise in MR images obtained by inverse Fourier transform of undersampled data. In wavelet domain energy leakage occurs particularly in detailed subbands after wavelet decomposition. Applications of standard magnitude based thresholding in the wavelet domain cannot accurately identify the wavelet coefficients' support. Wavelet coefficients of MR images have unique properties, like, non-Gaussianity and persistency. Hidden Markov tree (HMT) successfully model these properties of wavelet coefficients [9]. Using the wavelet domain HMT, we have retrieved the support information with better accuracy. It improves the quality of CS-MRI reconstruction or in other words reduce the amount of measurements to achieve the same quality of reconstruction.

2) 2D multi-slice or 3D MRI reconstruction:: In clinical practice 2D multi-slice or 3D data MRI is very common, where a significant number of slices are acquired from a small volume of body with zero or negligible inter-slice gap. Due to this reason, adjacent slices are highly correlated. One can estimate missing k-space samples of a particular undersampled slice from samples of the adjacent undersampled slices [10] which are relatively less undersampled. This simply means

that in the k-space multi-slice sequence, some slices may be heavily undersampled (H-slice) while keeping their neighbors lightly undersampled (L-slice). We assume that in a group of three adjacent slices center slice is the L-slice and the rest two are H-slices. So, the entire multislice sequence could be {...-H-L-H–H-L-H-... }. Assuming line type of k-space trajectories, in case of H-slice only a few consecutive center rows are acquired. On the other hand, in case of L-slice some center rows along with randomly selected periphery rows are acquired as well. To estimate missing samples of a target H-slice, we propose a 3D k-space interpolation scheme. First, we obtained a virtual L-slice (V-slice) from acquired samples of two neighboring L-slices, i.e. {...-H-L-H-V-H-L-H-... } . Second, we estimate the missing samples of a target H-slice from its adjacent V-slice and L-slice. After the interpolation both interpolated H-slices and L-slices contain similar amounts of k-space data. Here, our objective of the work is twofold i.e. to reduce the overall undersampling ratio for multi-slices MRI, followed by the CS reconstruction with both interpolated and originally acquired data, which results reduction in scan time without sacrificing the reconstruction quality.

In literature [11], it is reported that wavelet coefficients of MR image follow a quadtree structure also known as wavelet tree sparsity i.e. properties of any non-leaf detail coefficient are followed by corresponding four coefficients of the finer scale just below it. As in multi-slice MRI, adjacent images are highly correlated, trends of wavelet transform coefficients of adjacent images are similar. This inter-slice similarity in wavelet domain can be modeled as the wavelet domain forest sparsity i.e. a number of connected trees. Moreover, gradients of adjacent images are also similar in similar positions. To utilize this inter-slice similarity we propose a joint multi-slice reconstruction model consisting of wavelet domain forest sparsity and joint TV regularization terms. These group-sparsity based regularization terms enforce similarities among different coefficients of the same position in both wavelet domain as well as spatial domain gradients across adjacent slices during the iterative group soft-thresholding process.

3) 3D parallel MRI reconstruction:: In parallel MRI (pMRI), to speed up the scan time multiple receiver coils are used around the target field of view (FoV). A single MR image is reconstructed using information acquired from all receiver coils or channels. Depending on the use of coil sensitivity information, there are two types of approaches. Some methods explicitly require coil sensitivity profiles during reconstruction while others use the coil sensitivity information implicitly by estimating the same from the acquired data, known as auto-calibrating methods. The most important concern is that a small error in the sensitivity profile may lead to significant artifacts in the reconstructed image. Well known CS-based pMRI approaches, namely, iterative self-consistent parallel imaging reconstruction from arbitrary k-space (SPIRiT) [12] and efficient L1SPIRiT reconstruction (ESPIRiT) [13] are auto-calibrating methods. Due to the capability of CS-MRI reconstruction, some calibrationless work [14] are also reported i.e. they do not require any calibration information explicitly or implicitly during reconstruction.

For clinical implementation of the CS-MRI, computational time of reconstruction is one of the major issues. In practice we need to reconstruct hundreds of 2D slices of multi-channel data for 3D MRI within a few minutes without compromising the diagnostic quality. To address this issue, we have implemented wavelet forest sparsity and joint TV regularizations based CS-MRI reconstruction algorithm in heterogeneous parallel computing workstation for multi-slice multi-channel data. Most of the existing parallel MRI reconstruction methods require coil sensitivity profile information either explicitly or implicitly. For these methods the quality of reconstruction highly depends on the calibration data and the approach of sensitivity profile estimation. Our method is calibrationless, thus no dependence on calibration data i.e. more robust for clinical implementation.

All these works are evaluated with clinical MRI datasets. For evaluation we have considered both objective parameters as well as subjective evaluation by radiologist and image processing experts. Moreover 3D parallel MRI reconstruction work is implemented in heterogeneous parallel computing workstation equipped with multi-core processor and GP-GPU hardware.

REFERENCES

[1] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candes, J. K. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[3] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, 2007.

[4] J. Huang, S. Zhang, and D. N. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," *Medical Image Analysis*, vol. 15, no. 5, pp. 670–679, 2011.

[5] C. Chen and J. Huang, "Exploiting the wavelet structure in compressed sensing MRI," *Magnetic Resonance Imaging*, vol. 32, pp. 1377–1389, 2014.

[6] Y. Han, H. Du, W. Mei, and L. Fang, "MR image reconstruction with block sparsity and iterative support detection," *Magnetic Resonance Imaging*, vol. 33, no. 5, pp. 624 – 634, 2015.

[7] M. Panic, J. Aelterman, V. Crnojevic, and A. Pizurica, "Sparse recovery in magnetic resonance imaging with a Markov random field prior," *IEEE Transactions on Medical Imaging*, vol. 36, no. 10, pp. 2104–2115, 2017.

[8] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted L1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[9] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1056–1068, 2001.

[10] Y. Pang and X. Zhang, "Interpolated compressed sensing for 2D multiple slice fast MR imaging," *Ed. Jonathan A. Coles. PLoS ONE*, vol. 8, no. 2, pp. 1–5, 2013.

[11] C. Chen, Y. Li, and J. Huang, "Forest sparsity for multi-channel compressive sensing," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2803–2813, 2014.

[12] M. Lustig and J. Pauly, "SPIRiT: iterative self-consistent parallel imaging reconstruction from arbitrary k-space," *Magnetic Resonance in Medicine*, vol. 64, no. 2, pp. 457–471, 2010.

[13] M. Uecker, P. Lai, M. J. Murphy, P. Virtue, M. Elad, J. M. Pauly, S. S. Vasanawala, and M. Lustig, "ESPIRiT–an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA," *Magnetic Resonance in Medicine*, vol. 71, no. 3, pp. 990 –1001, 2014.

[14] P. J. Shin, P. E. Z. Larson, M. A. Ohliger, M. Elad, J. M. Pauly, D. B. Vigneron, and M. Lustig, "Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion," *Magnetic Resonance in Medicine*, vol. 72, pp. 959–970, 2014.

I. List of Publications

A. Journals

1) Sumit Datta and Bhabesh Deka, "Efficient Interpolated Compressed Sensing Reconstruction Scheme for 3D MRI," *IET Image Processing*, vol. 12, no. 11, pp. 2119-2127, Aug. 2018. (DOI: 10.1049/iet-ipr.2018.5473)
2) Bhabesh Deka, Sumit Datta, and Sanjeev Handique, "Wavelet Tree Support Detection for Compressed Sensing MRI Reconstruction," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 730-734, May 2018. (DOI: 10.1109/LSP.2018.2824251)
3) Sumit Datta and Bhabesh Deka, "Magnetic Resonance Image Reconstruction using Fast Interpolated Compressed Sensing," *Journal of Optics*, vol. 47, no. 2, pp. 154-165, Sept. 2017. (DOI: 10.1007/s12596-017-0428-8)

B. Book

1) Bhabesh Deka and Sumit Datta, "Compressed Sensing Magnetic Resonance Image Reconstruction Algorithms - A Convex Optimization Approach," Springer Series on Bio- and Neurosystems, Springer Singapore, 2019. (DOI: 10.1007/978-981-13-3597-6)

C. Book Chapters

1) Sumit Datta and Bhabesh Deka, "Multi-channel, Multi-slice, and Multi-contrast Compressed Sensing MRI Using Weighted Forest Sparsity and Joint TV Regularization Priors," Soft Computing for Problem Solving (SoCProS 2017) , *Advances in Intelligent Systems and Computing*, Springer, 2018.
2) Bhabesh Deka and Sumit Datta, "Weighted Wavelet Tree Sparsity Regularization for Compressed Sensing Magnetic Resonance Image Reconstruction," Advances in Electronics, Communication and Computing (ETAEERE 2016), *Lecture Notes in Electrical Engineering*, vol. 443, pp. 449–457, Springer, 2016.

D. Conferences

1) Sumit Datta and Bhabesh Deka, "Efficient Adaptive Weighted Minimization for Compressed Sensing Magnetic Resonance Image Reconstruction," in *Indian Conference on Computer Vision, Graphics and Image Processing* (ICVGIP) 2016, Guwahati, India
2) Sumit Datta, Bhabesh Deka, Helal Uddin Mullah and Sushant Kumar, "An Efficient Interpolated Compressed Sensing Method for Highly Correlated 2D Multi-slice MRI," in *International Conference on Accessibility to Digital World* (ICADW) 2016, Guwahati, India
3) Sumit Datta and Bhabesh Deka, "Magnetic Resonance Image Reconstruction using Fast Interpolated Compressed Sensing," in *International Conference on Light and Light based Technologies* (ICLLT) 2016, Tezpur, India.
4) Sumit Datta and Bhabesh Deka, " Interpolated Compressed Sensing for Calibrationless Parallel MRI Reconstruction" in *Twenty fifth National Conference On Communications* (NCC) 2019, Bangalore, India.

Early Diagnosis of Osteoporosis Using Metacarpal Radiogrammetry and Texture Analysis

Anu Shaju Areeckal[a], Sumam David S.[a], and Michel Kocher[b]

[a]Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, Karnataka, India

[b]Department of Industrial Technologies, Haute Ecole dIngenierie et de Gestion du Canton de Vaud (HEIG-VD), Yverdon-les-Bains, Switzerland

I. Introduction

Osteoporosis is a disease characterized by reduction in bone mass and micro-structural connectivity of bone, leading to increased risk to fragility fracture. Osteoporotic fractures affect one in three women and one in five men over the age of 50 [1]. It usually occurs at the lumbar spine, hip and wrist.

II. Background

The gold standard technique used for diagnosis of osteoporosis is determination of Bone Mineral Density (BMD) using Dual Energy X-ray Absorptiometry (DXA) [2]. According to the standard deviation of the measured BMD value from the reference mean BMD of young adult Caucasian women, also called T-score, bone loss is classified as normal, osteopenic, osteoporotic and severely osteoporotic. Quantitative Computed Tomography (QCT) and Quantitative Ultrasound (QUS) are other commonly used bone densitometers. QCT measures the volumetric trabecular bone density. QUS measures the differential reflections and attenuation of sound waves through bone. Digital X-ray Radiogrammetry (DXR) is a low cost method that measures bone density using automated radiogrammetry of the metacarpal bones from hand radiograph [3].

Although DXA is accurate and precise, it is expensive and not widely available in developing countries like India. QCT is expensive and has a high radiation exposure. QUS is cheap, but has low precision and reproducibility. DXR is not widely available in India. DXR measures only the properties of the cortical bone. Trabecular bone, being sensitive to changes due to osteoporosis, would help to improve the accuracy of the diagnostic technique. Hence, there is a need for a low cost and accurate technique for early diagnosis of osteoporosis.

III. Methodology

The aim of the research work is to develop a low cost prescreening tool for early diagnosis of osteoporosis using hand and wrist radiographs. Cortical radiogrammetric measurements of the third metacarpal bone is combined with trabecular texture features of the distal radius to train classifiers for diagnosis of people with low bone mass (T-score< -1). A graphical abstract of the proposed methodology is shown in Figure 1. A low cost technique is also proposed to measure the metacarpal cortical bone volume by using three dimensional (3D) reconstruction of metacarpal from hand radiographs in just three views.

An automatic technique for segmentation of third metacarpal bone from hand and wrist radiographs is proposed using intensity profiles and marker-controlled watershed segmentation. The radiographic images are preprocessed to remove noise and illumination variations. Markers are determined using automatically detected anatomical landmarks and watershed method is applied to detect the outer and inner bone edges of the third metacarpal bone. Cortical radiogrammetric features such as combined cortical thickness (CCT), percent cortical area (PCA), Barnett Nordin index (BNI), etc. are measured from the third metacarpal bone shaft.

The distal radius region-of-interest (ROI) is segmented using automatically detected anatomical landmarks and intensity profile. The largest square inscribed within the circular distal radius is used for texture analysis. Texture features of the trabecular bone ROI is obtained from histogram, Gray Level Co-occurrence matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Morphological Gradient Method (MGM), Laws's masks, Gabor filter, Local Binary Pattern (LBP) and Local Line Binary Pattern (LLBP) [4]–[9].

Significant features are determined using independent sample t-test and Pearson correlation. The significant features are divided into different feature sets and used to train classifiers, namely Artificial Neural Network (ANN), logistic regression classifier, Support Vector Machine (SVM) and k-Nearest Neighbour (KNN). Classifiers are trained using combined cortical and texture features for two sample population, Indian and Swiss. Data of 138 subjects from Indian sample population and 65 subjects from Swiss sample population are analyzed.

A low cost technique to measure the cortical volume of the metacarpal bone shaft using 3D reconstruction from hand X-ray images in three views, namely Postero-Anterior (PA) view, 45^0 and 135^0 oblique views, is proposed. The methodology is similar to the 3D reconstruction of femur bone using two biplanar views, proposed by [10]. The Computed Tomography (CT) scan of one subject is used to create a template model, from which subject-specific models of other people are reconstructed. The 3D reconstruction is done iteratively by registration of the projection and X-ray contours using Iterative Closest Point (ICP) and Self-Organizing Map (SOM), and deformation of the template model using Laplacian deformation. The outer and inner bone walls of the metacarpal are modeled separately and the cortical bone shaft is extracted. Cortical volumetric measurements of the third metacarpal bone shaft are determined.

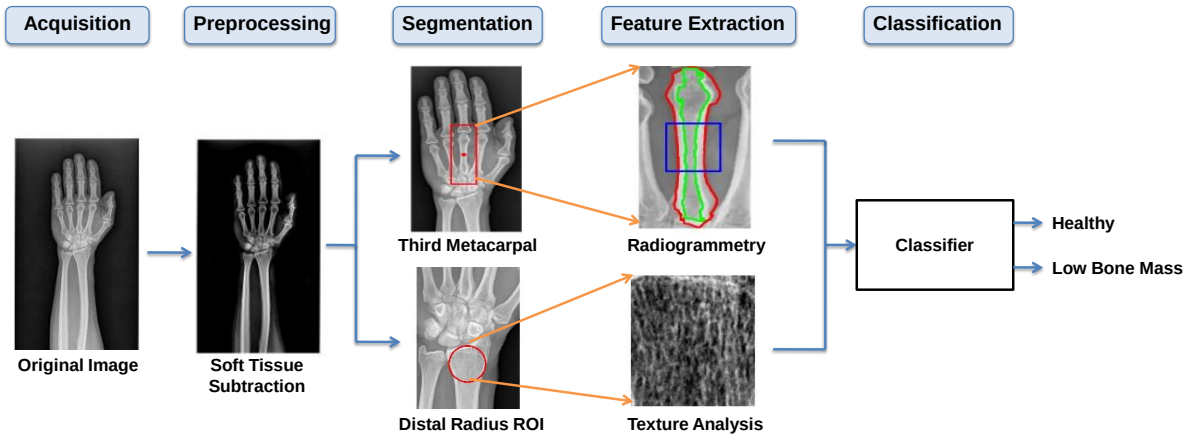


Fig. 1. Graphical abstract of the proposed methodology

IV. Simulation Results

The proposed segmentation method accurately detected the third metacarpal bone in 89% of Indian sample data and 78% of Swiss sample data. The proposed segmentation method shows better performance than segmentation using Active Appearance Model (AAM), when compared to manual segmentation of five images and ground truth radiogrammetric measurements of 14 images. The extracted cortical radiogrammetric features showed high discrimination ability in the healthy and low bone mass groups of both Indian and Swiss sample data. The cortical features are also significantly correlated with DXA-BMD of the lumbar spine.

The proposed segmentation method extracted the distal radius ROI in 93.5% of Indian sample data and 83% of Swiss sample data. In the Indian sample population, majority of the extracted texture features are significant with t-test and well-correlated with DXA-BMD of lumbar spine using Pearson correlation. The most significant cortical and texture features of GLRLM and Laws's masks are selected to train classifiers on Indian sample data. ANN classifier trained using holdout validation achieves test accuracy of 90.0%. A 10-fold cross validation using KNN achieves an accuracy of 81.7%. The linear regression model developed with the cortical and texture features achieves a significant correlation of 0.671 with DXA-BMD. Classifiers are also trained separately for Indian and Swiss sample population. ANN classifiers trained with cortical, histogram, GLCM and MGM features show test accuracy of 92.9% with Indian data and 90.9% with Swiss data. Weighted KNN shows test accuracy of 96.2%. Classifiers trained on LBP features and its variants did not show a good performance with test data.

The 3D reconstructed bone model of one subject is compared to its ground truth and relative volume error and P2S error is determined. The projections of the 3D reconstructed models are compared with manually segmented X-ray images and the mean error percentage in CCT error shows 11.18%.

The proposed method was developed using a small dataset due to limited funding sources. A larger dataset would help to improve the accuracy of the prescreening tool.

V. Conclusion

The proposed low cost prescreening tool can be used to identify people with low bone mass using hand and wrist radiographs. The combination of cortical radiogrammetric and trabecular texture features helps to improve the diagnostic ability of the tool.

This work is done in collaboration with Kasturba Medical College Hospital, Mangalore, India and University Hospital of Geneva, Switzerland.

References

[1] International Osteoporosis Foundation, "Asia-Pacific regional audit on epidemiology, costs and burden of osteoporosis in 2013," IOF Regionals 4^{th} Asia-Pacific Osteoporosis Meeting, Hong Kong, Tech. Rep., 2013.

[2] J. A. Kanis, E. V. McCloskey, H. Johansson, A. Oden, L. J. Melton, and N. Khaltaev, "A reference standard for the description of osteoporosis," *Bone*, vol. 42, no. 3, pp. 467–475, 2008.

[3] A. Rosholm, L. Hyldstrup, L. Baeksgaard, M. Grunkin, and H. Thodberg, "Estimation of bone mineral density by digital X-ray radiogrammetry: Theoretical background and clinical testing," *Osteoporosis International*, vol. 12, no. 11, pp. 961–969, 2001.

[4] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[5] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172–179, 1975.

[6] J. F. Veenland, "Texture analysis of the radiographic trabecular bone pattern in osteoporosis," Ph.D. dissertation, Erasmus University Rotterdam, Rotterdam, Netherlands, 1999.

[7] K. I. Laws, "Textured image segmentation," University of Southern California, Los Angeles, California, Tech. Rep., 1980.

[8] M. Kuse, Y.-F. Wang, V. Kalasannavar, M. Khan, and N. Rajpoot, "Local isotropic phase symmetry measure for detection of beta cells and lymphocytes," *Journal of Pathology Informatics*, vol. 2, 2011.

[9] A. Petpon and S. Srisuk, "Face recognition with local line binary pattern," in *Proceedings of 5^{th} International Conference on Image and Graphics (ICIG)*. IEEE, 2009, pp. 533–539.

[10] V. Karade and B. Ravi, "3D femur model reconstruction from biplane X-ray images: a novel method based on Laplacian surface deformation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 4, pp. 473–485, 2015.

LIST OF PUBLICATIONS

International Journals

1) Anu Shaju Areeckal, Nikil Jayasheelan, Jagannath Kamath, Sophie Zawadynski, Michel Kocher, and Sumam David S., "Early diagnosis of osteoporosis using radiogrammetry and texture analysis from hand and wrist radiographs in Indian population", *Osteoporosis International, Springer Nature*, vol. 29, no. 3, pp. 665-673, 2018, DOI: 10.1007/s00198-017-4328-1, SCI/WoS/Scopus indexed.
2) Anu Shaju Areeckal, Jagannath Kamath, Sophie Zawadynski, Michel Kocher and Sumam David S., "Combined radiogrammetry and texture analysis for early diagnosis of osteoporosis using Indian and Swiss data", *Computerized Medical Imaging and Graphics (CMIG), Elsevier*, vol. 68, no. 9, pp. 25-39, 2018, DOI: 10.1016/j.compmedimag.2018.05.003, SCIE/WoS/Scopus indexed.
3) Anu Shaju Areeckal, Michel Kocher and Sumam David S., "Current and emerging diagnostic imaging-based techniques for assessment of osteoporosis and fracture risk", *IEEE Reviews in Biomedical Engineering (RBME)*, vol. 12, no. 1, pp. 1, 2019, DOI: 10.1109/RBME.2018.2852620, Scopus indexed.

International Conferences

1) Anu Shaju Areeckal, Sumam David S., Michel Kocher, Nikil Jayasheelan, and Jagannath Kamath, "Fully automated radiogrammetric measurement of third metacarpal bone from hand radiograph", *11^{th} IEEE International Conference on Signal Processing and Communication (SPCOM)*, Bangalore, India, pp. 1-5, June 12-15, 2016, DOI: 10.1109/SPCOM.2016.7746608, WoS/Scopus indexed.
2) Mathew Sam, Anu Shaju Areeckal, and Sumam David S., "Early diagnosis of osteoporosis using active appearance model and metacarpal radiogrammetry", *13^{th} IEEE International Conference on Signal Image Technology and Internet Based Systems (SITIS)*, Jaipur, India, pp. 173-178, December 4-7, 2017, DOI: 10.1109/SITIS.2017.38, Scopus indexed.
3) Anu Shaju Areeckal, Mathew Sam, and Sumam David S., "Computerized radiogrammetry of third metacarpal using watershed and active appearance model", *19^{th} IEEE International Conference on Industrial Technology (ICIT)*, Lyon, France, pp. 1490-1495, February 20-22, 2018, DOI: 10.1109/ICIT.2018.8352401, Scopus indexed.
4) Avinash D. Jayakar, Gautham Sambath, Anu Shaju Areeckal and Sumam David S., "Cortical volumetry using 3D reconstruction of metacarpal bone from multi-view images", *4^{th} IEEE International Conference on Recent Advances in Computational Systems (RAICS)*, Kerala, India, December 6-8, 2018.

Large Scale Twin Support Vector Machine and its applications in Human Activity Recognition

Sweta Sharma
Department of Computer Science
South Asian University
New Delhi, India

Human Activity Recognition (HAR) is an active field of research in computer vision that deals with how computers can be made to gain a high-level understanding of activities from digital images or videos or data acquired through sensors. These systems are crucial for intelligent environment where they find extensive applications in surveillance systems, human-computer interaction, video analysis and annotations etc. The key challenges include the presence of extreme noises due to varying illumination, occlusion and intraclass differences among activities. Moreover, as vocabulary of the possible activities is extremely large, it leads to need for large training data, which makes the recognition task even more challenging. Beside this, learning better representation from training videos is yet another major challenge owing to the similarities between different activities. Traditionally, maximum margin based Support Vector Machine (SVM) [1] is used as an effective classification tool to deal with the recognition task in HAR systems. However, SVM suffers from the limitations of sensitivity to noise and high training time complexity as it requires $O(n^3)$ training time where n denotes the number of training samples. Over the past few decades, various improvements to traditional SVM have been suggested, such as Lagrangian Support Vector Machine (LSVM), Least Squares Support Vector Machine (LS-SVM) and Proximal Support Vector Machine (PSVM). One of the major breakthroughs among these is Twin Support Vector Machine (TWSVM) [3] which seeks two non-parallel hyperplanes that are proximal to their respective classes while simultaneously being at least a unit distance away from the other class. TWSVM exhibits better generalization performance and is almost four times faster than SVM [3]. The motivation behind this research work is to explore existing machine learning algorithms based on TWSVM and to develop new ones which could deliver better results than well-established methodologies while handling the key challenges of human activity recognition systems.

Usually, the features extracted from the sequence of activity videos are contaminated by the presence of inter-class noise between related activity classes along with high training and testing time complexity of the system. Our first approach to address these issues is by introducing a novel algorithm termed as Robust Least Squares Twin Support Vector Machine (RLS-TWSVM) which handles the heteroscedastic noise and outliers present in activity recognition framework by replacing the unit distance separation criteria in TWSVM by a self-optimizing parameter ρ, that is trained as a part of the optimization problem itself. This accords the model the flexibility to adjust the class representative hyperplanes according to the distribution of noise in training data. Further, to deal with the large size of training data, we propose an incremental learning approach for RLS-TWSVM. Also, we introduce the hierarchical framework with RLS-TWSVM to deal with multi-category activity recognition problem. In our subsequent work, we improved upon this model by finding a pair of parametric margin hyperplanes that automatically adjusts the parametric insensitive margin of the model as a function of training data which enables the learning model to capture non-uniformly distributed noise. The resulting model is termed as Robust Parametric Twin Support Vector Machine (RPTWSVM). RPTWSVM optimizes the model parameters by incorporating the noise information through a function rather than a fixed parameter.

Since different humans can perform the same actions differently leading to high intra-class variation thus, the activity recognition systems generally need a substantial amount of training data. Although TWSVM based classifiers are faster than conventional SVMs, these are not able to scale up to handle very large number of data samples during training as solving the corresponding quadratic optimization problem become infeasible. In contemplation of this drawback, we propose an efficient stochastic Quasi-Newton method based Twin Parametric Support Vector Machine termed as SQN-PTWSVM. Further, as many studies reveal that Hinge loss function does not only make the problem non-smooth leading to instability near the minima but also propagate noise-sensitivity to the problem which makes the stochastic optimization process deviate near the point of non-differentiability, we propose to use Pinball loss function in SQN-PTWSVM to minimize the training loss, which is also efficient and more robust to the presence of noise when compared to conventional hinge loss SVM for large-scale data scenarios. The resulting SQN-PTWSVM can effectively handle millions of training points in HAR framework. Unlike SG-TSVM [4], which has the limited competence to handle feature noise, SQN-PTWSVM is insensitive to noise, robust to re-sampling and leads to rapid convergence of the corresponding convex optimization problem which speeds up the training process as well.

As mentioned above, excellent capability of Pinball error

loss in handling feature noise present in data eventually leads to robust classification models. However, the main focus to ensure robustness to noise and outliers has led to limited sparsity in the classifier(s) which result in the increased testing time. In HAR framework, large testing time hinders the applicability of these systems in real-world scenarios. Thus, in our subsequent work, we develop two robust Pin-TWSVM based models termed as $\epsilon-TWSVM$ and $Flex-Pin-TWSVM$ that not only handle the noise present in data but also control model sparsity using user regulated and self-optimized insensitive zone respectively. The insensitive zone obtained in the model corresponds to the zero value of the dual variable and thus lead to simpler and sparser model.

Proceeding onto our subsequent works, we observed that Human activity recognition systems, like many other machine learning problems such as Content-Based Image Retrieval Systems, usually poses yet another interesting challenge. Although the acquisition of hundreds of hours of training videos is indeed an easy task but obtaining corresponding training label is difficult, tedious or sometimes just boring. In such cases, usually semi-supervised learning approaches such as Laplacian SVM [5], requiring limited label information, are employed. However, conventional semi-supervised learning approaches often have no explicit control over the choice or usefulness of labeled data available for training, hence to overcome this limitation, we propose a pool-based active learning framework using a fast semi-supervised classifier model termed as Fast Laplacian Twin Support Vector Machine ($FLap-TWSVM$) which identifies most informative examples to train the learning model. The resulting active learning framework $FLap-TWSVM_{AL}$ is faster than existing Laplacian twin support vector machine as it solves a smaller sized Quadratic Programming Problem (QPP) along with an Unconstrained Minimization Problem (UMP) to obtain decision hyperplanes which can also handle heteroscedastic noise present in the training data. Moreover, the aforementioned framework has been extended to deal with multi-activity recognition scenarios.

Following our approach to use limited labeled data information, we next dealt with activity recognition problem in an unsupervised manner. As the information in activity video sequences is distributed spatially in the form of two-dimensional matrices (elements of second-order tensor space), traditional vector-based approaches rely on low-dimensional features representation for identifying patterns and are thus prone to loss of useful information which is present in the spatial structure of the data. Hence, we propose a novel clustering framework, termed as Ternary Treebased Structural Least Squares Support Tensor Clustering (TT-SLSTWSTC), that builds a cluster model as a hierarchical ternary tree, where at each node non-ambiguous data is dealt separately from ambiguous data points using the proposed Ternary Structural Least Squares Support Tensor Machine (TS-LSTWSTM). The TS-LSTWSTM classifier considers the structural risk minimization of data alongside a symmetrical L2-norm loss function. The proposed clustering model helps in identifying

Dataset	RLS-TWSVM	SVM	Random forest classifier
Weizmann	100	100	100
UIUC1	99.8	98.3	
IXMAS	84.9	81.2	80.55

TABLE I
COMPARISON OF THE PERFORMANCE OF RLS-TWSVM CLASSIFIER ON FOUR STANDARD HAR DATASETS TO THE STATE OF ART METHODS

Prediction Accuracy	Classifier (%)	Type	Year
k-NN	89.66	Supervised	2008
k-NN	92.30	Supervised	2014
LS-TWSVM	85.56	Supervised	2014
RL-TWSVM	100	Supervised	2016
TWSVM	88.89	Supervised	2007
$TWSVM_{AL}$	82.22*	Semi-supervised+AL	2018
$FLap-TWSVM_{AL}$	93.33 *	Unupervised+AL	2018

* Less than 50% labeled sequences were used.

TABLE II
ACCURACY RATES OF DIFFERENT METHODS ON WEIZMANN DATASET

relevant patterns in matrix data as they take advantage of structural information present in multi-dimensional framework and reduce computational overheads as well. Also, initialization framework based on tensor k-means is used in order to overcome the instability disseminated by random initialization.

To compare the performance of our proposed models with other related state-of-art methods, we performed extensive experiments with standard machine learning datasets along with benchmark human activity recognition datasets. Moreover, various experiments were carried out on Content based Image Retrieval Problem, Handwritten digit recognition etc for large scale stochastic and tensor-based models. The experimental comparisons of our noise-robust RLSTWSVM on well-known activity recognition datasets (Refer Table I) show the outperformance of our proposed methods in terms of significantly better generalization performance and its capability to handle heteroscedastic noise and outliers. Moreover, the prediction accuracies of $FLap-TWSVM_{AL}$ with different supervised and semi-supervised classifiers have been compared in Table II. Results shows that using only most relevant and informative labeled data through $FLap-TWSVM_{AL}$ can improve the label requirements as well as training complexity of the resulting system when compared to traditional semi-supervised [5] and active learning [6] approaches.

Although our proposed techniques yield effective results, we observed that in hierarchical approaches to deal with the multi-category classification of activity classes suffered due to error propagation to lower nodes which was, however, handled with proper parameter tuning to some extent. Also, finding the kernel version of proposed algorithm was also a challenge in some of the above-mentioned methods as size of the kernel matrix become huge and it is computational complex task to tune the kernel parameter. This issue was taken care of through the use of rectangular kernel trick and data-driven kernels but more insightful analysis is still required. In future, we aim to improve our existing approach to deal with large size datasets by using low-dimensional embeddings for kernel matrices, thereby reducing the memory requirements and time requirements further. For this purpose, lower-order approximation methods can be incorporated in the optimization framework only. In addition to this, rather than

considering strict activity labels for each video frame, soft labels can be considered so that ambiguous activity poses can be assigned a set of related labels rather than an arbitrary one.

REFERENCES

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[2] L. Wang, Support vector machines: theory and applications. Springer Science & Business Media, 2005, vol. 177.

[3] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 5, pp. 905–910, 2007.

[4] Z. Wang, Y.-H. Shao, L. Bai, L.-M. Liu, and N.-Y. Deng, "Insensitive stochastic gradient twin support vector machines for large scale problems," *arXiv preprint arXiv:1704.05596*, 2017.

[5] Z. Qi, Y. Tian, and Y. Shi, "Laplacian twin support vector machine for semi-supervised classification," *Neural Networks*, vol. 35, pp. 46–53, 2012.

[6] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

I. LIST OF PUBLICATIONS

1. R. Rastogi and S. Sharma, "Fast Laplacian Twin Support Vector Machine with Active Learning for Pattern Classification," *Applied Soft Computing*, vol. 74, pp. 424-439, 2019.
2. S. Sharma and R. Rastogi, "Stochastic Conjugate Gradient Descent Twin Support Vector Machine for Large Scale Pattern Classification," In *Australasian Joint Conference on Artificial Intelligence*, pp. 590-602. Springer, Cham, 2018.
3. S. Sharma and R. Rastogi, "Maximum Margin Minimum Variance Twin Support Vector Machine for Pattern Classification," In *IEEE-Symposium Series on Computational Intelligence*, IEEExplore, IEEE, 2018.
4. S. Sharma and R. Rastogi, "Insensitive Zone based Pinball Loss Twin Support Vector Machine for Pattern Classification,"In *IEEE-Symposium Series on Computational Intelligence*, IEEExplore, IEEE, 2018.
5. R. Rastogi and S. Sharma, "Tree-Based Structural Twin Support Tensor Clustering with Square Loss Function," *Proceedings of Pattern Recognition and Machine Intelligence*, pp. 28-34, 2017.
6. R. Rastogi, S. Sharma, and S. Chandra, "Robust Parametric Twin Support Vector Machine for Pattern Classification," *Neural Processing Letters*, May 2017.
7. R. Khemchandani and S. Sharma, "Robust Parametric Twin Support Vector Machine and Its Application in Human Activity Recognition," *Proceedings of International Conference on Computer Vision and Image Processing*, pp. 193-203, Dec. 2016.
8. R. Khemchandani and S. Sharma, "Robust least squares twin support vector machine for human activity recognition," *Applied Soft Computing*, vol. 47, pp. 33-46, Oct. 2016.

Multidimensional Reflection Symmetry: Theory, Algorithms, and Applications

Rajendra Nagar, Electrical Engineering, IIT Gandhinagar
Supervisor: Dr. Shanmuganathan Raman, Associate Professor, IIT Gandhinagar

ABSTRACT

An object is called symmetric if it can be partitioned into more than one visually identical segments. Symmetry present in natural and man-made objects enriches the objects to be physically balanced, beautiful, easy to recognize, and easy to understand. The real world objects are stored in the computer memory mostly as digital images, point clouds obtained through devices such as Kinect and LiDAR, and triangle meshes. Characterizing and finding the symmetry has been an active topic of research in computer vision and computer graphics as physical objects form the basis for these research areas. Our main motivation of detecting symmetry is due to its importance in solving problems such as shape matching, retrieving the normal forms of objects, shape recognition, model compression, reconstruction of 3D models from images, real-time attention for robotic vision, and cultural heritage object reconstruction and restoration.

There exist approaches for reflection symmetry detection in digital images, point clouds, and triangle meshes. However, the performance of the existing approaches on the real world datasets still has to be improved for better solving the above important problems due to impurities added in the data while capturing, such as occlusion, non-rigid deformations, and sensor noise. Our goal is to develop algorithms for detecting reflection symmetries in digital image, point cloud, and triangle mesh in the presence of noise, occlusions, and non-rigid deformations. We have designed algorithms for the following problems.

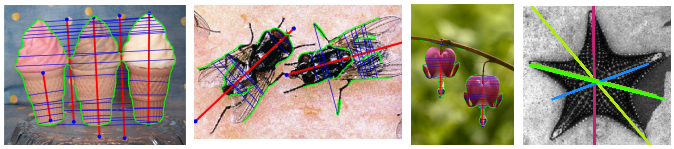


Fig. 1: Symmetry detection results on the dataset [4].

Multiple Symmetry Axes Detection in Real World Images. We find reflection symmetry axes of symmetric objects present in an image based on their salient feature points using multiple model fitting framework [iv] and boundary orientations using a graph embedding approach [vii]. We achieve state-of-the-art performance for multiple symmetry axes detection using the boundary orientation approach. The F-score for Elawady *et al.* [1] is 0.21 , Loy and Eklundh [3] is 0.32, the proposed multiple model fitting approach is 0.20, and the boundary orientation based approach is 0.37. In Fig. 1, we show some results.

SymmMap: Estimation of the 2-D Reflection Symmetry Map. We propose an approach to find the reflection symmetry map of an image [iii]. The reflection symmetry map represents a confidence score at every pixel having its mirror reflection pixel. The reflection symmetry map also identifies the mirror reflection pixel of each pixel. In Fig. 2, we show an example.

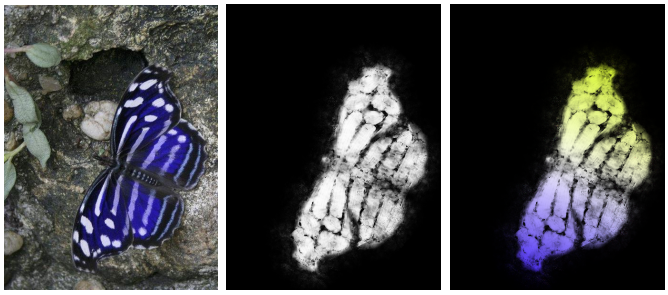


Fig. 2: For an image, the estimated reflection symmetry score map and the nearest mirror reflection field.

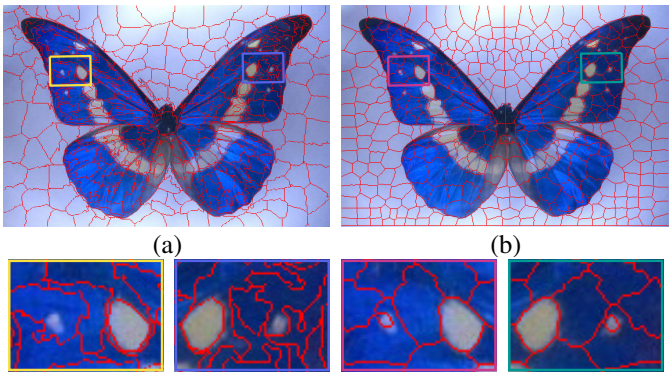


Fig. 3: (a) The result using [5], (b) result using the proposed SymmSLIC, and (c) zoomed mirror symmetric windows (top: MSLIC, bottom: ours).

SymmSLIC: Symmetry Aware Superpixel Segmentation. We introduce the concept of symmetry aware oversegmentation of digital images into superpixels [ii]. Oversegmentation of an image is a useful tool for solving various problems in computer vision. Reflection symmetry is an essential cue in understanding and grouping the objects in natural

scenes. Existing algorithms for estimating superpixels do not preserve the reflection symmetry of an object which leads to different sizes and shapes of superpixels across the symmetry axis. We propose an algorithm to over-segment an image through the propagation of reflection symmetry evident at the pixel level to the superpixel boundaries. Further, we show the importance of symmetry aware superpixel over-segmentation by using it to unsupervised symmetric object segmentation. In Fig. 3, we show a result compared with [5].

Global 3D Extrinsic Symmetry Detection in Point Clouds. We present two methods for detecting symmetry in 3D point clouds ([viii] and [v]). The first approach detects the symmetry without using any feature descriptors. The F-score for the approaches in [2], [6], [7], and the proposed approach on the dataset [4] are 0.67, 0.83, 0.73, and 0.84, respectively. In Fig. 4, we show some results. The second approach uses images used in a structure from motion framework for constructing the point cloud of the underlying object and thereby detecting the symmetry plane.

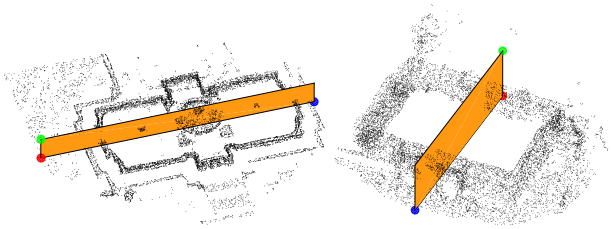


Fig. 4: Detected reflection symmetry plane using our approach in two real scans from the dataset in [4].

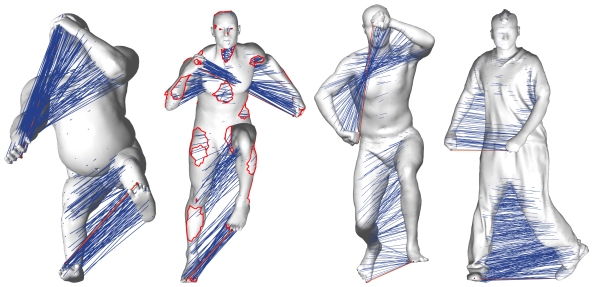


Fig. 5: Results of our approach on the TOSCA [8], SHREC16 [9], SCAPE [10], and Articulated [11] datasets.

Fast and Accurate Intrinsic Symmetry Detection. Intrinsic symmetry of 3D objects (represented as triangle meshes) is used as a prior in various important problems of shape analysis. Therefore, off-the-shelf fast and accurate detection of symmetry in challenging settings is required. Previous approaches are time-consuming and require improvement in accuracy on challenging datasets. We propose a functional map based approach [i]. In Fig. 5, we show some results. The computation time for the approaches in, [13], [14], [15], and the proposed approach on the SCAPE dataset [10] are 360, 60, 24, 8 minutes, respectively. The correspondences rate for the approaches in [12], [13], [14], [15], and the proposed approach on the SCAPE dataset [10] are 82%, 84.8%, 91.7%, 94.5%, and 97.5%, respectively.

Approximate Reflection Symmetry in d-Dimensional Point Set. The motivation behind detecting symmetry in higher dimensional spaces ($d > 3$) is inspired by the fact that many physical data points reside in the space of dimensions greater than three. For example, an RGB-D image captured using a Kinect sensor, which has become a major tool for the interaction of the human with the machines, has four dimensions at each pixel location. We develop the theory and present a method for detecting the reflection symmetry in d-dimensional point cloud without using feature descriptors [vi].

References

[1] Mohamed Elawady, Christophe Ducottet, Olivier Alata, Cécile Barat, and Philippe Colantoni. Wavelet-based reflection symmetry detection via textural and color histograms: Algorithm and results. In *IEEE ICCV Workshop*, pages 1734–1738. 2017.

[2] Marcelo Cicconet, David GC Hildebrand, and Hunter Elliott. Finding mirror symmetry via registration and optimal symmetric pairwise assignment of curves: Algorithm and results. In *IEEE ICCV Workshop*, pages 1759–1763. 2017.

[3] Gareth Loy and Jan-Olof Eklundh. Detecting symmetry and symmetric constellations of features. In *ECCV* 2006.

[4] Christopher Funk, Seungkyu Lee, Martin R. Oswald, Stavros Tsogkas, Wei Shen, Andrea Cohen, Sven Dickinson, and Yanxi Liu. 2017 iccv challenge: Detecting symmetry in the wild. In *IEEE ICCV Workshops*, 2017.

[5] Yong-Jin Liu, Cheng-Chi Yu, Min-Jing Yu, and Ying He. Manifold slic: A fast method to compute content-sensitive superpixels. In *IEEE CVPR*, pages 651–659, 2016.

[6] Aleksandrs Ecins, Cornelia Fermuller, and Yiannis Aloimonos. Detecting reflectional symmetries in 3d data through symmetrical fitting. In *IEEE ICCV Workshops*, Oct 2017.

[7] Pablo Speciale, Martin R Oswald, Andrea Cohen, and Marc Pollefeys. A symmetry prior for convex variational 3d reconstruction. In *ECCV*, pages 313–328. 2016.

[8] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.

[9] L Cosmo, E Rodolà, MM Bronstein, A Torsello, D Cremers, and Y Sahillioglu. Shrec16: Partial matching of deformable shapes. *Proc. 3DOR*, 2(9):12, 2016.

[10] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24, pages 408–416. ACM, 2005.

[11] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)*, volume 27, page 97. ACM, 2008.

[12] Vladimir G Kim, Yaron Lipman, Xiaobai Chen, and Thomas Funkhouser. Möbius transformations for global intrinsic symmetry analysis. In *Computer Graphics Forum*, volume 29, pages 1689–1700. Wiley Online Library, 2010.

[13] Vladimir G Kim, Yaron Lipman, and Thomas Funkhouser. Blended intrinsic maps. In *ACM Transactions on Graphics (TOG)*, volume 30, page 79. ACM, 2011.

[14] Xiuping Liu, Shuhua Li, Risheng Liu, Jun Wang, Hui Wang, and Junjie Cao. Properly constrained orthonormal functional maps for intrinsic symmetries. *Computers & Graphics*, 2015.

[15] Hui Wang and Hui Huang. Group representation of global intrinsic symmetries. In *Computer Graphics Forum*, volume 36, pages 51–61. Wiley Online Library, 2017.

PUBLICATIONS

Following publications are the outcomes of this thesis.

In Peer-Reviewed Conferences and Journals

[i] Rajendra Nagar and Shanmuganathan Raman. "Fast and Accurate Intrinsic Symmetry Detection", in European Conference on Computer Vision (ECCV), Munich, Germany, Sep. 8-14, 2018.

[ii] Rajendra Nagar and Shanmuganathan Raman. "SymmSLIC: Symmetry aware superpixel segmentation ". In IEEE ICCV Workshop on Detecting Symmetry in the Wild, Venice, Italy, 28 Oct 2017.

[iii] Rajendra Nagar and Shanmuganathan Raman, "SymmMap: Estimation of the 2-D Reflection Symmetry Map and its Applications". In IEEE ICCV Workshop on Detecting Symmetry in the Wild, Venice, Italy, 28 Oct 2017.

[iv] Rajendra Nagar and Shanmuganathan Raman, "Reflection Symmetry Axes Detection using Multiple Model Fitting", In IEEE Signal Processing Letters, Pages 1438-1442, Volume 24, Issue 10, 2017.

[v] Rajendra Nagar and Shanmuganathan Raman, "Revealing Hidden 3D Reflection Symmetry", In IEEE Signal Processing Letters, Pages 1776-1780, Volume 23, Issue 12, 2016.

[vi] Rajendra Nagar and Shanmuganathan Raman, "Approximate Reflection Symmetry in a Point Set: Theory, Algorithm, and Applications", IEEE Transactions on Signal Processing. [Under Revision]

[vii] Rajendra Nagar and Shanmuganathan Raman, "Embedding Symmetry as Cliques in a Graph", In ICASSP 2019. [Under Review]

[viii] Rajendra Nagar and Shanmuganathan Raman, "Robust 3D Reflection Symmetry Detection", IEEE Transactions on Visualization and Computer Graphics. [Under Review]

Low-Complexity Distributed Arithmetic based Pipelined VLSI Architectures for LMS Adaptive Filters

Mohd Tasleem Khan[1], and Shaik Rafi Ahamed[2]
Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
tasleem@iitg.ac.in[1], rafiahamed@iitg.ac.in[2]

I. ABSTRACT

Least-mean-square (LMS) algorithm is widely used in system identification, channel equalization, noise cancellation, and several other areas of digital signal processing due to its simplicity and ease of implementation. In most of the cases, LMS algorithm is not employed directly, rather a combination of LMS units in parallel or in series or in block is used to obtain the desired performance. For instance, the combination of two parallel LMS adaptive filters with different step-size is believed to have fast convergence and low steady state error, the series combination of two LMS adaptive filters connected in feedforward and feedback topology is believed to provide better performance against noise and interference and the block implementation of LMS adaptive filter is used to realize high order filter. Due to ubiquitous applications, LMS filters have several implementation complexity issues which need to be addressed. One of them is that its order primarily determines the number of multipliers. For high-throughput applications, the number of multipliers required for the implementation would be very high, while the critical path has to be made shorter, thus the real-time implementation of such filters for higher order is a challenging task. Pipelining is, therefore, necessary for LMS adaptive filter to achieve the throughput requirements, but it has an adverse effect on the convergence rate and steady-state error. Thus, obtaining low-complexity LMS adaptive filters with good convergence performance without compromising the throughput is an interesting area of research. Distributed arithmetic (DA) is an efficient multiplierless approach for implementation of LMS adaptive filter. Such implementations enable to realize LMS adaptive filters with low computational cost and high throughput. DA based implementations basically consist of a look-up table (LUT) followed by a shift-accumulate (SA) unit. But, the direct usage of DA for complexity reduction of adaptive filter, especially in high-throughput applications would be challenging, since time required to access LUT and to compute SA unit is significant. The modularity feature of DA makes it amenable for implementing on field-programmable gate arrays (FPGA) and the design of application specific integrated circuit (ASIC).

In this thesis, we first derived the optimal complexity pipelined architectures for LMS adaptive filter using offset-binary-coding (OBC) DA. Although, it is straightforward to pre-compute and store the filter partial products in LUT for the realization of non-adaptive filter such as FIR filter using DA, problem arises while generating them using hardware elements to overcome the access time of LUT. Since the number of hardware elements required to generate the partial products for LUT-less grow exponentially with filter order. For this purpose, we have implemented partial products of input samples serially by representing the filter coefficients in OBC-form. But, this produces non-OBC terms at the output during some initial clock cycles due to pipelined nature of filter, which are subsequently corrected in the error computation unit. The reason for choosing OBC scheme in the implementation of pipelined LMS adaptive filter is to exploit the redundancies between the partial products for high-radix applications. This is because high-radix DA reduces the time involved in SA unit for the computation of filter output. The effect of non-OBC terms are separately studied through adaptive equalization problem. Next, we applied two's complement (TC) DA treatment to the pipelined realization of convex combination of two parallel LMS adaptive filters which greatly improves the convergence performance. Unlike conventional LMS algorithm, the convex combination of two LMS filter can provide fast convergence and low steady-state error by transferring the filter coefficients from one LMS unit to other. However, the computational requirements are significantly higher due to two LMS units and coefficient transfer scheme. For this purpose, we employed a single DA based LMS adaptive filter with a new coefficient transfer scheme. Further reduction in computational complexity is achieved by sharing the partial products and employing bit-level coefficient update accumulators. The reduction in computational elements is utilized for transferring the filter coefficients by switching the step-size. As a result, it enhances the convergence properties with an efficient criterion through which the coefficients can be transferred from one LMS unit to another. Based on this, we exploited the correlation between the consecutive delayed error samples and compared their duration with a pre-defined time window. In the sequel,

an analytical expression for pre-defined window is derived in terms of filter parameters such as step-size, filter order and wordlength. Later, we considered low-complexity DA based VLSI implementation of adaptive filter for channel equalization problem in 5G communication system. In this work, we have employed two LMS adaptive filters in series with one in feedforward path and other in feedback path with a decision device. The overall system is designed to achieve throughput requirement of 5G communication system while maintaining computational requirements relatively lower than the best existing design. The design first utilizes the pipelined implementation of non-adaptive feedback filter using OBC. It is based on the observation that when radix-size becomes equal to the wordlength of coefficients, then the implementation of DA based LMS adaptive filter can be made SA-less. In this design, decisions are OBC coded to derive low-complexity SA-less architecture for feedback adaptive filter. The proposed architecture basically pre-computes and stores coefficients in two separate LUTs for complexity reduction. Further reduction in complexity is achieved by exploiting the symmetries between the stored contents of both the LUTs. The proposed architecture is pre-speed up by two, retimed and unfolded to meet the throughput requirements of 5G. To adapt the feedback coefficients, a novel strategy is presented to update the LUT contents of both the stages by adding with the contents of external LUT storing the decisions. The contents of decision LUT are updated before the filtering operation in every iteration. A new approach is presented at architectural-level to improve the convergence performance by employing a parallel error multiplexer. Lastly, a low-complexity hardware design of block pipelined least mean square (BLMS) adaptive filter is proposed for noise cancellation in in-ear headphones applications. Here, both physical LUT and SA unit are employed, and their effect on throughput performance are compensated by block processing. It utilizes OBC-DA for complexity reduction which stores the partial products of input samples in a LUT. The symmetries in LUT contents allowed them to split into two smaller LUTs, but requires the contents of external register to be updated along with the error computation. The splitted LUTs are shared to compute the filter output and coefficient-increment terms in the same block iteration. A novel strategy is developed to update LUT contents in fewer number of clock cycles as compared to the best existing design. To validate the performance of the proposed architectures, ASIC and FPGA implementations have been carried out to estimate area, power, throughput, number of flip-flops and slice LUTs.

II. Publications

A. Journal Publications (Published)

- **M.T. Khan** and **R.A. Shaik**, "Optimal Complexity Architectures for Pipelined Distributed Arithmetic based LMS Adaptive Filter", *IEEE Transaction on Circuits and Systems I*, Sept. 2018, DOI: 10.1109/ TCSI.2018.2867291 [Available Online, Early Access]
- **M.T. Khan** and **R.A. Shaik**, "An Energy Efficient VLSI Architecture of Decision Feedback Equalizer for 5G communication system", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 4, pp. 569-581, Dec. 2017. [Available Online]
- **M.T. Khan**, **R.A. Shaik** and Surya Prakash Matcha, "Improved Convergent Distributed Arithmetic based Low complexity Pipelined Least-Mean-Square Filter", *IET Circuits, Devices & Systems*, Apr. 2018, DOI: 10.1049/iet-cds.2018.0041 [Available Online]
- **M.T. Khan** and **R.A. Shaik** "High-Performance Hardware Design of Block LMS Adaptive Noise Canceller for In-ear Headphones", *IEEE Consumer Electronics Magazine*. [Accepted]

B. Journal Publications (Revision)

- **M.T. Khan** and **R.A. Shaik**, "High-Throughput, Fast-Convergent and Low-Steady State Error Pipelined Architecture of Adaptive DFE", *in IEEE Transactions on Circuits and Systems II*.
- **M.T. Khan** and **R.A. Shaik**, "High-Performance Pipelined Architecture of Distrib-uted Arithmetic based LMS Adaptive Filter", *in IEEE Signal Processing Letters*.
- **M.T. Khan** and **R.A. Shaik**, "Finite Precision Analysis of Two Non-Pipelined Distributed Arithmetic based LMS Adaptive Filters", *in IEEE Transaction on Circuits and Systems I.*

C. Conference Publications

- J. Hazarika, **M. T. Khan** and **R.A. Shaik**, "Low-Complexity Continuous-Flow Memory-Based FFT Architectures for Real-Valued Signals", *32nd International Conference on VLSI Design (VLSID 2019)* [Accepted, **Nominated for Best Paper**].
- **M. T. Khan** and **R.A. Shaik**, "Analysis and Implementation of Block Least Mean Square Adaptive Filter using Offset Binary Coding", *International Symposium on Circuits and Systems (ISCAS)-2018*, Florence, **Italy**.
- **M. T. Khan** and **R.A. Shaik**, "Enhanced Convergence Distributed Arithmetic Based LMS Adaptive Filter Using Convex Combination", *National Conference on Communication (NCC)-2018*, Hyderabad, India.
- **M. T. Khan** and **R.A. Shaik**, "Area and Power Efficient VLSI Architecture of Distributed Arithmetic Based LMS Adaptive Filter", *International Conference on VLSI Design-2018*, Pune, India.
- **M. T. Khan** and **R.A. Shaik**, "VLSI Realization of Low Complexity Pipelined LMS Filter Using Distributed Arithmetic", *International Conference TENCON-2017*, Penang, **Malaysia**.
- **M. T. Khan** and **R.A. Shaik**, "A New High Performance VLSI Architecture for LMS Adaptive Filter using Distributed Arithmetic", *International Symposium on VLSI (ISVLSI) 2017*, Bochum, **Germany**.
- **M. T. Khan** and **R.A. Shaik**, "VLSI Implementation of Throughput Efficient Distributed Arithmetic Based LMS Adaptive Filter", *International Symposium on VLSI design and Test (VDAT)-2017*, Roorkee, India.
- **M. T. Khan** and **R.A. Shaik**, "Forrest Brewer Low Complexity and Critical Path Based VLSI Architecture for LMS Adaptive Filter Using Distributed Arithmetic", *International Conference on VLSI Design-2017*, Hyderabad, India.
- **M. T. Khan** and **R.A. Shaik**, A. Chatterjee. "Efficient Implementation of Concurrent Lookahead Decision Feedback Equalizer using Offset Binary Coding, *International Symposium on VLSI Design and Test VDAT-2016*, Guwahati, India".
- **M. T. Khan** and **R.A. Shaik**, "Low Cost Implementation of Concurrent Decision Feedback Equalizer using Distributed

Arithmetic", *IEEE India International Confe-rence on Information Processing (IICIP)-2016*, Delhi, India.

- P. K. Sharma, **M. T. Khan** and **R.A. Shaik**, "An Alternative Approach To Design Reconfigurable Mixed Signal VLSI DA Based FIR Filter", *IEEE Students Technology Symposium (TechSym)-2016*, Kharagpur, India.

III. Acknowledgement

This work was supported by Special Manpower Development Programme for Chip to System Design (SMDP-C2SD) sponsored by the Ministry of Electronics & Information Technology (MeitY), Govt. of India.

IV. References

1) I. Present, "Cramming more components onto integrated circuits," *Readings in computer architecture*, vol. 56, 2000.
2) L. R. Vega and H. Rey, "A rapid introduction to adaptive filtering", *Springer Science & Business Media*, 2012.
3) S. Rudich and A. Wigderson, "Computational complexity theory", *American Mathematical Soc.*, 2004, vol. 10.
4) D. M. Markovic, "A power/area optimal approach to VLSI signal processing," University of California, Berkeley, 2006.
5) S. Haykin, Adaptive Filter Theory (3rd Ed.). Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.
6) G. Long, F. Ling, and J. G. Proakis, "The LMS algorithm with delayed coefficient adaptation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 9, pp. 13971405, 1989.
7) A. Croisier, D. Esteban, M. Levilion, and V. Riso, "Digital filter for PCM encoded signals," Dec. 4 1973, uS Patent 3,777,130.
8) A. Peled and B. Liu, "A new hardware realization of digital filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 6, pp. 456462, 1974.
9) D. J. Allred, H. Yoo, V. Krishnan, W. Huang, and D. V. Anderson, "LMS adaptive filters using distributed arithmetic for high throughput," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 7, pp. 13271337, 2005.
10) R. Guo and L. S. DeBrunner,"Two high-performance adaptive filter implementation schemes using distributed arithmetic," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 58, no. 9, pp. 600604, 2011.
11) M. S. Prakash and R. A. Shaik, "Low-area and high-throughput architecture for an adaptive filter using distributed arithmetic," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 11, pp. 781785, 2013.
12) P. K. Meher and S. Y. Park, "High-throughput pipelined realization of adaptive FIR filter based on distributed arithmetic," in 2011 *IEEE/IFIP 19th International Conference on VLSI and System-on-Chip. IEEE*, 2011, pp. 428433.
13) S. Y. Park and P. K. Meher, "Low-power, high-throughput, and low-area adaptive FIR filter based on distributed arithmetic," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 6, pp. 346350, 2013.
14) S. P. Matcha, "High performance architectures for adaptive equalizers using distributed arithmetic" Ph.D. dissertation, 2016.
15) R. Shaik, M. Chakraborty, and S. Chattopadhyay, "An efficient finite precision realization of the block adaptive decision feedback equalizer," in *IEEE International Symposium on Circuits and Systems, 2008.* ISCAS 2008. IEEE, 2008, pp. 19101913.
16) J. Arenas-Garca, M. Martnez-Ramon, A. Navia-Vazquez, and A. R. Figueiras-Vidal, "Plant identification via adaptive combination of transversal filters," *Signal Processing*, vol. 86, no. 9, pp. 24302438, 2006.
17) V. H. Nascimento and R. C. de Lamare, A low-complexity strategy for speeding up the convergence of convex combinations of adaptive filters, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012. IEEE, 2012, pp. 35533556.
18) L. Lu, H. Zhao, Z. He, and B. Chen, "A novel sign adaptation scheme for convex combination of two adaptive filters," *AEU-International Journal of Electronics and Communications*, vol. 69, no. 11, pp. 15901598, 2015.
19) F.-L. Luo and C. Zhang, "*Signal Processing for 5G: Algorithms and Implementations*," John Wiley & Sons, 2016.
20) W. Xiang, K. Zheng, and X. S. Shen, "*5G mobile communications*,", Springer, 2016.
21) M. Renfors and Y. Neuvo, "The maximum sampling rate of digital filters under hardware speed constraints," *IEEE Transactions on Circuits and Systems*, vol. 28, no. 3, pp. 196202, 1981.
22) S. Kasturia and J. H. Winters, "Techniques for high-speed implementation of nonlinear cancellation," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 5, pp. 711717, 1991.
23) C.-H. Lin, A.-Y. A. Wu, and F.-M. Li, "High-performance VLSI architecture of decision feedback equalizer for gigabit systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 9, pp. 911915, 2006.
24) C.-S. Lin, Y.-C. Lin, S.-J. Jou, and M.-T. Shiou, Concurrent digital adaptive decision feedback equalizer for 10GBase-LX4 ethernet system, in *IEEE Custom Integrated Circuits Conference, CICC'07, 2007. IEEE, 2007*, pp. 289292.
25) Y.-C. Lin, S.-J. Jou, and M.-T. Shiue, "High throughput concurrent lookahead adaptive decision feedback equaliser," *IET circuits, devices & systems*, vol. 6, no. 1, pp. 5262, 2012.
26) A. L. Pola, J. E. Cousseau, O. E. Agazzi, and M. R. Hueda, "A low-complexity decision feedforward equalizer architecture for high-speed receivers on highly dispersive channels," *Journal of Control Science and Engineering*, vol. 2013, p. 3, 2013.
27) S. Baghel and R. Shaik, "FPGA implementation of fast block LMS adaptive filter using distributed arithmetic for high throughput," in *Communications and Signal Processing (ICCSP), 2011 International Conference on. IEEE, 2011*, pp. 443447.
28) B. K. Mohanty, P. K. Meher, and S. K. Patel, "LUT optimization for distributed arithmetic based block least mean square adaptive filter," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 5, pp. 19261935, 2016.
29) B. K. Mohanty and P. K. Meher, "A high-performance energy-efficient architecture for FIR adaptive filter based on new distributed arithmetic formulation of block LMS algorithm," *IEEE Transactions on Signal Processing*, vol. 61, no. 4, pp. 921932, 2013.
30) S. M. Kuo, S. Mitra, and W.-S. Gan, "Active noise control system for headphone applications," *IEEE Transactions on Control Systems Technology*, vol. 14, no. 2, pp. 331335, 2006.
31) H.-S. Vu and K.-H. Chen, "A low-power broad-bandwidth noise cancellation VLSI circuit design for in-ear headphones," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*,vol. 24, no. 6, pp. 20132025, 2016.

Digital Predistortion Linearization for Multi-Band/Multi-Channel Software Defined Transmitters

Praveen Jaraut
Electronics and Communication Engineering
Indian Institute of Technology Roorkee
praveen.jaraut@ieee.org

***Abstract*—The most optimum PA linearization technique is Digital Predistortion (DPD). The complexity and numerical stability of the DPD model is still a huge challenge. This work investigates the issue of implementation complexity, numerical stability and feasibility of DPD model adaptation for low cost FPGAs for single band, multi-band and multi-channel transmission. Independent Component Analysis (ICA) as a novel algorithm level solution is proposed to enhance numerical stability of the state-of-the-art models for different carrier aggregated (intraband contiguous, intra-band non-contiguous and inter-band noncontiguous) LTE signals. The application of the ICA technique upon MP model reduces model complexity and improves numerical stability of the DPD model for CA LTE signals. In MIMO transmitters, a neural network (NN)-based DPD model has been proposed as an integral solution to compensate for crosstalk, PA nonlinearity, I/Q imbalance and dc offset imperfections simultaneously. The proposed NN DPD model provides a single-model digital mitigation solution to multi-branches of MIMO transmitters, which is suitable for higher order MIMO operation.**

I. Introduction

Wireless and mobile communication is evolving to offer newer services and higher data rates to more number of users within a limited radio-frequency (RF) spectrum. In order to meet the requirements, transceivers should support multi-standards, multiple bands and multiple-input multiple-output (MIMO) topology [1]. Nowadays, complex modulation schemes, such as orthogonal frequency-division multiplexing (OFDM) is used in Long-Term evolution (LTE) and World-wide Interoperability for Microwave Access (WiMAX) wireless technologies. These modulation schemes are spectrally efficient, but have a high peak-to-average power ratio (PAPR). The multi-carrier and carrier aggregation (CA) techniques are used to increase overall network capacity, data rates and achieve an allocation of the fragmented spectrum [2]. CA consists of combining various component carriers (CCs). In Long Term Evolution-Advanced (LTE-A), CA of up to five CCs, each of up to 20 MHz is possible. This places very challenging requirements for radio frequency (RF) front-end specifications in terms of power efficiency for base stations. When base station connects to user equipments (UEs) in down-link operation, it requires high power and this requirement is met using Power Amplifier (PA) [3]. PA should operate near saturation region to satisfy the exacting requirements on power efficiency. However, PA behaves nonlinearly in saturation region and produces distortions. This undesirable effect due to non-linearity of the PA can be overcome by digital predistortion (DPD), which is also the most popular technique for linearization of PA [4]. The DPD comprises of nonlinear digital model which pre-distorts the incoming signal, which in turn cancels the distortion generated by PA. For baseband level predistortion, a baseband signal should be passed through DPD linearizer implemented in Digital Signal Processing (DSP). In an ideal case, it is the PA's inverse transfer function as shown in Fig. 1 [4]. The resultant transfer function of DPD+PA system is a linear function.

Several behavioral models consider static nonlinearities only [5], [6] but due to a poor decoupling in FET gate and drain as well as poor decoupling in BJT base and collector, electrical memory effects are present in PAs. The thermal memory effect is also present in PA due to junction temperature. Many behavioral methods such as Volterra and Memory Polynomial (MP) have been widely used to account for these memory effects [4]. Several variations of MP and Volterra models such as generalized memory polynomial (GMP) and Modified Volterra Series or dynamic Volterra series were proposed [4]. However, when a wideband signal is used, more number of coefficients are required in their DPD models due to prominent memory effects. Therefore, DPD techniques with lower complexity are required to linearize the RF PA output especially for multi-channel communication systems such as multi-band and MIMO signal transmitters.

This paper is organized as follows. Section-II describes the procedure for behavioral modeling required in DPD technique. Section II presents a novel ICA method for the DPD models for CA LTE signals to numerical stability of DPD models. Section-III presents a neural network (NN)-based DPD as an integral solution to compensate for crosstalk, PA nonlinearity, I/Q imbalance and DC offset imperfections simultaneously in MIMO transmitters.

II. Proposed ICA based DPD technique for low resource consumption using CA LTE Signals

The non-linear distortions generated by PA for intra-band and inter-band CA signals results in intermixing of in-band carrier components (CC) of CA signal as well as in-band

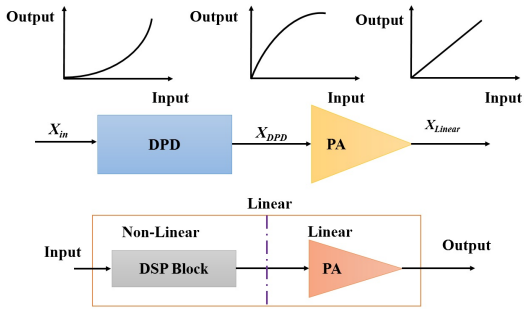


Fig. 1. The fundamental concept of DPD [4].

IMDs. The output of PA becomes very noisy and it does not follow second-order statistics like a gaussian distribution. The steps followed by fast ICA technique are as follows [7], [8]:

1) Let $\mathbf{U}$ is a $(M+1)(K)\times L$ observation data matrix where, L is number of input samples.
2) Center the data matrix $\mathbf{U}$ around its mean value $\mathbf{U} \leftarrow \mathbf{U}-\mathrm{E}\{\mathbf{U}\}$
3) Whiten the data $\mathbf{U}$ matrix $\mathbf{W}=\mathbf{Q}\boldsymbol{\Lambda}^{-1/2}\mathbf{Q}^{H}\mathbf{U}$ where, $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{H}=\mathrm{E}\left\{\mathbf{U}\mathbf{U}^{H}\right\}$
4) Initialize a random matrix $\mathbf{p}$ whose dimension is $S\times(M+1)(K)$ and $\mathbf{p}$ is such that $\|\mathbf{p}\|=1$
5) Update $\mathbf{p}$
   a. $\mathbf{p} \leftarrow \mathrm{E}\{\mathbf{W}f(\mathbf{pW})\}-\mathrm{E}\{f'(\mathbf{pW})\}\mathbf{p}$ where, $f(y)=y^3$
   b. $\mathbf{p} \leftarrow \mathbf{p}/\|\mathbf{p}\|$
6) The independent component matrix $\mathbf{V}=(\mathbf{pW})^{T}$
7) $\mathbf{y}=\mathbf{VD}$ where $\mathbf{D}=[d_0,d_1,\cdots,d_s]^{T}$ is a $S\times 1$ vector of ICA-based coefficients and can be calculated using the pseudo-inverse method.
8) Using the pseudo-inverse method $\hat{\mathbf{D}}$ is extracted as $\hat{\mathbf{D}}=\left(\mathbf{V}^{H}\mathbf{V}\right)^{-1}\mathbf{V}^{H}\mathbf{y}=\mathbf{V}^{+}\mathbf{y}$

After extracting the coefficients, the inverse modeling performance is evaluated in terms of normalized mean square error (NMSE) and adjacent channel error power ratio (ACEPR).

For the case of MP model, the dimensions of the observation matrix $\mathbf{V}$ of MP-ICA model is reduced from the $L\times(M+1)(K)$ to $L\times S$. The pruning percentage $(1-S/((M+1)(K)))\times 100$ of MP-ICA is selected such that MP-ICA's inverse modeling performance NMSE must be nearly equal to MP's NMSE performance.

It is to be noted that the proposed method is significantly different from the Principal Component Analysis (PCA) method. PCA projects the data along the eigen vectors, which have the highest variance. The eigen value of the covariance matrix of the observation matrix measures the variance of the data along an eigen vector. The steps followed by PCA technique are already defined in previous chapter. It is applied upon MP and 2D-MP model and termed as MP-PCA and 2D-MP-PCA. The pruning percentage of MP-PCA is selected such that MP-PCA's inverse modeling performance NMSE must be nearly equal to MP's NMSE performance.

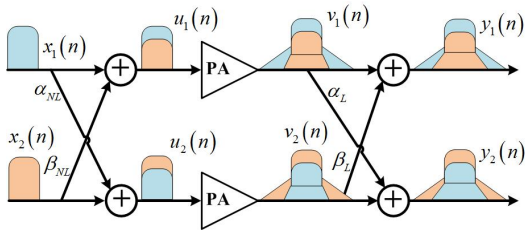


Fig. 2. MIMO transmitters with nonlinear crosstalk and linear crosstalk.

PCA technique uses covariance i.e. it is based on second-order statistics/Gaussian distribution. While Fast-ICA uses fourth-order statistics i.e. it identifies components for non-gaussian signals too. Therefore, PCA technique is inadequate and ICA can be applied to identify the independent components of intermixed CA noisy signals. The observation matrix $\mathbf{V}$ of MP-ICA does not have nonlinear geometric terms thus its condition number would be reduced leading to numerical stability of the solution.

III. Proposed NN based DPD Model for Mitigation of Imperfections in MIMO Transmitters

A. MIMO Transmitter

Direct-conversion transmitter is known to suffer ill-effects of gain imbalance, phase imbalance and LO leakage on the transmitted signal. In MIMO transmitters, signals are transmitted at the same carrier frequency in different transmitters' paths. Hence, the effect of above-mentioned impairments magnifies, when the signal is distorted due to PA nonlinearity and multi-branch crosstalk.

B. Crosstalk

Crosstalk is induced due to the coupling effects between different transmitters' paths or leakage through the common LO. Crosstalk can be categorized as linear and nonlinear. Fig. 2 shows the MIMO transmitters with nonlinear crosstalk and linear crosstalk. Linear crosstalk occurs after the PA. In Fig. 2, coupling factors α_L and β_L denotes the effects of linear crosstalk. Nonlinear crosstalk occurs before PA. This crosstalk effect is amplified when the composite signal passes through a nonlinear PA. In Fig. 2, coupling factors α_{NL} and β_{NL} denotes the effects of nonlinear crosstalk. Generally, MIMO crosstalk has a coupling factor between -15 dB to -30 dB [9].

C. I/Q Imbalance

In a MIMO transmitter, I/Q imbalance occurs due to a mismatch between the in-phase (I) and quadrature-phase (Q) signal paths in the modulator. For $P\times P$ MIMO transmitters, let $x_p(n)$ denote the p^{th} transmitter path baseband input signal, where $p=1,2,\cdots P$. Due to I/Q imbalance, signal at the output of quadrature modulator is

$$\hat{x}_p(n)=a_p x_p(n)+b_p x_p^{*}(n) \tag{1}$$

where

$$a_p=\cos(\theta_p/2)+j\varepsilon_p\sin(\theta_p/2) \tag{2}$$

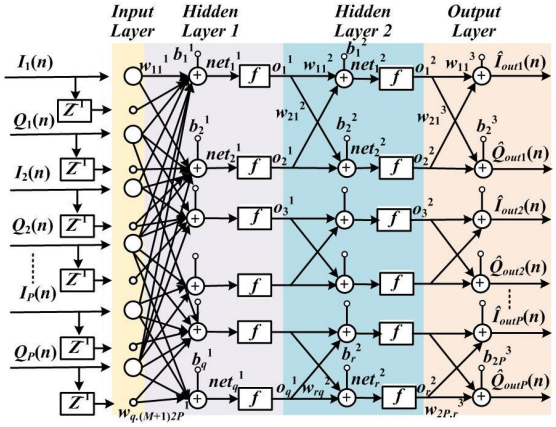


Fig. 3. Real-valued time-delay feedforward backpropagation-based Neural Network.

$$b_p = \varepsilon_p \cos(\theta_p/2) + j \sin(\theta_p/2) \tag{3}$$

In (2) and (3), ε_p and θ_p represents the p^{th} transmitter path's gain imbalance and phase imbalance respectively. For a balanced modulator, ε_p=1 and θ_p=0°. *I/Q* imbalance can be represented by a metric known as image rejection ratio (IRR) [9]. It is defined as

$$\Gamma_p \ \ (\text{dB}) = 20\log_{10}(b_p/a_p) \tag{4}$$

Generally, IRR in RF transmitters ranges from -20 dB to -40 dB [9].

D. Proposed Neural Network based DPD Model

Fig. 3 shows the real-valued time-delay NN [9]. The input vector contains present and past values of I and Q. The input vector is defined as

$$\begin{aligned} X = [&I_1(n), \cdots, I_1(n-m), Q_1(n), \cdots, Q_1(n-m), \\ &I_2(n), \cdots, I_2(n-m), Q_2(n), \cdots, Q_2(n-m), \\ &\cdots, I_P(n), \cdots, I_P(n-m), Q_P(n), \cdots, Q_P(n-m)] \end{aligned} \tag{5}$$

where $I_P(n)$, $Q_P(n)$ are the I and Q components of base-band input signal of P^{th} transmitter branch at n^{th} time instant. z^{-1} represents the unit delay operator. The feedforward back-propagation neural network is used.

1) Feedforward Propagation: During feedforward computation, data propagates from neurons of a lower layer to upper layer. As shown in Fig. 3, two hidden layers are used in this neural network. The net input in layer l+1is given by

$$net_j^{l+1} = \sum_{i=1}^{q} w_{ji}^{l+1} o_i^l + b_j^{l+1} \tag{6}$$

where w_{ji}^{l+1} represents the synaptic weight between the i^{th} input from the previous layer to the j^{th} neuron of the present layer. Initially, weights are set in the interval of [-0.8, 0.8] and during backward propagation, weights are adjusted to reduce the error. q represents the total number of neurons in the previous layer and b_j^{l+1} denotes bias of the j^{th} neuron in the $l+1^{th}$ layer. The output of neuron j at $l+1^{th}$ layer is

$$o_j^{l+1} = f\left(net_j^{l+1}\right) \tag{7}$$

The hidden layers have the hyperbolic tangent function, as the activation function, f. It maps the nonlinearity between -1 and 1. The output of any layer works as an input to the next layer. The outputs of hidden neurons are linearly summed up at the output layer.

2) Backward Propagation: During backward propagation, the performance index for the NN is calculated as

$$V = \tfrac{1}{2} \sum_{n=1}^{N} \left\{ e_n^T e_n \right\} \tag{8}$$

where e is the error between the actual baseband outputs ($I_{outP}(n)$, $Q_{outP}(n)$) and the outputs from output-layer neurons of the NN model ($\hat{I}_{outP}(n)$, $\hat{Q}_{outP}(n)$) of P^{th} transmitter branch of MIMO.

Then the Levenberg-Marquardt algorithm is used, which is an approximation to Gauss-Newtons method. According to this algorithm, the parameter V is minimized with respect to a parameter $\mathbf{u}$ which depends on synaptic weights and biases. During backward propagation $\mathbf{u}$ is updated as

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \left[\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}\right]^{-1}\mathbf{J}^T\mathbf{e} \tag{9}$$

where

$$\mathbf{u} = \left[w_{11}^1 \cdots w_{q.2P(M+1)}^1 b_1^1 \cdots b_q^1 w_{11}^2 \cdots w_{2Pr}^3 b_1^3 \cdots b_{2P}^3\right] \tag{10}$$

where $\mathbf{J}$ is the Jacobian matrix calculated over error matrix $\mathbf{e}$ with respect to $\mathbf{u}$. q and r are the numbers of neurons of two hidden layers. M and P are the memory depth and number of transmitter's branches in MIMO. Whenever V increases, μ is multiplied by some factor β. Whenever V decreases, μ is divided by some factor β. Initially μ and β are set equal to 0.01 and 10 respectively. The whole procedure is iterated until the good performance is achieved by NN.

REFERENCES

[1] P. B. Kenington, "RF and Baseband Techniques for Software Defined Radio," Artech House, Inc., Norwood, MA, 2005.

[2] 3GPP RP-091440, "Work Item Description: Carrier Aggregation for LTE," Nokia Corporation, Sanya, P.R. China, Dec. 2009.

[3] S. Cripps, "RF power amplifers for wireless communications," 2nd ed. Boston: Artech House, 2006.

[4] F. M. Ghannouchi, and O. Hammi, "Behavioral modeling and predistortion," *IEEE Microw. Mag.*, vol.10, no.7, pp. 52–64, Dec. 2009.

[5] K. J. Muhonen, M. Kavehrad, and R. Krishnamoorthy, "Look-up table techniques for adaptive digital predistortion: a development and comparison," *IEEE Trans Veh Technol.*, vol. 49, no. 5, pp. 1995-2-002, Sep. 2000.

[6] K. Rawat, M. Rawat, and F. M. Ghannouchi, "Compensating IQ imperfections in hybrid RF/digital predistortion with an adapted look up table implemented in an FPGA," *IEEE Trans. Circuits Syst. II Exp. Briefs*, vol. 57, no. 5, pp. 389–393, May 2010.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, "ICA by Maximization of Nongaussianity," in *Independent Component Analysis*, John Wiley & Sons, 2001, ch. 8, sec. 8.3.5, pp. 188–192.

[8] P. Jaraut, M. Rawat, and P. Roblin, "Digital Predistortion technique for Low resource consumption using Carrier Aggregated 4G/5G Signals," *IET Microw. Antennas Propag.*, to be published. DOI: 10.1049/iet-map.2018.5608.

[9] P. Jaraut, M. Rawat, and F. M. Ghannouchi, "Composite Neural Network Digital Predistortion Model for Joint Mitigation of Crosstalk, I/Q Imbalance, Nonlinearity in MIMO Transmitters," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 11, pp. 5011–5020, Nov. 2018.

Low complexity transmission in few mode fibers using limited feedback of principal modes

Jinesh C. Jacob
Department of Electrical Engineering
Indian Institute of Technology Bombay

Kumar Appaiah
Department of Electrical Engineering
Indian Institute of Technology Bombay

I. ABSTRACT

Mode division multiplexing (MDM) using few mode fibers (FMFs) is a promising approach to resolve the capacity crunch in the modern optical communication system. However, in practical FMFs, the refractive index profile of the fiber is perturbed along the length of the fiber due to various physical effects, viz. core ellipticity, bends, twists, internal and external stresses, etc. This perturbation introduces coupling among the modes of the fiber and power is redistributed randomly among the modes during propagation. We know that each mode of the FMF is associated with a unique spatial distribution of electric field and a frequency dependent wave number. Since frequency dependence of the propagation characteristic of each mode is different, optical pulses will travel with different group velocities through modes. As the result of mode coupling, an optical pulse launched into one specific mode will couple into other modes during the propagation, and several pulses, with different delays, appear at the fiber output. This causes modal dispersion which can be compensated after at the receiver using digital filters. We go on to show that, feeding back some of this information permits a significant reduction in the complexity of the receiver when optical transmission strategies are used.

In our work, we have proposed different types of limited feedback strategies for the FMFs under different mode coupling regimes. To make the effective communication using the limited feedback, we need to identify the parameters of the FMF link that should be estimated at the receiver and sent back to the transmitter. Much like CSI feedback in wireless communication, we can use mode coupling matrix of FMF (similar to channel matrix in wireless terminology) or singular vectors of the mode coupling matrix as the choice for feedback. However, in the case of FMF communication, where the communication is performed using a single carrier (laser source) and the modulation bandwidth is usually very large, the frequency dependent mode coupling matrix or its singular vectors will be varied across the bandwidth of operation and it demands large amount of feedback overhead. In [1], the eye opening at the receiver is used as a feedback quantity, and the parameters of a spatial light modulator (SLM) at the transmitter are adaptively adjusted according the feedback information. In this work, we have used special type of electric field patterns, known as "principal modes", to facilitate the limited feedback.

In [2], it has been shown that principal modes (PMs) are linear combinations of ideal modes, which are representing propagating waves within the optical fiber. For a particular FMF section, there are as many PMs as ideal modes. The important property of the PMs is the first order frequency independent nature, i.e if an a pulse is launched into the FMF with a spatial pattern to one of the PMs, then the output field pattern remains unchanged and the pulse can be recovered without significant dispersion. This permits the use of a large bandwidth around the center frequency of the irrespective of the frequency dependent mode coupling without the need for additional dispersion compensation. PMs are the eigen vectors of the group delay matrix $\mathbf{F}$ which is given as

$$\mathbf{F} = -j\mathbf{U}^H(\omega)\frac{\partial \mathbf{U}}{\partial \omega}\bigg|_{\omega=\omega_0}$$

where $\mathbf{U}(\omega)$ is the frequency dependent mode coupling matrix and ω_0 is the laser frequency. $(.)^H$ denotes the Hermitian operator. The propagation of a PM pattern through a FMF with negligible mode dependent loss (MDL) can be written as (within the correlation bandwidth of PM)

$$e^{-j\phi(\omega)}\mathbf{p}_{\text{out}} \approx \mathbf{U}(\omega)\mathbf{p}_{\text{in}}$$

where $\mathbf{p}_{\text{in}}$ and $\mathbf{p}_{\text{out}}$ are input and output PM patterns respectively. Within the correlation bandwidth of PMs (the bandwidth for which output PM pattern remains unchanged), $\phi(\omega) \approx \tau\omega$ where τ is the group delay of the PM which is shown to be eigen value of $\mathbf{F}$ matrix. The FMF systems which use the PM patterns as the input patterns instead of conventional LP modes (ideal modes) are experimentally demonstrated [3]. PM based FMF systems use programmable SLMs which can be controlled using computer generated electrical signals to the modulate spatial distribution and state of polarization (SOP) of arbitrary optical input signals incident on them.

Initially, we have used the limited feedback of PM pattern to mitigate modal dispersion in FMF systems. For this feedback model, the transmitter and receiver should equipped with an SLM and a spatial filter (SF) respectively. The PM pattern will be estimated at the receiver using some pilot based

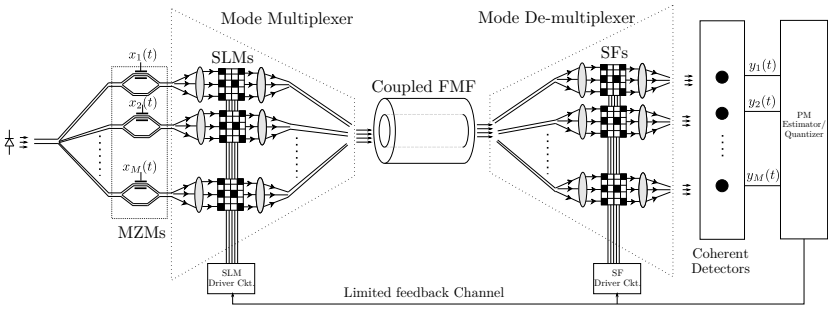


Fig. 1. Proposed MDM architecture for FMFs using limited feedback of principal modes (PMs). PM estimator/quantizer estimate the PM patterns and quantize these patterns using the the codebooks which was generated based on the mode coupling strength. The feedback information used to update the SLMs and SFs.

techniques. The estimated PM pattern is quantized using vector quantization (VQ) techniques. Since the statistical distribution of PM patterns is unknown for the particular mode coupling strength, the vector codebook is generated using trained PM patterns. These trained PMs are produced from FMF multi-section model which is adapted from the FMF model given in [4]. The statistical mode coupling properties of the FMF model should match with physical FMF. The important fact is that codebook generation is a one time process for a given mode coupling strength and therefore, complexity of codebook generation does not affect the system performance significantly. The estimated PM pattern is quantized against PM codebook and the label of "closest" the codevector is fed back to the transmitter. The SLM driver circuit at the transmitter will generate the electric signal for the SLM to produce the pattern, represented by the feedback bit sequence.

We have used two types of PM codebooks. The first one is generated using the Linde-Buzo-Gray algorithm (LBG), which uses the trained PMs to generate the codebook. It was found that this approach is very effective for weakly coupled FMFs where each PM distribution resembles a cloud around the ideal LP modes, and PMs are highly correlated. In the second codebook generation strategy, we have used the fact that the distribution of the group delay operator $\mathbf{F}$ approaches a Gaussian unitary ensemble (GUE) if the FMF is in the strong coupling regime [5]. Since the eigenvectors of the GUE matrix can be treated as being "uniformly" distributed points on a unit sphere in multidimensional complex space, each PM pattern of strongly coupled FMFs can be represented as a point on the Grassmannian manifold. We have used the Gassmannian line packing algorithm to generate the PM code book for strongly coupled FMFs. Our results suggest that quantization controls modal dispersion and the performance of the quantized PMs is within 2 dB of the unquantized PMs with only 6 bit quantization.

Subsequently, we have extended the concept of PM feedback in to mode division multiplexing (MDM) systems. In the conventional approach of MDM where LP mode patterns are used for multiplexing, the cross-talk, induced from mode coupling, is equalized electrically using MIMO-DSP filters. The MDM architecture which uses the limited feedback of PMs is shown in Figure 1. For FMFs with negligible MDL, the PM patterns are also orthogonal patterns, like ideal modes, and we can accomplish cross-talk free spatial multiplexing by transmitting through PMs even in the presence of mode coupling and with little or no MIMO processing. However, the information about PMs, estimated at the receiver, should be made available at the transmitter, which is unaware about PMs. In this work, unlike single PM feedback, we need to quantize multiple PM patterns simultaneously. This problems boils down to the quantization of unitary matrix where each column of the unitary matrix represents the PM pattern. In this work, we have converted the unitary matrix into set of scalar parameters using Givens rotation [6]. These scalar values are quantized and fed back to transmitter. The SLMs and SFs are updated according to the quantized PM patterns.

Presently, we are working on the concept of PM feedback in the DWDM (dense wavelength division multiplexing) in FMFs. The DWDM combined with MDM will significantly increase the data carrying capacity of the fiber. However, mode coupling will be different for different WDM channels, and needs to be considered separately for each channel.

REFERENCES

[1] R. Panicker, J. P. Wilde, J. M. Kahn, D. F. Welch, I. Lyubomirsky *et al.*, "10 $\times$ 10 Gb/s DWDM transmission through 2.2-km multimode fiber using adaptive optics," *IEEE Photon. Technol. Lett.*, vol. 19, no. 15, pp. 1154–1156, 2007.

[2] S. Fan and J. M. Kahn, "Principal modes in multimode waveguides," *Optics Letters*, vol. 30, no. 2, pp. 135–137, 2005.

[3] J. Carpenter, B. J. Eggleton, and J. Schröder, "Observation of Eisenbud–Wigner–Smith states as principal modes in multimode fibre," *Nature Photonics*, vol. 9, no. 11, pp. 751–757, 2015.

[4] M. Shemirani, W. Mao, R. Panicker, and J. Kahn, "Principal Modes in Graded-Index Multimode Fiber in Presence of Spatial and Polarization-Mode Coupling," *J. Lightw. Technol.*, vol. 27, no. 10, pp. 1248–1261, 2009.

[5] M. B. Shemirani and J. M. Kahn, "Higher-order Modal Dispersion in Graded-Index Multimode Fiber," *J. Lightw. Technol.*, vol. 27, no. 23, pp. 5461–5468, 2009.

[6] J. C. Roh and B. D. Rao, "An efficient feedback method for MIMO systems with slowly time-varying channels," in *IEEE Wireless Communications and Networking Conference, 2004*, vol. 2. IEEE, 2004, pp. 760–764.

PUBLICATIONS

Journals

- J. C. Jacob, R. K. Mishra and K. Appaiah, *"Quantization and Feedback of Principal Modes for Dispersion Mitigation and Multiplexing in Multimode Fibers,"* in IEEE Transactions on Communications, vol. 64, no. 12, pp. 5149-5161, Dec. 2016. doi: 10.1109/TCOMM.2016.2605693

Conferences

- J. C. Jacob and K. Appaiah, *"Multiplexing using principal modes in spliced MMFs with mode dependent losses,"* 2017 Twenty-third National Conference on Communications (NCC), Chennai, 2017, pp. 1-6. doi: 10.1109/NCC.2017.8077069
- J. C. Jacob and K. Appaiah, *"Limited feedback and interpolation of principal modes in spatially multiplexed WDM fiber links,"* presented in 2018 International conference on signal processing and communications (SPCOM), Bengaluru, 2018
- R. Mishra, J. C. Jacob and K. Appaiah, *"Quantization and feedback of principal modes for high speed multimode fiber links,"* 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, 2016, pp. 1-6. doi: 10.1109/ICC.2016.7510928
- R. Mishra, J. C. Jacob and K. Appaiah, *"Limited feedback of principal modes for high speed multimode fiber links,"* 2015 Workshop on Recent Advances in Photonics (WRAP), Bangalore, 2015, pp. 1-4. doi: 10.1109/WRAP.2015.7805962

Low Cost RF Predistortion for Carrier Aggregated Ultra-Wideband Signals

Karan Gumber, *Student Member, IEEE* and Meenakshi Rawat, *Member, IEEE*

***Abstract*— RF-Predistorter (RF-PD) is an attractive solution for power amplifier in 5G base station driven by carrier aggregated signals and it also provide a gateway to repeaters for long distance communication, where baseband information is not readily available. For ultra-wideband (UWB) signals, existing linearization techniques become intractable due to high cost and power overhead consumption. Proposed RF-PD eliminates the use of power hungry data converters and FPGAs but still provide worthy linearization using cost and energy efficient passive components. Particularly, ZX60V-63+ PA is stimulated with two application scenario (a) non-contiguous 160MHz 8 component carrier (CC) centered at 800MHz, each CC occupying 20MHz instantaneous bandwidth, and (b) contiguous 8 CC centered at 2400MHz. Proposed RF-PD delivers an adjacent channel power ratio of -52.19dBc with an improvement of 16.22dB at 6.7dB back-off for non-contiguous signal. For contiguous signal, an ACPR improvement of 12.07dB is achieved at 6.3dB back-off.**

***Index Terms*—Adjacent channel power ratio (ACPR), carrier aggregation, linearization, power amplifier, predistorter.**

I. Introduction

WITH the advent of 5G era, we expect that the whole civilization will go through a revolutionary change. This change will make 5G fundamentally divergent from preceding mobile generation. A unique event was witnessed by us in 2016, in which global internet traffic in 2016 will be equal to 44 times to the volume in 2005. An idea to shift toward 5G is based on the challenges that are not effectively addressed by 4G i.e. higher data rates, low latency, and power consumption.

Forthcoming 5G anticipates the use of broader bandwidth through the application of carrier aggregation (CA). Undeniably, aggregating the multiple CC has contributed to enhancing the data rates and user throughput. In release 10, the standardized framework of long term evolution (LTE) allows the aggregation of 5 CC to 8CC that makes 100MHz-160MHz signal. But this framework is expanded in Release 13, where goal is to aggregate 32CC that makes 640MHz signal [1], [2].

If the existing digital predistortion (DPD) is employed for 160MHz signal, which means at least 800MHz linearization bandwidth is required. Such a high bandwidth will remarkably increase the crisis in system design. DPD also requires knowledge of digital baseband signal at the input and computational speed of the digital circuit limits the operational bandwidth [3]. This proves to be a limitation of the RF repeater system, where linearization needs to be applied to incoming RF signal before amplifying and transmitting it to the next station. Hence it become essential to come up with a linearization technique that alleviate the need of high speed data converters, wideband transmitter and receiver chains for processing UWB signal.

K. Gumber and M. Rawat are with the Communication group, Electronics and Communication Engineering Department, Indian Institute of Technology, Roorkee, Uttarakhand, India. (e-mail: gumberkaran88@gmail.com, rawatfec@iitr.ac.in).

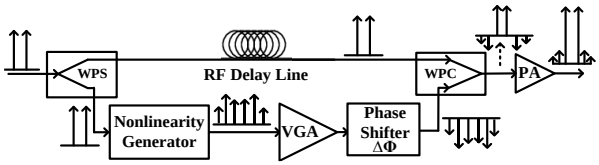


Fig. 1. Conventional Analog predistortion circuit.

In this letter, an ultra-broadband RF-PD is proposed that not only provide sufficient linearization to UWB signals, but also enhance the system efficiency by reducing the power overhead consumption. It is able to linearize 8CC UWB signal which is not possible with conventional analog predistorter (CAPD). Therefore, proposed method is desirable for such repeaters and transmitters, where, the non-linear PD function and corrections are applied in the analog domain without gathering the history of baseband. The only tradeoff is that due to the non-ideality of passive components, effective linearization performance is limited as compared to DPD.

II. Circuit Configuration

The block diagram of CAPD is shown in Fig. 1, in which nonlinear path composed of phase shifter (P.S), nonlinearity/cubic element and variable gain amplifier (VGA). In the CAPD, PS was used after nonlinear element to optimize the phase of the expanded signal containing all the necessary intermodulation distortion (IMD) components. To deteriorate the nonlinearity of PA, the new IMD product will be introduced by a combination of VGA and cubic element. If 160MHz signal is given at the input of CAPD, to compensate fifth order IMD (IMD5), the nonlinear element expand it to 800MHz. This expanded bandwidth signal is given at the input of PS. As PS is a narrow band component, it is unable to capture and adjust the phase of 800MHz UWB signal.

This drawback of CAPD is addressed in this letter, in which narrow band PS is used in conjunction with IMD generator to linearize UWB signal. Proposed RF-PD architecture is shown in Fig. 2, in which there is mutual exchange in the position of PS and IMD generator. This change is in favor to capture and adjust the phase of UWB signal. With this slight modification instead of capturing 800MHz which is expanded spectrum, PS

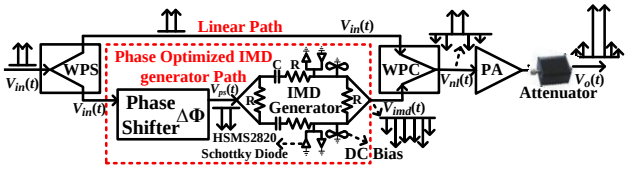


Fig. 2. Proposed ultra-broadband RF-Predistortion circuit.

only capture 160MHz which is the original bandwidth of input signal. The description of components used in RF-PD is as:

A. *Wilkinson Power Splitter (WPS)*

A WPS (1:2) following the input port is used to split the input RF signal into two paths i.e. linear and phase optimized IMD generator path. The input signal $V_{in}(t)$ is given as:

$$V_{in}(t) = A(t)\left[\cos\left(\omega_1 t + \phi_{in}\left[A(t)\right]\right) + \cos\left(\omega_2 t + \phi_{in}\left[A(t)\right]\right)\right] \quad (1)$$

where $A(t)$ and ϕ_{in} is the amplitude and phase of input signal.

B. *Phase Shifter*

As shown in Fig. 2, the output of WPS is given at the input of PS. Phase of the input signal is adjusted by PS in such a way that the IMD produced by IMD generator circuit counteracts the nonlinear distortion of PA. By adjusting control voltage of PS, different combination of phase characteristics can be obtained. Depending on the phase of fundamental signal adjusted by PS, the nonlinearity of IM generator can be tailored to match the nonlinearity of PA over wide frequency range. For proper cancellation of PA IMD, signal at the output of PS have opposite phase w.r.t input signal i.e. $\phi_{PS} = 180° + \phi_{in}$.

$$V_{PS}(t) = A(t)\left[\cos\left(\omega_1 t + \phi_{PS}\left[A(t)\right]\right) + \cos\left(\omega_2 t + \phi_{PS}\left[A(t)\right]\right)\right] \quad (2)$$

$$= -A(t)\left[\cos\left(\omega_1 t + \phi_{in}\left[A(t)\right]\right) + \cos\left(\omega_2 t + \phi_{in}\left[A(t)\right]\right)\right] \quad (3)$$

-ve sign in above equation represents phase reversal. The capturing bandwidth of PS in proposed RF-PD is equal to the bandwidth of original input signal, whereas in CAPD it is usually five times more than the signal bandwidth.

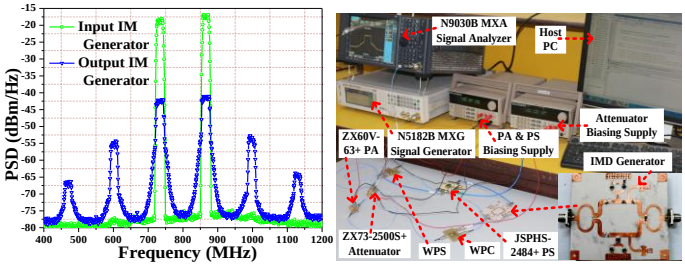
C. *IMD Generator*

The capability of generator is judge by its ability to cancel the fundamental signal and to generate higher order IMDs. A pair of diodes connected in anti-parallel fashion generate odd order IMD and cancel even order IMD, while RC filter bank cancel the fundamental signal. The value of R and C was extracted in ADS by tuning and optimization that provide best fundamental signal cancellation. IMD generator provide fundamental signal cancellation up to 25dB for 8CC non-contiguous signal as shown in Fig. 3(a). It also generate in-band as well as out of band IMD components. The output of IM generator is given as: (taking only 3rd order into account)

$$V_{IM}(t) = a_1 V_{PS}(t) + a_3 V^3{}_{PS}(t) \quad (4)$$

$$V_{IM}(t) = \frac{-3\Delta A(t) a^3}{4}\left[\begin{pmatrix}\cos(2\omega_2 - \omega_1)t \\ +\phi_{in}\left[\Delta A(t)\right]\end{pmatrix} + \begin{pmatrix}\cos(2\omega_1 - \omega_2)t \\ +\phi_{in}\left[\Delta A(t)\right]\end{pmatrix}\right] \quad (5)$$

The fundamental signal, third order IMDs that lies at frequency $2\omega_1 + \omega_2$ and $2\omega_2 + \omega_1$ and third order harmonics that lies at frequency $3\omega_1$ *and* $3\omega_2$ can be easy filtered out.



(a) (b)

Fig. 3. (a) Analysis of IM Generator. (b) Test-bench for data extraction.

D. *Wilkinson Power Combiner (WPC)*

At WPC, the signal from linear and IMD generator path combined with phase difference of 180°. The IMD components in (5) that is generated by IMD path, cancel the IMD components of PA due to its opposite phase.

III. Brief Survey on Previously Proposed RF PD

Several analog methods have already been developed to linearize PA. Among these a linearization technique based on phase correction is proposed in [1] that provides a correction of only 4.2dB for one carrier WCDMA signal. The RF PD that has the capability to linearize 200MHz 64 QAM UWB signal was first reported in [2]. It employs very simple circuitry but provide ACPR correction of only 5.5dB which is not worthy. Table I compare the performance parameters and improvement yielded by proposed RF-PD and other works [3]-[7].

IV. Experimental Verification

For the experimental validation, we implemented a RF-PD that can effectively suppress PA intermodes from 700MHz to 2.5GHz.The test-bench of RF-PD was shown in Fig. 3(b). The signal is uploaded to signal generator (MXG N9030B) via Matlab, which is further given at the input port of WPS. One output of WPS is given directly to WPC and other output is given at the input of PS. PS adjusts the phase of fundamental signal with the help of control voltage. The amplitude and phase of intermodes generated by IM generator were controlled by PS. For demonstration WPS, IM generator and WPC are fabricated in-house. Measurement results are carried out in two phases:

8CC 160MHz non-contiguous signal: Linearization of 160MHz UWB signal is very demanding and challenging task. Proposed RF-PD linearization scheme is applied to non-contiguous 10000001 where binary1 indicates on and 0 indicates off with total instantaneous bandwidth of 160MHz. A narrow-band JSPHS-1000+ 180° Voltage Variable PS, is used to adjust the phase of 8CC non-contiguous signal centered at frequency 800MHz. Fig. 4(a) shows the power spectral density (PSD) of non-contiguous 8CC LTE signal with and without RF-PD. When proposed RF-PD is applied to linearize ZX60-V63+ PA using 160MHz LTE signal, an ACPR of -52.19dBc is achieved with an improvement of over 16.22dB at 6.7dB

TABLE I
PERFORMANCE COMPARISON OF VARIOUS APD LINEARIZER

Reference	BW of PD /Freq.	SIGNAL USED& BW OF SIGNAL	ACPR & IM IMPROVEMENT	PD HARDWARE INCLUDES	KEY POINTS
[3]	-/2.4GHz	Two tone signal with 20MHz spacing WCDMA 5MHz	16dB@6dB backoff (IM3) 13dB@10dB backoff	Two Directional Couplers Two PA's.	It eliminate VM, attenuators and delay lines. Measured BW of circuit is only 80MHz, hence, it is not a strong candidate for future 5G system. Using two PA's makes the circuit costly.
[4]	1.5 GHz - 2.4GHz	Two tone signal with 2MHz spacing	17dB to 25dB (IM3)	WPC, Two voltage variable attenuator, WPS, Two driver amp, broadband PS, coupler.	Complex circuit but provide good linearization. Performance is only limited to IM3; Delay adjustment for modulated signal is tricky challenge.
[5]	-/2.14GHz	Two tone signal with 5MHz spacing WCDMA 20MHz	≈ 36dB(IM3) ≈ 23dB(IM5) 16.4dB	Four hybrid coupler, two VM, two delay lines, gain amp, two error generator circuit (EGC).	Cascading of two APD provide excellent linearization but increase cost and complexity; compensate memory effects; simultaneously cancel IM3 and IM5.
[6]	-/2.14GHz	WCDMA 5MHz WCDMA 20MHz	25dB 20.3dB	Hybrid coupler, WPC, WPS, delay lines, PA, EGC, VM.	Three nonlinear paths provide good linearization but increase the cost and complexity of system.
[7]	2MHz-600MHz	Two tone signal with 1MHz spacing	4.9dB	Capacitor and a biased Schottky diode.	Linearization performance is not worthy; Compact APD with frequency selective characteristics.
Proposed	700MHz – 2.5GHz	8CC non-contiguous 8CC contiguous	16.22dB 12.07dB	WPC, WPS, PS and IM generator	Linearize 8CC signal, hence it is a strong candidate for 5G; Linearization performance is worty.

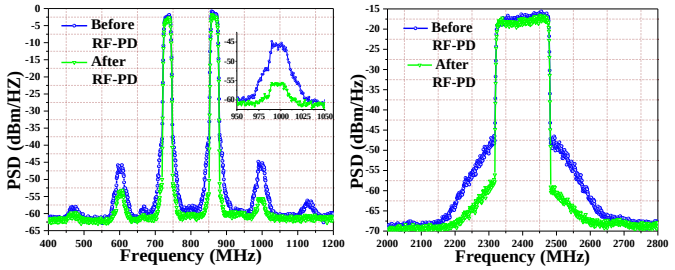


Fig. 4. Measured power spectrum densities of PA with and without the proposed RF PD (a) non-contiguous 8CC. (b) contiguous 8CC signal.

back-off as compared to PA without linearization. It can also be appreciated from Fig. 4(a), that RF-PD also suppress out-of-band IMDs that reduces co-channel interference. However, it achieves the significant ACPR improvement by compensating for in-band and out-of-band IMDs.

8CC 160MHz contiguous signal: A narrow band PS JSPHS-2484+ is used to optimize the phase of 8CC contiguous signal centered at frequency 2.4GHz. Fig. 4(b) shows the PSD of 8CC LTE signal with and without proposed RF-PD. When proposed linearization is applied to linearize PA, an ACPR of -45.02dBc is achieved with an improvement of over 12.07dB at 6.3dB back-off as compared to PA without linearization.

V. CONCLUSION

Proposed RF-PD linearizer will be a powerful incentive to adopt 5G base station, multi-antenna PA design and multi-carrier repeater system. There is a strong demand for the linearizer that has simple circuit configuration, minimal power consumption and provide adequate linearity for UWB signal without gathering baseband information. The performance of linearizer has been verified with 160MHz LTE and two tone signal. Experimental results validate the dexterity of proposed RFPD by providing a linearization performance comparable with conventional and transistor based APD. The designed PD lead to 16dB improvement in ACPR for non-contiguous signal at the center frequency of 800MHz, and more than 12dB improvement for contiguous signal centered at 2400MHz. To the best of the authors' knowledge, this achieved linearization for 160MHz UWB signal implemented using passive components is the highest reported in the literature so far.

REFERENCES

[1] U. Kim and Y. Kwon, "A high efficiency SOI CMOS stacked-FET power amplifier using phase based linearization," *IEEE Microw. Compon. Lett.*, vol. 24, no. 12, pp. 875-877, Dec. 2014.

[2] N. Rostomyan, J.A. Jayamon, and P.M. Asbeck, "15 GHz Doherty power amplifier with RF predistortion linearize in CMOS SOI," *IEEE Trans. Microw. Theory Techn*., Dec., 2017, DOI: TMTT. 2017.2772785

[3] Q. Cai, W. Che, K. Ma, and M. Zhang, "A simplified transistor based analog predistorter for a GaN power amplifier," *IEEE Trans. Circuits Syst. II: Express Briefs*, DOI: TCSII.2017.2735022.

[4] H. park, H. Yoo, S. Kahng, and H. Kim, "Broadband tunable third-order IMD cancellation using LHTL based phase shifter," *IEEE Microw. Compon. Lett.*, vol. 25, no. 7, pp. 478-480, Jul. 2015.

[5] Y-S. Lee, M-W. Lee, S-H Kam, and Y-H. Jeong, "A high linearity wideband power amplifier with cascaded third order analog predistorters," *IEEE Microw. Compon. Lett.*, vol. 20, no. 2, pp. 112-114, Feb. 2010.

[6] Y-S. Lee, M-W. Lee, S-H Kam, and Y-H. Jeong, "A transistor based analog predistorter with unequal delays for memory compensation," *IEEE Microw. Compon. Lett.*, vol. 19, no. 11, pp. 743-745, Nov. 2009.

[7] M. Seo, K. Kim, M. Kim, H. Kim, J. Jeon, M-K Park, H. Lim, and Y. Yang, "Ultra broadband linear power amplifier using a frequency selective analog predistorter," *IEEE trans. Circuits Syst. II: Express Briefs*, vol. 58, no. 5, pp.264-268, May 2011.

List of own Publications

[1] K. Gumber and M. Rawat, "Low cost RFin-RFout predistorter linearizer for high power amplifiers and ultra-wideband signals," IEEE Trans. Instrum. Meas., vol. 67, no. 9, pp. 2069-2081, Mar. 2018.

[2] K. Gumber and M. Rawat, "A modified hybrid RF predistorter linearizer for ultra-wideband 5G systems," IEEE J. Emerg. Sel. Topics Circuits Syst., vol. 7, no. 4, pp. 547-557, Dec. 2017.

[3] K. Gumber and M. Rawat, "Modified RF_{in}-RF_{out} broadband predistorter for 5G communication system," IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, May 2018.

[4] K. Gumber, P. Jaraut, M. Rawat, and K. Rawat, "Digitally assisted analog predistortion technique for power amplifier," IEEE 88th Microwave Measurement Conference (ARFTG), Austin, TX, USA, Dec. 2016.

[5] K. Gumber and M. Rawat, "Digital predistorter design using linear splines and its fixed point implementation," Asia-Pacific Microwave Conference (APMC), New Delhi, India, Dec. 2016.

[6] K. Gumber and M. Rawat, "Ultra Broadband RF predistorter supporting carrier aggregation for future 5G" IEEE Trans. Microw. Theory Tech., (Reject & Resubmit).

Estimation and Optimization of Design Parameters in Diffusive Molecular Nanonetworks

Satish K. Tiwari and Prabhat K. Upadhyay
Discipline of Electrical Engineering
Indian Institute of Technology Indore
Indore, 453552, Madhya Pradesh, India
Email: phd1401102014@iiti.ac.in and pkupadhyay@iiti.ac.in

I. Abstract

Advancements in the field of nanotechnology require an efficient and reliable communication between artificial or genetically engineered bio-nanomachines having limited operational capabilities. The size, power consumption and computational requirement of transceiver and constraint on antenna size make the electromagnetic (EM) wireless communication inappropriate at nanoscale. Moreover, macroscale communication inside tunnels, pipelines, or saline water environment can be inefficient due to the increased conductivity resulting in large attenuation of EM signals. In view of the limitations of the existing EM wave-based technologies, communications among nanomachines require different paradigms.

This thesis focuses on a new energy-efficient communication paradigm at nanoscale, known as molecular communication (MC) [1], [2], which uses chemicals or molecules as the information carrier. Owing to its biocompatibility in general, MC has found application in nanomedicine for the targeted drug delivery and the cutting-edge *in vivo* biomedical applications [3] such as diagnosis, therapy and the monitoring of diseases. However, MC suffers from residual noise, reduced communication range, high latency and lower molecular throughput. As such, we investigate the aforesaid problems for the diffusive molecular communication (DMC) systems.

A. Prior Work and Motivation

One of the major challenges in DMC system is the residual noise or inter-symbol interference (ISI) which occurs due to the emission of a new symbol before the previously emitted molecules disappear at the reception node. Residual noise is a dominant source of error which degrades or bottlenecks the performance of DMC systems, especially at high data rates. Several methods have been proposed in the literature [4]-[10] to control the level of residual noise, e.g., enzymes deployment, ISI-reducing modulation techniques, symbol interval optimization, designing detectors and equalizers to reduce the residual noise, and so on. However, no work have used signal-to-noise ratio (SNR) estimate as a measure to mitigate residual noise. Based on the estimate of SNR at the receiver nanomachine (R_xN), design parameters such as size of propagation area, symbol duration, number of messenger molecules per symbol, etc., can be adjusted to control the level of residual noise in DMC systems.

Another limitation of DMC system is that the information-bearing molecular signal experiences high attenuation with large propagation delay resulting in reduced communication range and lower molecular throughput. This is because, in general, the propagation time increases with the square of communication distance, while the peak molecular count decreases with cube of the communication distance. Thereby, such phenomena necessitate a relay deployment in DMC systems with distant receivers. Various molecular relaying strategies such as sense-and-forward, amplify-and-forward (AF) and decode-and-forward (DF) have been explored in the prevailing MC literature [11]-[17] to enhance the communication distance.

Realizing that the transmitted molecular pulse attenuates rapidly in comparison to wireless EM pulse, the received molecular signals in a DMC system may become less distinguishable. This effect would be more prominent in a DMC system that employs modulation scheme having reduced molecular gap between the adjacent symbols (e.g., budget limited system with higher order modulation) or when relay nanomachine (RN) is placed closer to R_xN. In such scenarios, the aforementioned relaying strategies may suffer from error propagation (due to hard decision in DF) or increased molecular noise (due to amplification of received signal in AF) at the RN. On the contrary, forwarding a soft information of symbol released by the transmitter nanomachine (T_xN) may help in performance enhancement of relay-assisted DMC systems.

More importantly, design parameters such as detection threshold, relay location and number of molecules at each of the transmitting nodes should be chosen appropriately in order to get optimal performance of the system. If they are not chosen appropriately, either probability of false alarm or probability of miss detection will increase. Methodologies such as bisection method for a quasi-convex problem and MATLAB-based interior point algorithm for a convex problem have been proposed in the literature [14], [17], [18] to optimize the detection threshold in DMC systems. These methodologies either give suboptimal threshold or optimal one whereof the convergence depends on choosing the smallest interval that contains the optimal threshold. In addition to above works, research is needed to develop efficient optimization techniques for DMC systems that can guarantee optimal detection threshold in few iterations. Such techniques are required especially for applications requiring real-time optimization, considering the limited memory and processing capability of the nanomachines.

Above all, prior works [14], [19], [20] have considered optimization of molecules allocation and relaying positioning as separate problems, however their joint optimization for the given detection thresholds has not been yet investigated. Note that molecules are limited resource because of the finite availability of molecule synthesizing energy and limited storage capabilities of the reservoir. In fact, such investigation in the context of bio-nanosensor networks [21] is essential for their practical implementation of local drug delivery system where drug molecules may be expensive and their large amount can

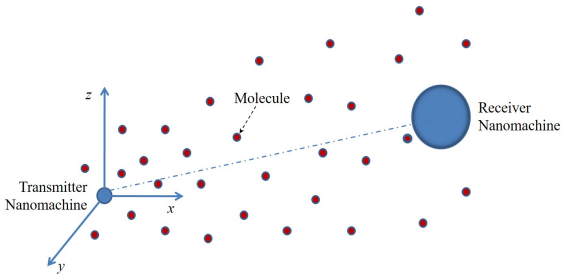


Fig. 1. Signal propagation and schematic diagram for a DMC System.

have ill effects on healthy parts of the body. More importantly, the emission of an arbitrary number of molecules would increase multi-source interference for other nanomachines present in the medium. Therefore, optimal allocation of these molecules and RN placement for the given detection thresholds would reduce the network error without increasing much the complexity of the nanomachines (i.e., nanomachines neither need to update their detection thresholds nor require high computational cost involved with the maximum likelihood (ML) detection).

B. Solution Approach

With regard to addressing the first problem related to this thesis, first we try to mitigate residual noise by adjusting design parameters based on SNR estimate at the R_xN. Considering molecular concentration-encoded signal with residual noise, we present the SNR estimation using ML principle with sampled observations. Moreover, we derive the Cramer-Rao lower bound (CRLB) and assess the performance of our proposed estimator in terms of its mean square error (MSE). Then as a solution approach for the second problem of this thesis, we propose and investigate a new relaying strategy known as estimate-and-forward (EF) that improves communication range, molecular throughput and latency of the DMC system. The proposed relaying scheme forwards an estimate of the transmitted number of molecules, which is derived using ML principle. Thereafter, for solution of the third problem, we came up with a new approach that yields optimal detection threshold for convex optimization problem in DMC system using logarithmic barrier followed by modified Karush-Kuhn-Tucker conditions, Newton Raphson method, and finally rounding the solution to the nearest integer value. Finally, we solve the joint molecules allocation and relaying positioning optimization problem by using block coordinate descent algorithm (BCDA) which relies on the concept of fixing all the parameters except one and finding its optimal value that minimizes the objective function. This process is repeated until all the parameters converge.

C. Analytical or Empirical Results

We provide closed-form analytical expressions of the SNR estimator and CRLB of an unbiased estimator for a DMC system (III). The proposed SNR estimator came out to be a function of symbol duration. Interestingly, estimated SNR increases for the increasing symbol duration. This is intuitive since the expected number of information molecules increases while that of residual molecules decreases at the R_xN. However, as symbol duration influences the data rate, a system designer can carefully adjust other parameters (like transmitted

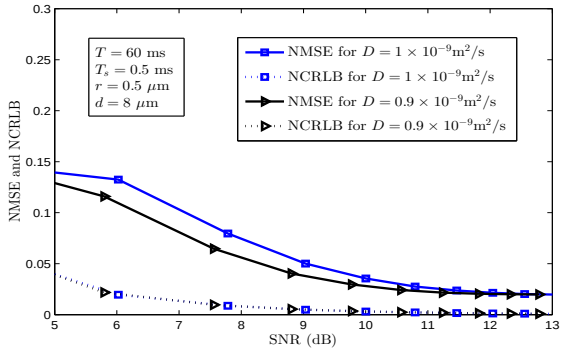


Fig. 2. Normalized MSE and CRLB as a function of SNR (dB) (for details go through (III)).

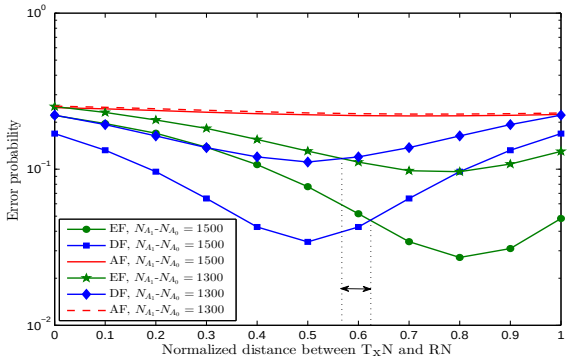


Fig. 3. Error probability of EF, DF, and AF relaying schemes as a function of normalized distance between T_xN and RN (for details go through (I), (V)).

number of molecules) to further reduce the residual noise in order to achieve the required SNR level. Moreover, the proposed SNR estimator is found to be asymptotically optimal for large symbol duration. Such SNR estimation can facilitate the design of a reliable and memoryless DMC system. Further, we assess the performance of EF relaying by deriving analytical expressions for the molecular throughput and end-to-end error probability (I), (V). Our results demonstrate that the proposed EF scheme performs better than the existing AF scheme and also better than the competitive DF scheme when the molecular gap is marginal or when RN is positioned near the R_xN. Moreover, it is found that the EF relaying boosts molecular throughput with the same molecular budget as allocated to the baseline direct transmission. More importantly, EF scheme can be used for *in vivo* diagnosis of abnormalities based on the estimate of released number of specific biomarkers. Above all, without increasing the drug molecules which may be expensive, the proposed relaying strategy boosts the throughput of the drug delivery system for the continued remedial effects. Thereafter, we provide a sequence of analytical formulations leading to the eventual expression for optimal detection threshold (I), (II). Our numerical results illustrate that the proposed threshold optimization technique is more effective than existing approaches which provided either suboptimal detection threshold or optimal one with sluggish convergence for the same value of tolerance. Above all, our analysis and the simulation results reveal that the optimal detection threshold can be achieved efficiently with utmost accuracy. Eventually, our iterative algorithm provides joint

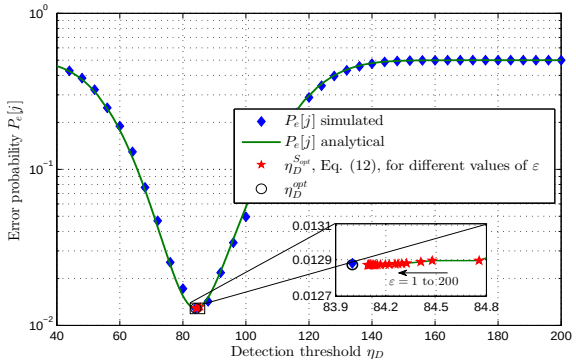


Fig. 4. Error probability $P_e[j]$ and central points $\eta_D^{S_{opt}}$ as a function of detection threshold η_D and accuracy parameter ε respectively (for details go through (II)).

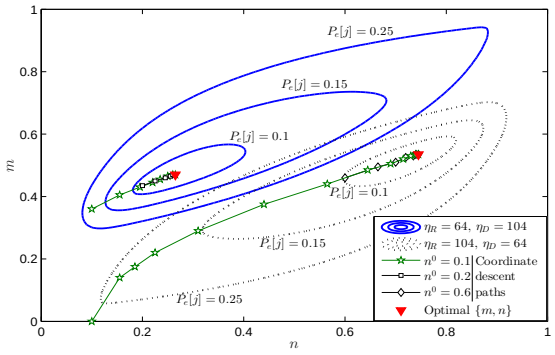


Fig. 5. Contours or level curves of $P_e[j]$ illustrating coordinate descent paths, where m represents relay positioning factor and n denotes the fraction of molecular budget allocated to T_xN (for details go through (IV)).

optimal molecular resource allocation and relay location for the given detection thresholds (IV). Numerical and simulation results reveal the improvement in error performance when molecules distribution and relay placement are in accordance with their joint optimal value. Moreover, it is found that as the relay detection threshold increases, more molecules are needed to be allocated to the source while relay need to be placed closer to the destination in order to satisfy the optimization criteria. We demonstrate the effectiveness of our joint optimization solution through 3D and contour plots illustrating the convergence time. Above all, our analysis helps in designing a reliable and budget limited DMC system with minimal computational requirements at the receiving nanomachines.

ACKNOWLEDGMENT

This work was supported by Council of Scientific and Industrial Research (CSIR), Government of India under the project (No.22(0763)/18/EMR-II).

REFERENCES

[1] T. Nakano, A. Eckford, and T. Haraguchi, *Molecular Communication.* Cambridge University Press, 2013.

[2] N. Farsad, H. B. Yilmaz, A. Eckford, C.-B. Chae, and W. Guo, "A comprehensive survey of recent advancements in molecular communication," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 3, pp. 1887-1919, Feb. 2016.

[3] Q. H. Abbasi, K. Yang, N. Chopra, J. M. Jornet, N. A. Abuali, K. A. Qaraqe, A. Alomainy, "Nano-communication for biomedical applications: A review on the state-of-the-art from physical layers to novel networking concepts," *IEEE Access*, vol. 4, pp. 3920-3935, Jul. 2016.

[4] D. Kilinc and O. B. Akan, "Receiver design for molecular communication," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 12, pp. 705-714, Dec. 2013.

[5] A. Noel, K. C. Cheung, and R. Schober, "Improving receiver performance of diffusive molecular communication with enzymes," *IEEE Trans. NanoBiosci.*, vol. 13, no. 1, pp. 31-43, Mar. 2014.

[6] M. U. Mahfuz, D. Makrakis, and H. T. Mouftah, "Characterization of intersymbol interference in concentration-encoded unicast molecular communication," *in Proc. Canadian Conference on Electrical and Computer Engineering(CCECE)*, May 2011.

[7] H. B. Yilmaz, N.-R. Kim, and C.-B. Chae, "Effect of ISI mitigation on modulation techniques in molecular communication via diffusion," *Proc. ACM NANOCOM*, pp. 3:1-3:9, May 2014.

[8] B. Tepekule, A. E. Pusane, H. B. Yilmaz, C.-B. Chae, and T. Tugcu, "ISI Mitigation Techniques in Molecular Communication," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 1, no. 2, pp. 202-216, Jun. 2015.

[9] B. Li, M. Sun, S. Wang, W. Guo, and C. Zhao, "Low-complexity noncoherent signal detection for nanoscale molecular communications," *IEEE Trans. Nanobiosci.*, vol. 15, no. 1, pp. 3-10, Jan. 2016.

[10] G. Chang, L. Lin, and H. Yan, "Adaptive detection and ISI mitigation for mobile molecular communication," *IEEE Trans. Nanobiosci.*, vol. 17, no. 1, pp. 21-35, Jan. 2018.

[11] A. Einolghozati, M. Sardari, and F. Fekri, "Relaying in diffusion-based molecular communication," *in Proc. IEEE ISIT*, pp. 1844-1848, Jul. 2013.

[12] A. Einolghozati, M. Sardari, and F. Fekri, "Decode and forward relaying in diffusion-based molecular communication between two populations of biological agents," *in Proc. IEEE ICC*, pp. 3975-3980, Jun. 2014.

[13] A. Ahmadzadeh, A. Noel, and R. Schober, "Analysis and design of two-hop diffusion-based molecular communication networks," *in Proc. IEEE GLOBECOM*, pp. 2820-2825, Dec. 2014.

[14] A. Ahmadzadeh, A. Noel, and R. Schober, "Analysis and design of multi-hop diffusion-based molecular communication networks," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 1, no. 2, pp. 144-157, Jun. 2015.

[15] X. Wang, M. D. Higgins, and M.S. Leeson, "Relay analysis in molecular communications with time-dependent concentration," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1977-1980, Nov. 2015.

[16] A. Ahmadzadeh, A. Noel, and R. Schober, "Amplify-and-forward relaying in two-hop diffusion-based molecular communication networks," *in Proc. IEEE GLOBECOM*, pp. 1-7, Dec. 2015.

[17] N. Tavakkoli, P. Azmi, and N. Mokari, "Performance evaluation and optimal detection of relay-assisted diffusion-based molecular communication with drift," *IEEE Trans. NanoBiosci.*, vol. PP, no. 99, Nov. 2016.

[18] Y. Fang, A. Noel, N. Yang, A. W. Eckford, and R. A. Kennedy, "Convex optimization of distributed cooperative detection in multi-receiver molecular communication," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 3, no. 3, pp. 166-182, Sep. 2017.

[19] C. Jiang, Y. Chen, K. J. Ray, "Nanoscale molecular communication networks: A game-theoretic perspective," *EURASIP J. Adv. Signal Process.*, Dec. 2015.

[20] N. Tavakkoli, P. Azmi, and N. Mokari, "Optimal positioning of relay node in cooperative molecular communication networks," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5293-5304, Dec. 2017.

[21] N. Agoulmine, K. Kim, S. Kim, T. Rim, J.-S. Lee, and M. Meyyappan, "Enabling communication and cooperation in bio-nanosensor networks: toward innovative healthcare solutions," *IEEE Wireless Commun.*, vol. 19, no. 5, Oct. 2012.

LIST OF PUBLICATIONS

(I) S. K. Tiwari and P. K. Upadhyay, "Estimate-and-forward relaying in diffusion-based molecular communication networks: Performance evaluation and threshold optimization," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 3, no. 3, pp. 183-193, Sept. 2017.

(II) S. K. Tiwari, T. R. T. Reddy, and P. K. Upadhyay, "Error performance optimization using logarithmic barrier function in molecular nanonetworks," *IEEE Commun. Lett.*, vol. 21, no. 11, pp. 2408-2411, Nov. 2017.

(III) S. K. Tiwari and P. K. Upadhyay, "Maximum likelihood estimation of SNR for diffusion-based molecular communication," *IEEE Wireless Commun. Lett.*, vol. 5, no. 3, pp. 320-323, Jun. 2016.

(IV) S. K. Tiwari, T. R. T. Reddy, P. K. Upadhyay, and D. B. da Costa, "Joint optimization of molecular resource allocation and relay positioning in diffusive nanonetworks," *IEEE Access*, vol. 6, no. 1, pp. 67681-67687, Nov. 2018.

(V) S. K. Tiwari and P. K. Upadhyay, "Estimate-and-forward relaying in molecular communication using brownian motion with drift," in *Proc. International Conference on Signal Processing and Communication (SPCOM)*, IISc Bangalore, India, Jul. 2018, pp. 502-506.

Online multivariate resource usage prediction in cloud datacenters

Shaifu Gupta, Dileep A.D. and Timothy A. Gonsalves
School of Computing and Electrical Engineering
Indian Institute of Technology Mandi
Kamand - 175005, Himachal Pradesh, India
shaifu_gupta@students.iitmandi.ac.in and addileep,tag@iitmandi.ac.in

***Abstract*—Effective resource provisioning requires prediction of future resource usage trends. Training prediction models once is not sufficient to capture the variability in cloud workloads.**

An important goal of the study is to improve resource usage prediction for better resource management using different multivariate resource usage prediction models and propose sparse methods for online adaptation of multivariate resource usage prediction models.

I. Introduction and Motivation

Users dynamically enter and leave cloud datacenters to access services offered by the cloud. This makes resource provisioning a challenging problem and results in over-provisioning and under-provisioning issues where the service providers end up allocating either too many or too few resources respectively.

Insight into future usages of different resources helps in making better decisions for allocating, scheduling of resources and services to clients. Resource usage prediction is formulated as a time series analysis problem where temporal patterns in the past are analyzed to determine the expected trends in upcoming resource usage. However, cloud resource workloads are highly heterogeneous and dynamically varying in nature and therefore, fixed set of history based trained models are not suitable for modeling the dynamically varying trends.

The main contributions of the work are: (i) Performance and stability aware joint framework for multivariate resource usage prediction, (ii) Online adaptation of resource usage prediction models using gradient descent and Levenberg-Marquardt methods, (iii) Sparse models for adaptation of resource usage prediction models in real time.

II. Prior Work

Different time series models have been proposed for resource usage prediction in cloud datacenters. Prominent methods for resource usage prediction in cloud datacenters use Markov models [1], linear autoregressive models [2], neural networks [3], and LSTM models [4] for future resource usage prediction in cloud. Most of the existing methods utilize past resource usage of the desired resource metric itself for prediction and ignore the effects of other features. Therefore, in our approach we focus on multivariate models for future resource prediction and propose sparse multivariate models for real time adaptation and prediction of resource usage. An overview of the proposed approaches is presented in next subsection.

III. The solution approach

Multivariate approaches attempt to closely model the reality where each decision involves analysis of more than one variable. In our work, we proposed a performance and stability aware joint framework for multivariate resource usage prediction. In this framework, we analyzed multivariate extensions of different state-of-the-art prediction models ([1], [2], [3], [4]) along with multiple feature selection methods. Since cloud workloads are highly time varying in nature, we proposed two desirable characteristics for the selection of relevant features by a feature selection technique. These are (i) the prediction performance and, (ii) sensitivity of the feature selection technique to changes in the data. Based on the analysis, we observed that a combination of Granger causality based feature selection technique and bidirectional long short term memory network (BLSTM) works best.

Adaptive models play an important role for modeling the real-time processes that exhibit time-variant behavior. Real time adaptation of resource usage prediction models is essential to model the changing trends in the workloads and generate more accurate predictions. We observed that BLSTM models generate best predictions as compared to other models but have a large number of learnable parameters due to the three gating functions involved as well as deep stacking of layers. The presence of a large number of trainable parameters increases the computational overhead of real time adaptation of these models. Therefore, we propose sparse BLSTMs for real time resource usage prediction and adaptation in cloud.

The core idea for obtaining a sparse resource usage prediction model is to start from a pre-trained dense network and remove less influential parameters of the model to build a parameter-sparse model. Magnitude based and error based pruning methods have been analyzed to introduce sparsity into the network. In magnitude based pruning methods, absolute value of weights reflect the significance of the associated connections. Whereas, in error based methods, gradient of the training error with respect to the weights is used to determine the significance of the connections. Under these two categories, ℓ^1 regularization, optimal brain damage and hybrid

of these two methods have been analyzed for introducing sparsity in the models. An iterative hard thresholding based approach is used with three sparsification methods for pruning the weights of the model. In the iterative approach, the weights of the model are pruned in multiple iterations rather all at once. The pruning and readjustment stages are used alternately in each iteration. The pruning stage removes less significant connections from the network. In the readjustment stage, the remaining sparse net work is retrained to adjust the remaining weight connections and compensate for the connections that have been removed. The built sparse model is then used for resource usage prediction and adaptation in real time.

In our work, we studied the online adaptation of resource usage prediction models based on autoregressive integrated moving average models (ARIMA), non-linear autoregressive neural networks and BLSTM networks for resource usage prediction models to capture the changes in incoming workload patterns and update the model. We use gradient descent and Levenberg-Marquardt methods for online adaptation of the prediction models. An observation window of the resource usage values is defined and after every observation window, the error between the predicted and newly observed actual resource usage is used to adjust the learned weights of the models to generate predictions ahead.

The results of the proposed prediction methods are validated on Google cluster trace [5] and presented in next section.

IV. Results

To evaluate the effectiveness of our framework, we apply it to the Google cluster trace. Google cluster trace presents resource workload data of a cluster of 12,500 machines. In this work, the resource usage metrics are analyzed to predict the expected resource workload on the cluster. CPU usage is one of the key resource metrics and is usually the first place we look when we observe the sign of jobs slowing down. Therefore, we focus on prediction of future CPU usage trends. In addition to CPU usage, several other resource metrics based on memory usage and disk input/output are used for multivariate CPU usage prediction. In the present work, all the 13 monitored resource usage metrics are chosen for analyzing their behavior during the study.

In the present study, we divide the data in training, validation and test set. The 7 days of resource workload data is used for training and validating the performance of prediction models. Resource usage values are aggregated at 10 second time interval. We use a machine with 40 Intel Xeon CPU cores and 128 GB RAM for training. Root mean square error (RMSE) is used to evaluate the accuracy of prediction results.

Table I compares the predictions of the dense BLSTM model and different sparse BLSTM models generated using three sparsification methods. The predictions of different models are analyzed at 60, 120 and 180 steps ahead. From the Table, it is observed that sparse BLSTM models have higher prediction error than the dense model. In our analysis, we discard 70% of the connections to build sparse models, this

TABLE I
RMSE of CPU usage prediction by different BLSTM models.

Model	60 steps	120 steps	180 steps
BLSTM-dense	0.0095	0.0184	0.0255
BLSTM-ℓ^1	0.0250	0.0243	0.0300
BLSTM-OBD	0.0179	0.0348	0.0563
BLSTM-ℓ^1-OBD	0.0245	0.0250	0.0353

results in an increase in the prediction error of sparse models. Among the sparse models, overall, the ℓ^1 regularization method generates the better set of predictions as they are not much affected by increase in prediction steps, making it optimal for the service providers to adapt the model at any adaptation window.

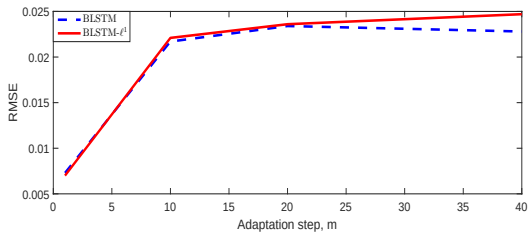


Fig. 1. Comparison of RMSE in CPU forecast using BLSTM and BLSTM-ℓ^1 at different adaptation steps.

Figure 1 compares the error in the predictions of the dense and sparse BLSTM model after adaptation at different steps. From the Figure, it is seen that at higher steps, the predictions of the BLSTM-ℓ^1 are bad as compared to that of the dense model. But as we decrease the steps and adapt the model more frequently, the predictions of the sparse model move closer to the dense model. Comparing the adaptation times of the sparse and dense models, it is seen that the execution time for sparse model is 60% lesser as compared to the dense model.

V. Conclusion

Cloud workloads are highly time varying. Real-time adaptation helps in improving the predictions and aids the service providers to get the resources ready for a possible change in resource usage trends. Multivariate sparse BLSTM models are proposed for online prediction in cloud datacenters. It is observed that the prediction error in sparse BLSTM models is slightly higher than the dense model. But sparse BLSTM model improves upon the accuracies of prediction after regular adaptations. The time required for sparse and dense BLSTM models is compared and it is observed that being sparse, real time adaptations are fast in the pruned model.

References

[1] Z. Gong, X. Gu, and J. Wilkes, "PRESS: PRedictive Elastic ReSource Scaling for cloud systems," in *International Conference on Network and Service Management (CNSM)*, 2010, pp. 9–16.

[2] Q. Zhang, M. F. Zhani, S. Zhang, Q. Zhu, R. Boutaba, and J. L. Hellerstein, "Dynamic energy-aware capacity provisioning for cloud computing environments," in *International Conference on Autonomic Computing (ICAC)*, New York, NY, USA, 2012, pp. 145–154.

[3] F. Caglar and A. Gokhale, "iOverbook: Intelligent resource-overbooking to support soft real-time applications in the cloud," in *7th International Conference on Cloud Computing (CLOUD)*. Anchorage, AK, USA: IEEE, 2014, pp. 538–545.
[4] B. Song, Y. Yu, Y. Zhou, Z. Wang, and S. Du, "Host load prediction with long short-term memory in cloud computing," *The Journal of Supercomputing*, pp. 1–15, 2017.
[5] C. Reiss, J. Wilkes, and J. L. Hellerstein, "Google cluster-usage traces: format + schema," Nov. 2011.

List of publications

I. S. Gupta, A.D. Dileep and T.A. Gonsalves,"Fractional Difference based Hybrid Model for Resource Prediction in Cloud Network", in proceedings of 5^{th} *ACM International Conference on Network, Communication and Computing (ICNCC 2016)*, Japan, pp. 93-97, 2016.

II. S. Gupta and A.D. Dileep,"Online Adaptation Models for Resource Usage Prediction in Cloud Network", in proceedings of 23^{rd} *IEEE National Conference on Communications (NCC 2017)*, Chennai, pp. 1-6, 2017.

III. S. Gupta and A.D. Dileep,"Resource Usage Prediction of Cloud Workloads using Deep Bidirectional Long Short Term Memory Networks", in proceedings of 11^{th} *IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS 2017)*, Bhubaneswar, pp. 1-6, 2017.

IV. S. Gupta, N. Muthiyan, S. Kumar, A. Nigam, and A.D. Dileep, "A Supervised Deep Learning Framework for Proactive Anomaly Detection in Cloud Workloads", in proceedings of 14^{th} *IEEE India Council International Conference (INDICON 2017)*, Roorkee, pp. 1-6, 2017.

V. S. Gupta, Dileep A.D. and T.A. Gonsalves, "A Joint Multivariate Feature Selection Framework for Resource Workload Prediction in Cloud using Stability and Prediction Performance", Journal of Supercomputing, Springer, Volume 74, Issue 11, pp. 6033-6068, 2018.

Small World Models for Development of Wireless Sensor Network Services

Om Jee Pandey and Rajesh M. Hegde
Department of Electrical Engineering
Indian Institute of Technology, Kanpur, India
Email: {ojpandey, rhegde}@iitk.ac.in

***Abstract*—**A wireless sensor network (WSN) [1], [2] is a collection of large number of low-cost, low-power, and smart wireless sensor nodes, deployed over a geographical area for monitoring temperature, pressure, humidity, light, vibrations, and seismic events. There are several ubiquitous WSN applications in our daily lives like health care monitoring, environmental surveillance, natural disaster prevention, water quality controlling, forest fire detection, and marine navigation [3]–[5]. The advent of Internet of Things (IoT) and Cyber-Physical Systems (CPS) have further increased the significance of WSN applications [6]–[10]. Wireless sensor networks middleware [11]–[13] generally refers to the layer in between the application layer and the hardware, and is very crucial in the efficient design of a WSN application. Services that run at this interface are called WSN services [14], [15]. Node localization [16]–[18], data gathering [19]–[21], time synchronization [22]–[24], and low-latency and energy-balanced data transmission [25]–[27] are some of the WSN services that are vital in real time and surveillance applications such as CPS, IoT, and context-aware pervasive systems. Efficient WSN services for CPS and IoT can lead to the rapid development of smart homes, smart cities, smart energy grids, intelligent transportation, and climate smart agriculture [28]–[31].

In order to develop efficient WSN services, several new methods have been proposed for node localization [32], [33], clustering of sensor nodes [34], [35], new routing algorithms [36], [37], adaptive duty cycles [38], [39], usage of mixed-transmission models [40], [41], placement of intelligent gateways [42], usage of mobile elements [25], [43], [44], and bypassing holes [45]. However, introduction of small world characteristics [46]–[55] in a WSN has hitherto not been used in this context. A small world network [56]–[59] is typically characterized by a low average path length and high average clustering coefficient and has been widely used to model social networks. However, small world models [60], [61] have not been used in the development of wireless sensor network. The primary objective of this thesis is to develop small world models for WSN services like sensor node localization [62], data gathering, energy-balancing, data transmission delay, and time synchronization. A small world WSN reduces the number of hops required for data transmission, and maximal utilization of sensor nodes closer to the sink. Other advantages of small world WSN include efficient time synchronization, increased network lifetime, and reduced data latency. Small world WSN developed in tandem with novel routing strategies leads to improved WSN services for various applications.

A historical background to small world networks [56]–[59] in social, biological, transportation, and computer networks is first provided in the thesis to highlight its significance. Subsequently, the development of small world models in the context of WSN is discussed. The problem of small world WSN development is formulated as an optimization problem with constraints on both average path length and average clustering coefficient. Generic solutions to the problem are also outlined with some illustrative examples over a WSN testbed.

This thesis makes three key contributions to the state of art in WSN services. The first contribution of this thesis is in the development of novel localization methods over small world WSN. Three new methods for sensor node localization over small world WSN are proposed. Small world models are developed in these methods by introducing new radio links in the network using heterogeneous sensor nodes and by imposing certain constraints on the average path length, average clustering coefficient, and localization error. The first localization method over small world WSN is developed for both model based and model free radio propagation and utilizes a constrained least squares approach. On the other hand, the second localization method is cooperative and imposes a maximal differential constraint to obtain the small world model. Additionally, a time synchronized localization method over small world WSN is also proposed. Performance evaluation in terms of localization error, power consumption, and bandwidth and anchors required for node localization is conducted. Cramer-Rao lower bound (CRLB) analysis for node location parameters is used to provide insights into the error bounds obtained using the proposed methods. Experimental results obtained indicate that the proposed small world model yields improved localization results when compared to results obtained over a regular WSN.

The second contribution of this thesis is in the development of novel method for joint localization and data gathering over small world wireless sensor network. An optimal data MULE (Mobile Ubiquitous LAN Extension) allocation method is proposed for small world WSN development in this work. This method computes both the optimal number of data MULEs and their placement in the network by minimizing an objective function which is the normalized-weighted sum of network pa-

rameters. A cooperative localization and data gathering method using the principle of multidimensional scaling is then realized over the small world WSN. The performance of the proposed method is evaluated by conducting exhaustive analysis of power consumption, bandwidth required, localization error, data latency, and throughput. Experimental results obtained indicate a significant improvement on several evaluation parameters when compared to results obtained on conventional WSN.

The third contribution of this thesis is the development of low-latency and energy-balanced data transmission over cognitive small world wireless sensor network. A small world model is developed in this work by iteratively adding a new set of radio links between the nodes in a network by maximizing the probability of link addition. The probability of link addition is itself calculated from six crucial network parameters which leads to the development of a cognitive small world model. A new data routing method over this small world wireless sensor network is also proposed by optimizing energy cost of the radio links in the network. Experiments are conducted using simulations and real node deployments over a WSN testbed. Experimental results obtained indicate that the proposed cognitive small world model improves energy balancing, network lifetime, and reduces data latency when compared to conventional methods in literature.

The performance of the key methods proposed in this thesis in the context of WSN services is motivating enough to be used in medium scale WSN and in practical applications like IoT and CPS among others. Development of small world models in the context of large scale WSN testbeds needs to be investigated further. The use of frequency selective radios in realistic WSN testbeds can also be explored in future. Advantages of small world phenomena can also be investigated in the context of distributed detection and estimation over WSNs. Certain other array signal processing techniques like collaborative transmit beamforming can be used to develop small world WSN to further improve the quality of WSN services.

Index Terms— : Wireless sensor network (WSN), small world characteristics (SWC), small world network (SWN), small world wireless sensor network (SW-WSN), WSN services, sensor node localization, time synchronization, data gathering, low-latency data transmission, energy-balancing, network lifetime maximization, cognition.

LIST OF PUBLICATIONS

- **In Peer Reviewed International Journals (Published)**
  1) Om Jee Pandey, and Rajesh M. Hegde, "*Low-Latency and Energy-Balanced Data Transmission over Cognitive Small World WSN.*" **IEEE Transactions on Vehicular Technology**, vol. 67, no. 8, pp. 7719-7733, August 2018.
  2) Om Jee Pandey, Akshay Mahajan, and Rajesh M. Hegde, "*Joint Localization and Data Gathering over a Small-World WSN with Optimal Data MULE Allocation.*" **IEEE Transactions on Vehicular Technology**, vol. 67, no. 7, pp. 6518-6532, July 2018.
  3) Om Jee Pandey, and Rajesh M. Hegde, "*Node Localization over Small World WSNs using Constrained Average Path Length Reduction.*" **Elsevier Ad Hoc Networks**, vol. 67, pp.87-102, December 2017.
  4) Om Jee Pandey, and Rajesh M. Hegde, "*Cooperative Localization over Small World WSN using Optimal Allocation of Heterogeneous Nodes.*" **IET Wireless Sensor Systems**, vol. 8, no. 4, pp. 162-169, August 2018.
  5) Om Jee Pandey, Richika Sharan, and Rajesh M. Hegde, 2017, "*Localization in Wireless Sensor Networks Using Visible Light in Non-Line of Sight Conditions.*" **Springer, Wireless Personal Communications**, vol. 97, no. 4, pp. 6519-6539, August 2017.

- **In Peer Reviewed International/National Conferences (Published)**
  1) Om Jee Pandey, Akshay Mahajan, and Rajesh M. Hegde, "*Cooperative Localization in Small World Wireless Sensor Networks.*" **In 9th IEEE International Conference on Communication Systems and Networks (COMSNETS)**, pp. 391-392, January 2017.
  2) Om Jee Pandey, Ashutosh Kumar, and Rajesh M. Hegde, "*Localization in Wireless Sensor Networks with Cognitive Small World Characteristics.*" **In 22nd National Conference on Communication (NCC)**, pp. 1-6, March 2016.

REFERENCES

[1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer networks*, vol. 52, no. 12, pp. 2292–2330, 2008.
[2] B. Rashid and M. H. Rehmani, "Applications of wireless sensor networks for urban areas: A survey," *Journal of Network and Computer Applications*, vol. 60, pp. 192–219, 2016.
[3] G. Xu, W. Shen, and X. Wang, "Applications of wireless sensor networks in marine environment monitoring: A survey," *Sensors*, vol. 14, no. 9, pp. 16 932–16 954, 2014.
[4] A. Z. Abbasi, N. Islam, Z. A. Shaikh *et al.*, "A review of wireless sensors and networks' applications in agriculture," *Computer Standards & Interfaces*, vol. 36, no. 2, pp. 263–270, 2014.
[5] S. Kurt, H. U. Yildiz, M. Yigit, B. Tavli, and V. C. Gungor, "Packet size optimization in wireless sensor networks for smart grid applications," *IEEE Trans. on Ind. Electronics*, vol. 64, no. 3, pp. 2392–2401, 2017.
[6] F.-J. Wu, Y.-F. Kao, and Y.-C. Tseng, "From wireless sensor networks towards cyber physical systems," *Pervasive and Mobile computing*, vol. 7, no. 4, pp. 397–413, 2011.
[7] J. A. Stankovic, "Research directions for the internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 3–9, 2014.
[8] S. Li, L. Da Xu, and S. Zhao, "5g internet of things: A survey," *Journal of Industrial Information Integration*, 2018.
[9] F. Tao, J. Cheng, and Q. Qi, "Iihub: An industrial internet-of-things hub toward smart manufacturing based on cyber-physical system," *IEEE Trans. on Ind. Informatics*, vol. 14, no. 5, pp. 2271–2280, 2018.
[10] S. Li, Q. Ni, Y. Sun, G. Min, and S. Al-Rubaye, "Energy-efficient resource allocation for industrial cyber-physical iot systems in 5g era," *IEEE Trans. on Ind. Informatics*, vol. 14, no. 6, pp. 2618–2628, 2018.
[11] W. B. Heinzelman, A. L. Murphy, H. S. Carvalho, and M. A. Perillo, "Middleware to support sensor network applications," *IEEE network*, vol. 18, no. 1, pp. 6–14, 2004.
[12] S. Hadim and N. Mohamed, "Middleware: Middleware challenges and approaches for wireless sensor networks," *IEEE distributed systems online*, vol. 7, no. 3, pp. 1–1, 2006.

[13] R. A. Alshinina and K. M. Elleithy, "A highly accurate deep learning based approach for developing wireless sensor network middleware," *IEEE Access*, 2018.

[14] S. Krco, M. Johansson, V. Tsiatsis, I. Cubic, K. Matusikova, and R. Glitho, "Mobile network supported wireless sensor network services," in *Mobile Adhoc and Sensor Systems, 2007. MASS 2007. IEEE International Conference on*. IEEE, 2007, pp. 1–3.

[15] L. Cheng, J. Niu, C. Luo, L. Shu, L. Kong, Z. Zhao, and Y. Gu, "Towards minimum-delay and energy-efficient flooding in low-duty-cycle wireless sensor networks," *Computer Networks*, vol. 134, pp. 66–77, 2018.

[16] F. Xiao, W. Liu, Z. Li, L. Chen, and R. Wang, "Noise-tolerant wireless sensor networks localization via multinorms regularized matrix completion," *IEEE Trans. on Veh. Technl.*, vol. 67, no. 3, pp. 2409–2419, 2018.

[17] Y. Kim, B. Lee, H. So, and S.-C. Kim, "Localization technique considering position uncertainty of reference nodes in wireless sensor networks," *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1324–1332, 2018.

[18] A. K. Alavijeh, M. H. Ramezani, and A. K. Alavijeh, "Localization improvement in wireless sensor networks using a new statistical channel model," *Sensors and Actuators A: Physical*, vol. 271, pp. 283–289, 2018.

[19] C. Zhan, Y. Zeng, and R. Zhang, "Energy-efficient data collection in uav enabled wireless sensor network," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 328–331, 2018.

[20] Y. Li, X. Zheng, J. Liu, N. Hu, and G. Yang, "Application of energy-efficient data gathering to wireless sensor network by exploiting spatial correlation," *Sensors and Materials*, vol. 30, no. 3, pp. 577–585, 2018.

[21] C.-F. Cheng and C.-F. Yu, "Mobile data gathering with bounded relay in wireless sensor networks," *IEEE Internet Things J.*, 2018.

[22] K. S. Yıldırım, R. Carli, and L. Schenato, "Adaptive proportional–integral clock synchronization in wireless sensor networks," *IEEE Trans. on Contr. Syst. Technol.*, vol. 26, no. 2, pp. 610–623, 2018.

[23] T. Qiu, Y. Zhang, D. Qiao, X. Zhang, M. L. Wymore, and A. K. Sangaiah, "A robust time synchronization scheme for industrial internet of things," *IEEE Trans. on Ind. Inf.*, vol. 14, no. 8, pp. 3570–3580, 2018.

[24] K. Xie, Q. Cai, and M. Fu, "A fast clock synchronization algorithm for wireless sensor networks," *Automatica*, vol. 92, pp. 133–142, 2018.

[25] A. Kinalis, S. Nikoletseas, D. Patroumpa, and J. Rolim, "Biased sink mobility with adaptive stop times for low latency data collection in sensor networks," *Information fusion*, vol. 15, pp. 56–63, 2014.

[26] Z.-t. Li, Q. Chen, G.-m. Zhu, Y.-j. Choi, and H. Sekiya, "A low latency, energy efficient mac protocol for wireless sensor networks," *Int. J. of Distr. Sens. Net.*, vol. 11, no. 8, p. 946587, 2015.

[27] N. Jan, N. Javaid, Q. Javaid, N. Alrajeh, M. Alam, Z. A. Khan, and I. A. Niaz, "A balanced energy-consuming and hole-alleviating algorithm for wireless sensor networks," *IEEE Access*, vol. 5, pp. 6134–6150, 2017.

[28] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things journal*, vol. 1, no. 1, pp. 22–32, 2014.

[29] L. Dan, C. Xin, H. Chongwei, and J. Liangliang, "Intelligent agriculture greenhouse environment monitoring system based on iot technology," in *Intelligent Transportation, Big Data and Smart City (ICITBS), 2015 International Conference on*. IEEE, 2015, pp. 487–490.

[30] S. P. Mohanty, U. Choppali, and E. Kougianos, "Everything you wanted to know about smart cities: The internet of things is the backbone," *IEEE Consumer Electronics Magazine*, vol. 5, no. 3, pp. 60–70, 2016.

[31] C. Choi, C. Esposito, H. Wang, Z. Liu, and J. Choi, "Intelligent power equipment management based on distributed context-aware inference in smart cities," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 212–217, 2018.

[32] S. K. Gharghan, R. Nordin, A. M. Jawad, H. M. Jawad, and M. Ismail, "Adaptive neural fuzzy inference system for accurate localization of wireless sensor network in outdoor and indoor cycling applications," *IEEE Access*, vol. 6, pp. 38 475–38 489, 2018.

[33] O. Cheikhrouhou, G. M Bhatti, and R. Alroobaea, "A hybrid dv-hop algorithm using rssi for localization in large-scale wireless sensor networks," *Sensors*, vol. 18, no. 5, p. 1469, 2018.

[34] Y. Zhang, J. Wang, D. Han, H. Wu, and R. Zhou, "Fuzzy-logic based distributed energy-efficient clustering algorithm for wireless sensor networks," *Sensors*, vol. 17, no. 7, p. 1554, 2017.

[35] P. Nayak and B. Vathasavai, "Energy efficient clustering algorithm for multi-hop wireless sensor network using type-2 fuzzy logic," *IEEE Sensors Journal*, vol. 17, no. 14, pp. 4492–4499, 2017.

[36] T.-T. Huynh, T.-N. Tran, C.-H. Tran, and A.-V. Dinh-Duc, "Delay constraint energy-efficient routing based on lagrange relaxation in wireless sensor networks," *IET Wirel. Sens. Sys.*, vol. 7, no. 5, pp. 138–145, 2017.

[37] A. Khasawneh, M. S. B. A. Latiff, O. Kaiwartya, and H. Chizari, "A reliable energy-efficient pressure-based routing protocol for underwater wireless sensor network," *Wireless Networks*, vol. 24, no. 6, pp. 2061–2075, 2018.

[38] Z. Chen, A. Liu, Z. Li, Y.-j. Choi, and J. Li, "Distributed duty cycle control for delay improvement in wireless sensor networks," *Peer-to-Peer Networking and Applications*, vol. 10, no. 3, pp. 559–578, 2017.

[39] C. Jiang, T.-S. Li, J.-B. Liang, and H. Wu, "Low-latency and energy-efficient data preservation mechanism in low-duty-cycle sensor networks," *Sensors*, vol. 17, no. 5, p. 1051, 2017.

[40] W. Guo, Z. Liu, and G. Wu, "An energy-balanced transmission scheme for sensor networks," in *Proceedings of the 1st international conference on Embedded networked sensor systems*. ACM, 2003, pp. 300–301.

[41] T. Liu, T. Gu, N. Jin, and Y. Zhu, "A mixed transmission strategy to achieve energy balancing in wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2111–2122, 2017.

[42] W. Youssef and M. Younis, "Intelligent gateways placement for reduced data latency in wireless sensor networks," in *Commun., 2007. ICC'07. IEEE Int. Conf. on*. IEEE, 2007, pp. 3805–3810.

[43] C. Wang, J. Li, F. Ye, and Y. Yang, "A mobile data gathering framework for wireless rechargeable sensor networks with vehicle movement costs and capacity constraints," *IEEE Transactions on Computers*, vol. 65, no. 8, pp. 2411–2427, 2016.

[44] Y.-H. Kao, B. Krishnamachari, M.-R. Ra, and F. Bai, "Hermes: Latency optimal task assignment for resource-constrained mobile computing," *IEEE Trans. on Mobile Compu.*, vol. 16, no. 11, pp. 3056–3069, 2017.

[45] F. Zhou, G. Trajcevski, R. Tamassia, B. Avci, A. Khokhar, and P. Scheuermann, "Bypassing holes in sensor networks: Load-balance vs. latency," *Ad Hoc Networks*, vol. 61, pp. 16–32, 2017.

[46] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[47] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[48] M. E. Newman and D. J. Watts, "Renormalization group analysis of the small-world network model," *Physics Letters A*, vol. 263, no. 4, pp. 341–346, 1999.

[49] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the national academy of sciences*, vol. 97, no. 21, pp. 11 149–11 152, 2000.

[50] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, pp. 845–845, 2000.

[51] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical review letters*, vol. 87, no. 19, p. 198701, 2001.

[52] A. Helmy, "Small worlds in wireless networks," *Communications Letters, IEEE*, vol. 7, no. 10, pp. 490–492, 2003.

[53] D. Cavalcanti, D. Agrawal, J. Kelner, and D. Sadok, "Exploiting the small-world effect to increase connectivity in wireless ad hoc networks," in *Int. Conf. on Telecommun.* Springer, 2004, pp. 388–393.

[54] G. Sharma and R. R. Mazumdar, "A case for hybrid sensor networks," *IEEE/ACM Trans. on Netw.*, vol. 16, no. 5, pp. 1121–1132, 2008.

[55] A. Banerjee, R. Agarwal, V. Gauthier, C. K. Yeo, and H. Afifi, "A self-organization framework for wireless ad hoc networks as small worlds," *IEEE Trans. on Veh. Technol.*, vol. 61, no. 6, pp. 2659–2673, 2012.

[56] M. Jovanović, F. Annexstein, and K. Berman, "Modeling peer-to-peer network topologies through small-world models and power laws," in *IX Telecommunications Forum, TELFOR*, 2001, pp. 1–4.

[57] P. Sen, S. Dasgupta, A. Chatterjee, P. Sreeram, G. Mukherjee, and S. Manna, "Small-world properties of the indian railway network," *Physical Review E*, vol. 67, no. 3, p. 036106, 2003.

[58] R. Chitradurga and A. Helmy, "Analysis of wired short cuts in wireless sensor networks," in *Pervasive Services, 2004. ICPS 2004. Proceedings. The IEEE/ACS International Conference on*. IEEE, 2004, pp. 167–177.

[59] C.-J. Jiang, C. Chen, J.-W. Chang, R.-H. Jan, and T. C. Chiang, "Construct small worlds in wireless networks using data mules," in *Sensor Networks, Ubiquitous and Trustworthy Computing, 2008. SUTC'08. IEEE International Conference on*. IEEE, 2008, pp. 28–35.

[60] A. Barrat and M. Weigt, "On the properties of small-world network models," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 13, no. 3, pp. 547–560, 2000.

[61] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[62] N. Jiang, C. Huan *et al.*, "A localization scheme of wireless sensor networks based on small world effects." *Advances in Information Sciences & Service Sciences*, vol. 3, no. 11, 2011.

libVNF: Library to build Virtual Network Functions

Priyanka Naik, Mythili Vutukuru
Department of Computer Science and Engineering, Indian Institute of Technology, Bombay
Email: {ppnaik, mythili}@cse.iitb.ac.in

I. Problem Statement

Network Function Virtualization(NFV) is a trend to replace network functions traditionally hosted on dedicated hardware to be virtualized into software components running on virtual machines. NFV is cost-efficient and provides the flexibility of scaling on demand. Recent techniques of high-performance network stacks has driven the adoption of NFV. The network functions being considered for virtualization range from middleboxes operating at the layer 3 of the network stack like network address translator(NAT), load balancers to components working at the application layer which terminate transport connections like telecommunication middleboxes.

Our work focuses on building high performance virtual network functions(VNFs). For NFV to be a cost saving paradigm, a VNF developer should just focus on the core processing logic. The other parts of the VNF code which are common across most VNFs, like communication and state management should be handled by an optimized framework. We examined an industry grade code of a telecommunication network function (LTE EPC) and observed that about 38% of the code amounted to the read and write of packets. To aid building of such VNFs, we have built a library (libVNF). Unlike existing frameworks for VNF development, our library

- can be used for the development of L2/L3 middleboxes as well as VNFs that are transport layer endpoints
- seamlessly supports multiple network stacks in the backend
- enables distributed implementation of VNFs via functions for distributed state and replica management

libVNF is a C++ library which provides high-level API functions for efficient network communication, for building both L2/L3 middleboxes as well as VNFs that act as transport layer endpoints. The libVNF library implements these API functions over multiple backend network stacks; we currently support the Linux kernel stack and the mTCP [1] userspace stack (over the netmap [2]/DPDK [3] kernel bypass mechanisms). The libVNF communication API is asynchronous and event-driven, in order to leverage event-driven multicore-scalable network stacks like mTCP. VNF developers embed packet processing logic within callback functions that are invoked by the library when events (e.g., packet arrivals) occur on connections. Unlike prior work, libVNF significantly eases the management of application state across callbacks by exposing the abstraction of an application-layer request to the VNF developer. With libVNF, VNF developers can store state across multiple connections and callbacks in a single request object that is efficiently managed by the library. libVNF can also be used to easily build distributed clustered implementations of a VNF; our API provides functions to store and retrieve state in a shared datastore, and functions to aid scale-out/in notification of a component.

We built several VNFs using our library to show that our API is expressive enough to cater to a wide variety of VNFs, from a simple L3 load balancer to complex VNFs that make up the EPC and IMS subsystems in the packet core of mobile telecom networks. We found that using our library to build VNFs saved up to 50% lines of code in the VNFs, and the performance of VNFs built over libVNF is within 10% of optimized VNFs built without the library. Further, the performance of VNFs built with our library scales well with increasing replicas of a VNF, and with increasing CPU cores within a single replica.

II. Related Work

A range of frameworks have been proposed to build VNFs. But these frameworks cannot be used to easily build scalable implementations of complex VNFs such as the EPC components for the following reasons.

First, most frameworks (e.g., [4], [5], [6]) focus on providing richer APIs for packet header manipulation in L2/L3 middleboxes, and do not have abstractions that support the development of VNFs with transport layer endpoints such as the EPC components. Second, frameworks that support the transport layer endpoint abstraction (e.g., [1], [7]) expose an event-driven socket-like API over a multicore-scalable userspace stack to VNF developers. However, it is well known that writing event-driven code is complicated by the fact that the processing of a single application-layer request is split across multiple callbacks, resulting in the VNF developer having to maintain significant state across callbacks [8]. Therefore, building VNFs with transport layer endpoints over such frameworks still requires significant developer effort.

Third, while some frameworks exist to build horizontally scalable middleboxes (e.g., [9], [10], [11], [12]), none of them comes with a transport layer stack that enables the development of VNFs with transport layer endpoints. Finally,

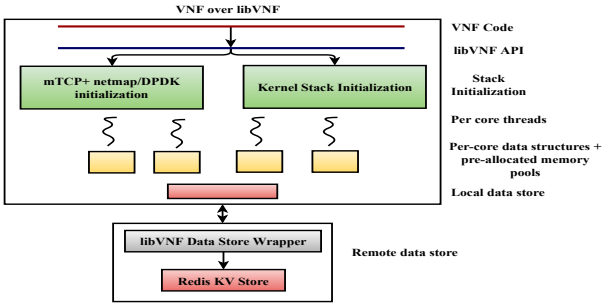


Fig. 1: VNF Architecture with libVNF.

Fig. 2: Example VNF service chain.

existing frameworks force the VNF developer to choose a network stack apriori (e.g., the regular Linux kernel stack or a kernel bypass stack), and VNFs written on one stack are not easily portable to another. However, our experiments with different types of VNFs show that there is no one right choice when it comes to network stacks. For example, when processing a CPU-intensive workload in the VNFs that comprise the IP Multimedia Subsystem (IMS) in telecommunication networks, a kernel bypass mechanism that is non-trivial to setup and configure achieves a throughput gain of only 1.4× over the more readily available Linux kernel stack. On the other hand, for the I/O intensive LTE EPC gateway VNFs, the kernel bypass stack improves throughput by 33× as compared to the Linux kernel stack, because the impact of reducing kernel I/O processing overheads is more pronounced with an I/O intensive workload. These results indicate that VNF developers would benefit from the ability to easily switch and experiment with multiple network stacks, to pick the one that is best suited to the application needs. However, existing VNF development frameworks do not provide this flexibility, and switching network stacks requires significant changes to the VNF code.

III. Solution Approach

The libVNF API consists of 19 functions that are implemented in around 2500 lines of code in an open-source library [13]. Figure 1 illustrates the architecture of a VNF built over our API. Upon initialization, the library spawns one event-driven application thread per core, and the threads use cache-optimized per-core data structures for lockfree operation wherever possible. The library maintains per-core preallocated pools of packet buffers to buffer incoming and outgoing packets of the VNF. We also employ optimizations such as batched transfer of packets to/from the network card wherever possible. The library also preallocates memory pool to manage the request state which is difficult to manage in event-driven stack. Developing applications over non-blocking event-driven APIs such as ours is harder than when using blocking APIs, because application state that would have been on the process stack in the blocking case must now be manually marshalled across callbacks by the developer of the event-driven application [8]. This can be illustrated using a simple example of a service chain consisting of VNFs A, B, and C, shown in Figure 2. Let the communication between the VNFs proceed as follows: `A` generates requests to `B` over a TCP connection. To process `A`'s request, `B` first opens a connection to `C`, sends a request, and waits to get a response back. After receiving `C`'s reply, `B` proceeds to do some computation based on data received from `A` and `C`, and then sends a reply back to `A`. In event-driven stacks, when `B` replies to `A`'s request within the callback invoked upon receiving `C`'s reply, the packet received from `A` is no longer on the stack. Therefore, when not using libVNF, `B` must explicitly allocate memory and maintain data-structures to track this state, and this problem only gets worse as the VNFs get more complex. To ease this process of managing state across callbacks, libVNF provides the abstraction of a request object. A request object can be used to store application state of a single request across callbacks and connections, e.g., VNF `B` can embed the packet received from `A` during the callback within a request object, and retrieve it to generate a reply to `A` in another callback.

libVNF makes it easy to develop VNFs that are implemented as a cluster of replicas for fault tolerance and scalability. Our state management API functions can be used to transparently store and retrieve state across VNF replicas, and our orchestration API can be used to monitor and manage the replicas.

IV. Results

Our evaluation of libVNF comprised of two parts. First, we ran microbenchmarks and showed that the performance of VNFs built with our library is usually within 10% of the performance of optimized implementations built without using the library, and that the performance scales well, both with increase in number of CPU cores in the VNF, as well as increase in the number of replicas of a distributed VNF implementation. Second, we build several complex real-world VNFs over our library, and show that our API is expressive enough to develop a wide variety of VNFs, and that using our library results in up to 50% reduction in development effort (as measured by LoC).

Table I shows that there was low performance overhead in implementing the IMS and EPC VNFs over the library. In case of Layer 3 load balancer, the impact of the additional API calls by the library is higher than the simple logic of the VNF that does very little work per packet, but this overhead was negligible in comparison with the VNF processing in more complex VNFs (IMS and EPC).

VNF	Perf. Overhead of libVNF	LoC saved
IMS (IP Multimedia Subsystem)	3.4%	42%
EPC (LTE-Evolved Packet Core)	5.5%	38%
LB (Layer3 load balancer)	14%	52%

TABLE I: VNF development with and without libVNF.

The experiments we conducted with VNF prototypes built over our API show that, libVNF enables the development of high performance VNFs, saves significant development effort, and introduces minimal overhead. We believe that a library such as ours, if widely used for VNF development, can significantly accelerate the adoption of NFV. libVNF is available on github: https://github.com/networkedsystemsIITB/libVNF

REFERENCES

[1] E. Y. Jeong, S. Woo, M. Jamshed, H. Jeong, S. Ihm, D. Han, and K. Park, "mtcp: A highly scalable user-level tcp stack for multicore systems," in *Proc. of NSDI'14*, 2014.

[2] L. Rizzo, "netmap: A novel framework for fast packet i/o," in *Proc. of USENIX ATC'12*, 2012.

[3] "Intel Data Plane Development Kit," http://dpdk.org/, 2018.

[4] A. Panda, S. Han, K. Jang, M. Walls, S. Ratnasamy, and S. Shenker, "Netbricks: Taking the v out of nfv," in *Proc. of OSDI'16*, 2016.

[5] "Yet Another Network Function Framework," https://www.slideshare.net/MichelleHolley1/ new-model-for-cloud-network-function-development-yanff, 2018.

[6] "Vector packet processing," https://github.com/FDio/vpp, 2018.

[7] "Transport Layer Development Kit," https://github.com/FDio/tldk, 2018.

[8] A. Adya, J. Howell, M. Theimer, W. J. Bolosky, and J. R. Douceur, "Cooperative task management without manual stack management," in *Proc. of ATEC '02*, 2002.

[9] S. Rajagopalan, D. Williams, H. Jamjoom, and A. Warfield, "Split/merge: System support for elastic execution in virtual middleboxes," in *Proc. of NSDI'13*, 2013.

[10] A. Gember-Jacobson, R. Viswanathan, C. Prakash, R. Grandl, J. Khalid, S. Das, and A. Akella, "Opennf: Enabling innovation in network function control," in *Proc. of SIGCOMM'14*, 2014.

[11] M. Kablan, A. Alsudais, E. Keller, and F. Le, "Stateless network functions: Breaking the tight coupling of state and processing," in *Proc. of NSDI'17*, 2017.

[12] S. Woo, J. Sherry, S. Han, S. Moon, S. Ratnasamy, and S. Shenker, "Elastic scaling of stateful network functions," in *Proc. of NSDI'18*, 2018.

[13] "libvnf," https://github.com/networkedsystemsIITB/libVNF, 2018.

V. PUBLICATIONS

- Priyanka Naik, Akash Kanase, Trishal Patel, Mythili Vutukuru, "libVNF: Building Virtual Network Functions Made Easy", ACM Symposium on Cloud Computing (SoCC), 2018.
- Pratik Satapathy, Jash Dave, Priyanka Naik, Mythili Vutukuru, "Performance comparison of state synchronization techniques in a distributed LTE EPC", IEEE Conference on NFV-SDN, 2017.
- Priyanka Naik, Mythili Vutukuru, "libVNF: A Framework for Building Scalable High Performance Virtual Network Functions", ACM Asia-Pacific Workshop on Systems (APSys), 2017.
- Priyanka Naik, Dilip Kumar Shaw, Mythili Vutukuru, "NFVPerf: Online Performance Monitoring and Bottleneck Detection for NFV", IEEE Conference on NFV-SDN, 2016.
- Sugata Sanyal and Priyanka Naik, Increasing Security in Cloud Environment, Annals of Faculty Engineering Hunedoara International Journal of Engineering, vol. no. XI, 2013, pp. 237-240.
- Priyanka Naik and Sugata Sanyal Prover and Verifier Based Password Protection: PVBPP, arxiv.org, arXiV: 1212.6059, 2012

Channel Selection in Dynamic Networks of Unknown Size

Rohit Kumar, NIT Delhi, rohitkumar@nitdelhi.ac.in

I. Problem Definition

The dynamic spectrum access (DSA) algorithms aim to maximize network throughput by ensuring orthogonal channel allocation among secondary users (SUs) in cognitive radio network. Most of the existing DSA algorithms need prior knowledge of the number of active SUs and thus may not be suitable for battery operated resource constrained SU terminals in the infrastructure-less cognitive radio terminals. Another major drawback of these algorithms is that they assume the static network with a fixed number of SUs in the network throughout the horizon and extension of these algorithms for dynamic network is challenging. Whereas the dynamic network is the more practical scenario in which the SUs may enter or exit the network anytime without any prior agreement. Thus, our goal is to design the DSA algorithms for the dynamic networks of unknown size.

II. Relevant State-of-the-Art

The DSA has been well studied for static networks with the need of prior knowledge of number of SUs, U [1–3]. [4] and [5] are the works which do not need the prior knowledge of U but these algorithms also don't consider the dynamic networks scenario. To the best of our knowledge, MEGA [6] and DMC algorithms [7] are the only works which considered the dynamic network with unknown U. DMC algorithm [7] shows to outperform MEGA [6] and is computationally efficient and thus existing state-of-the art. DMC runs in epoch and each epoch consists of two stages: 1) Learning stage- SUs randomly select the channels to observe number of collisions on them which is used to estimate U, 2) MC stage- all SUs are orthogonalized in one of the top channels. DMC[7] has the drawbacks that it suffers from large number of collisions leading to significant loss of throughput and wastage of transmission power. In addition, it also restricts the entry/exit of the SUs during the learning phase of every epoch.

III. Proposed Work

We consider decentralized network consisting of U number of non-cooperative SUs competing for N channels such that $K \geq N$. The probability of vacancy of a channel i is governed by some mean $\mu_i \in [0,1]$, $i \in \{1, 2, ..N\}$ and is initially unknown to all the SUs and assumed to be i.i.d distributed across time slots. The proposed algorithm runs in epochs and restarts after each epoch. The frequency of epoch is set based on duration of the horizon, T. The algorithm requires that all SUs restart the epochs at same time which can be achieved through a global clock as discussed in [7]. The proposed algorithm in run in three phases namely, 1) Channel characterization (CC) phase 2) User Estimation (UE) Phase and 3) Orthogonalization (OR) phase sequentially as shown in Fig. 1. The SUs who enter the network at the start of the horizon i.e. $t = 0$ run all these phases in the first epoch as shown in Fig. 2 (a). Next epoch onwards, these SUs run the UE and OR phase but don't repeat the CC phase. On the other hand, SU entered late in the network, say $t = T_{EN}$ as shown in Fig. 2 (b) completes its CC phase using the long sensing model (Fig. 2 (d)) where it first senses the presence of PU followed by the presence of SU and transmits if both are absent so that they do not collide with the existing SUs. Thereafter it finds a collision free vacant channel and stays on the same channel till the end of OR phase of the next epoch. Afterwards it again finds a collision free vacant channel and stays on on the same channel till the end of that epoch. In the subsequent epoch, it runs only UE and OR phase in each epoch. Now we explain each phase in details.

*1) **CC Phase**:* This phase starts with the random selection of channels by each SU. It continues till a collision free transmission is found by that SU. Next time slot onwards, the SU starts sequential hopping in which the SU selects a channel with next higher index (upto modulo N). In each time slot, the SU performs the channel sensing of the selected channel and confirms whether it is vacant. If found vacant, the SU transmits on that channel. Based on the sensing output, each SU stores the count of the number of times a channel, i is selected (say, S_i) and out of that how many times it has been sensed as vacant (say, V_i). This phase runs for $T_{CC} = (T_{RH} + T_{SH})$ number of time slots. At the end, each SU estimates the channel statistics μ_i by the ratio of V_i to S_i and then uses it to rank the channels.

Lemma 1. *If each SU selects the channel uniformly randomly for* $T_{RH} \geq \frac{\log\left(\frac{\delta_1}{N}\right)}{\log\left(1-\frac{\theta}{4}\right)}$ *time slots, then all the SUs are on non-overlapping channels with probability* $\geq 1 - \delta_1$ $(\delta_1 \in (0,1))$ *where* $\sum_{i=1}^{N} \mu_i/N > \theta$.

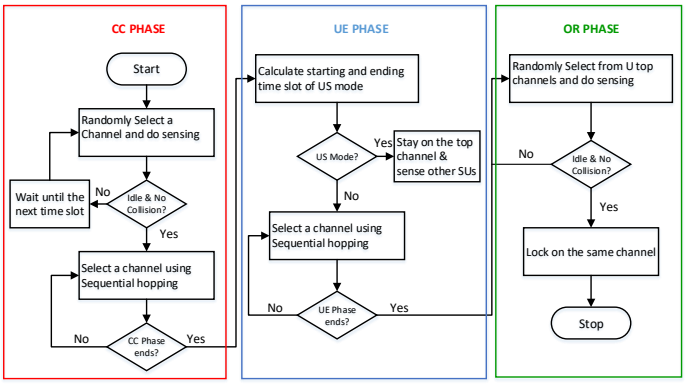


Fig. 1: Different phases of the proposed algorithm for each SU.

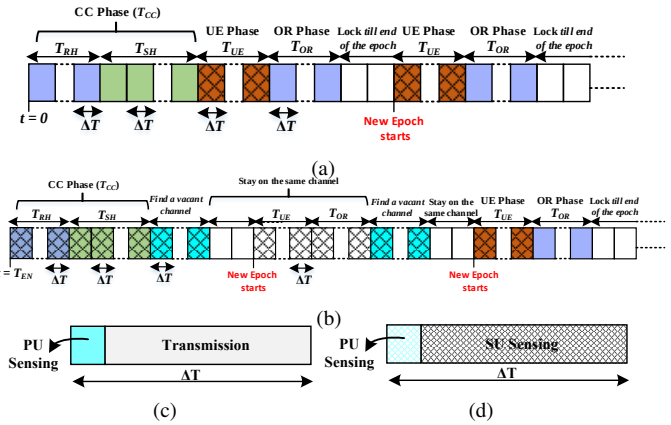


Fig. 2: Phases followed by the SUs entered at (a) $t = 0$, (b) $t = T_{EN}$, (c) Sensing model for SUs entered at $t = 0$ in any phase except UE phase, and (d) Sensing model for new SUs in CC and UE phase.

Lemma 2. *After initial T_{RH} slots, if each SU selects the distinct channel via collision-free sequential hopping for $T_{SH} = \frac{2 \cdot N}{\epsilon^2} \cdot \ln\left(\frac{2 \cdot N^2}{\delta_2}\right)$, then with probability $\geq 1 - \delta_2$ ($\delta_2 \in (0,1)$) all the SUs will have ϵ-correct ($\forall \epsilon > 0$) ranking of channels.*

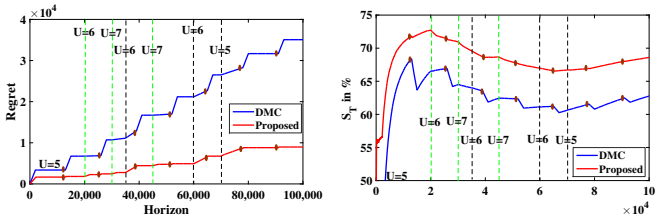
*2) **UE Phase**:* In this phase, the SUs work in either of two modes: - 1) Hopping mode in which SUs performs sequential hopping for the purpose of channel selection. 2) User Sensing (US) mode in which the SU stays on the top channel and performs long sensing (Fig. 2 (d)). A SU can shift to the US mode only once in the whole UE phase. In addition, only one SU can work in the US mode at a time. Once a SU enters into the US mode, it stops sequential hopping after reaching the top channel. Next time slot onwards, SU stays on the top channel and senses the other SUs who are in hopping mode. And thus does the correct estimation of U based on this count.

Lemma 3. *In $T_{UE} \geq N\left(N \cdot \left\lceil \frac{\log \delta_3}{\log(1-\theta)} \right\rceil - 1\right)$ ($Using \mu_1 \geq \theta$) time slots of UE phase, all the SUs will correctly estimate the number of active SUs in the network with probability $\geq 1 - \delta_3$ ($\delta_3 \in (0,1)$).*

*3) **OR Phase**:* In this phase, each SU hops onto one of the top U channels selected uniformly at random in each time slot till it observes a collision free transmission. Once such a channel is found, the SU locks on the same channel till the end of that epoch.

Lemma 4. *In $T_{OR} \geq \frac{\log\left(\frac{\delta_4}{N}\right)}{\log\left(1-\frac{\theta}{4}\right)}$ time slots of OR phase, all the SUs will orthogonalize in one of the top channels with probability $\geq 1 - \delta_4$.*

The performance guarantee of the proposed algorithm is provided under the following additional assumptions : 1) Number of SUs entering and leaving is bounded or at most sub-linear in T, 2) New SU does not enter in the UE and OR phase of an epoch. If number of active SUs changes frequently, no learning may be possible and regret is linear. The first assumption restricts this behaviour. The second assumption ensures that the existing SUs get correct estimation of U. Our theoretical analysis shows that the duration of UE and OR phase is a small



(a) (b)

Fig. 3: Average regret of the proposed and DMC algorithms, (b) Total vacant spectrum utilization of the proposed and DMC algorithms.

portion of the total time horizon T, and thus this assumption can not be assumed over restrictive.

Such assumptions have also been used in state-of-the-art DMC algorithm. In DMC, it is further assumed that new SU can not enter during the learning phase of every epoch. In addition, SU can not leave the network during the learning phase whereas we allow the SUs to leave during the corresponding CC phase due to separate UE phase for U estimation.

IV. Simulations Results and Analysis

Simulation results shown in Fig. 3 (a) and (b) compares the proposed algorithm with the state-of-the-art DMC algorithm [7] in terms of total throughput loss, i.e., average regret and vacant spectrum utilization (S_T) in %. We consider $N = 8$ with channel statistics, $\mu_n = \{0.29, 0.36, 0.43, 0.50, 0.57, 0.64, 0.71, 0.78\}$. T_{CC} for the proposed and T_O for DMC algorithm is considered as 2000. We mark the time of entry and exit of the SU with a green and black dashed line, respectively. Start of a new epoch is indicated with a marker. In Fig. 3, we consider five SUs in the beginning with $T_{EP} = 13000$. After 20000 and 35000 time slots each, a new SU enters the network and at $t = 40000$, one SU leaves the network. The leaving SU is chosen at random from the set of the existing SUs. A new SU enters the network at $T = 45000$ whereas a SU leaves the network at $T = 60000$ and $T = 70000$ each. As expected, the proposed algorithm offers lower regret than DMC due to the use of sequential hopping approach in the CC phase in comparison to the random hopping used in the learning phase of DMC. Also, DMC incurs regret which is linear in time during the learning phase of every epoch whereas we don't repeat this phase in every epoch. Whenever a new SU enters the network, the regret increases during that epoch due to its CC phase. Also, the regret increases during an epoch if SU locked in one of the optimum channels leaves the network. Otherwise, it remains constant after the OR phase of the epoch. This also means that the vacant spectrum utilization of the proposed algorithm is higher than the DMC as evident from Fig. 3 (b).

V. Conclusions and Future Works

Future works include the extension of the proposed algorithm for the scenario where some of the users are not faithful and may deviate from the given algorithm and for the scenario where the channel conditions are different for all SUs.

REFERENCES

[1] Y. Gai, B. Krishnamachari, "Distributed Stochastic Online Learning Policies for Opportunistic Spectrum Access," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, Dec. 2014

[2] M. Zandi, M. Dong and A. Grami, "Distributed Stochastic Learning and Adaptation to Primary Traffic for Dynamic Spectrum Access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1675–1688, Mar. 2016.

[3] L. Besson and E. Kaufmann, "Multi-Player Bandits Models Revisited," *arXiv:1711.02317v1 [stat.ML]*, Nov. 2017.

[4] Rohit Kumar, A. Yadav, S. J. Darak, and M. Hanawal, "Trekking Based Distributed Algorithm for Opportunistic Spectrum Access in Infrastructure-less Network," in *16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2018.

[5] H. Joshi, R. Kumar, A. Yadav and S. J. Darak, "Distributed Algorithm for Dynamic Spectrum Access in Infrastructure-less Cognitive Radio Network, " in *Proc. of IEEE Wireless Communications and Networking Conference(WCNC)*, Barcelona, Spain, 2018.

[6] O. Avner and S. Mannor, "Concurrent Bandit and Cognitive Radio Networks," in *Machine Learning and Knowledge Discovery in Databased*, pp. 66–81, Springer, April 2014.

[7] J. Rosenski, O. Shami and L. Szlak, "Multi-Player Bandits a Musical Chairs Approach," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1–9, New York, USA, 2016.

PUBLICATIONS

- Rohit Kumar, S. J. Darak, Ajay K. Sharma, and Rajiv Tripathi, "**Two-Stage Decision Making Policy for Opportunistic Spectrum Access and Validation on USRP Testbed**", ***Wireless network, Springer***".

- Rohit Kumar, S. J. Darak, Ankit Yadav, Ajay K. Sharma, and Rajiv Tripathi, "**Channel Selection for Secondary Users in Decentralized Network of Unknown Size", *IEEE Communication Letter***.

- Rohit Kumar, A. Yadav, S. J. Darak and M. Hanawal, "**Trekking Based Distributed Algorithm for Opportunistic Spectrum Access in Infrastructure-less Network**," *16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks* (**WiOpt 2018**), Shanghai, China, May 2018.

- S. Sawant, M. Hanawal, S. J. Darak and Rohit Kumar, "Distributed Learning Algorithms for Coordination in a Cognitive Network in Presence of Jammers," *16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks* (**WiOpt 2018**), Shanghai, China, May 2018.

- H. Joshi, Rohit Kumar, Ankit Yadav, and S. J. Darak, "**Distributed Algorithm for Dynamic Spectrum Access in Infrastructure-less Cognitive Radio Network",** *IEEE Wireless Communications and Networking Conference* (**WCNC 2018**), Barcelona, Spain, April 2018.

- Rohit Kumar, S. J. Darak, Ajay K. Sharma, and Rajiv Tripathi, "**Two-Stage Decision Making Policy for Opportunistic Spectrum Access**", *International Conference on Big Data and Advanced Wireless technologies* (**BDAW 2016),** Blagoevgrad, Bulgaria, Nov. 2016.

- Rohit Kumar, "**Distributed Algorithms for Opportunistic Spectrum Access in Decentralized Cognitive Radio Network"**, accepted for Graduate forum in 10th International Conference on Communication Systems and Networks **(COMSNETS),** January 2018, Bangalore, India. (**Secured place in Top 3 paper presenter**).

Fair Subchannel Allocation Algorithms for the Inter Cell Interference Coordination with Fixed Transmit Power Problem

Vaibhav Kumar Gupta, and Gaurav S. Kasbekar
Indian Institute of Technology, Bombay

I. Summary of Our Contributions

We study the Inter Cell Interference Coordination (ICIC) problem in a multi-cell OFDMA based cellular network employing universal frequency reuse. In each cell, only a subset of the available subchannels are allocated to mobile stations (MS) in a given time slot so as to limit the interference to neighboring cells; also, each base station (BS) uses a fixed transmit power on every allocated subchannel. The objective is to allocate the available subchannels in each cell to the MSs in the cell for downlink transmissions taking into account the channel qualities from BSs to MSs as well as traffic requirements of the MSs so as to maximize the weighted sum of throughputs of all the MSs. First, we show that this problem is NP-Complete. Next, we show that when the potential interference levels to each MS on every subchannel are above a threshold (which is a function of the transmit power and the channel gain to the MS from the BS it is associated with), the problem can be optimally solved in polynomial-time via a reduction to the matching problem in bipartite graphs. We also formulate the ICIC problem as a non-cooperative game, with each BS being a player, and prove that although it is an ordinal potential game in two special cases, it is *not* an ordinal potential game in general. Also, we design two heuristic algorithms for the general ICIC problem: a greedy distributed algorithm and a simulated annealing (SA) based algorithm. The distributed algorithm is fast and requires only message exchanges among neighboring BSs. The SA algorithm is centralized and allows a tradeoff between quality of solution and execution time via an appropriate choice of parameters. Our extensive simulations show that the total throughput obtained using the better response (BR) algorithm, which is often used in game theory, is very small compared to those obtained using the SA and greedy algorithms; however, the execution time of the BR algorithm is much smaller than those of the latter two algorithms. Finally, the greedy algorithm outperforms the SA algorithm in dense cellular networks and requires only a small fraction of the number of computations required by the latter algorithm for execution [8].

Next, we study the problem of inter cell interference coordination (ICIC) with fixed transmit power in OFDMA-based cellular networks, in which each base station (BS) needs to decide as to which subchannel, if any, to allocate to each of its associated mobile stations (MS) for data transmission. In general, there exists a trade-off between the total throughput (sum of throughputs of all the MSs) and fairness under the allocations found by resource allocation schemes. We introduce the concept of $\tau - \alpha-$fairness by modifying the concept of $\alpha-$fairness, which was earlier proposed in the context of designing fair end-to-end window-based congestion control protocols for packet-switched networks. The concept of $\tau - \alpha-$fairness allows us to achieve arbitrary trade-offs between the total throughput and degree of fairness by selecting an appropriate value of α in $[0, \infty)$. We show that for every $\alpha \in [0, \infty)$ and every $\tau > 0$, the problem of finding a $\tau - \alpha-$fair allocation is NP-Complete. Also, we propose a simple, distributed subchannel allocation algorithm for the ICIC problem, which is flexible, requires a small amount of time to operate, and requires information exchange among only neighboring BSs. We investigate via simulations as to how the algorithm parameters should be selected so as to achieve any desired trade-off between the total throughput and fairness [9].

We use *Jain's fairness index*, which was proposed in [11] and has been extensively used in the networking literature, *e.g.,* in [2], [6], [17], as a fairness metric. For different values of α we found the allocation that maximizes the system utility function by exhaustive search over all possible combinations of subchannel allocation to all the MSs of the system. The total throughput decreases and fairness index (FI) increases with α. We found that the values of total throughput and fairness index FI are not very sensitive to the value of τ. For the distributed $\tau - \alpha-$fair subchannel allocation algorithm, we first find the value of the parameter p_0 (for termination of the algorithm), say p_0^*, that results in the maximum total throughput under the allocation found by the algorithm. The total throughput is maximized for medium values of p_0. Therefore, the total throughput first increases then decreases and fairness index increases with as p_0 increases. We empirically found that the value of the parameter p_0 (say p_0^*) which gives close to maximum total throughput in terms of the parameters K, M

Authors email addresses are vaibhavgupta@ee.iitb.ac.in, and gskasbekar@ee.iitb.ac.in respectively.

and N is given by the following expression:

$$p_0^* = \begin{cases} 1 + \frac{M}{2(NK)}, & \text{if } M \leq K \times N, \\ 1 + \frac{\log(NK)}{2 \log M}, & \text{otherwise.} \end{cases} \quad (1)$$

Further, within the range $p_0 \in [0, p_0^*]$, the total throughput (respectively, fairness) is maximized at $p_0 = p_0^*$ (respectively, $p_0 = 0$). However, recall that $\alpha = 0$ (respectively, $\alpha = \infty$) corresponds to maximum total throughput (respectively, fairness) and minimum fairness (respectively, total throughput). This motivates us to set p_0, in terms of α, as:

$$p_0 = \frac{1}{\frac{1}{p_0^*} + \alpha}. \quad (2)$$

In summary, the choice of p_0 in (2) ensures that as α increases from 0 to ∞, the total throughput (respectively, degree of fairness) of the allocation found using the algorithm decreases (respectively, increases).

II. Related Work

The ICIC with variable transmit power problem is considered in [15], [18], [19]. The ICIC with fixed transmit power problem was studied in [5], [13], [16]. However, in all of the above papers [5], [13], [16], some or all of the allocation decisions are taken at a central controller that is connected to all the base stations; our greedy and BR algorithms do not require the availability of such a controller. No proof was provided to show that this problem is NP-hard in the literature. In contrast, a rigorous proof of the fact that the ICIC with fixed transmit power problem is NP-complete is provided in our work. Recently, the problem of allocation of subchannels to mitigate inter-cell interference has been extensively studied using the theory of *potential games*. The game was shown to be an exact potential game. Frequency selective channel gains were not considered in [1], [4], [7], [21]. In contrast, the model in our work takes into account frequency selective channel gains. Also, the utility function of a player (BS) in the channel allocation game was considered to be the total time delay experienced by the BS to transmit the data in [1], the negative of the total interference received at the BS in [4], [10], [20], [22] and MOS in [21]. In contrast, we have considered the total throughput of all the MSs associated with a BS to be its utility function. We show that the game in this work is not a potential game in general, which is surprising, since the games in [1], [4], [7], [10], [22], [21], [20], which are similar, are potential games.

Most of the proposed resource allocation schemes to address the ICIC problem consider maximizing the total throughput, *i.e.,* the sum of throughputs of all the MSs in the system, while completely neglecting the aspect of fairness [3], [8], [12], [14], [16], [19]. In the context of cellular systems, fairness means that each MS, irrespective of its channel gain (which is a measure of the quality of the channel from the BS to the MS), has an equal chance of being allocated each of the available subchannels, *i.e.,* no MS is preferred over the other MSs while allocating a subchannel in the system. Maximization of the total throughput results in high throughput of the MSs with good channel gain values; however, this is at the cost of low throughput of the MSs with poor channel gain values such as MSs at the cell boundaries [6]. On the other hand, if lower (respectively, higher) throughputs were assigned to MSs with good (respectively, poor) channel gains, then it would lead to better fairness, but at the expense of the total throughput. So, there exists a trade-off between the total throughput and fairness of resource allocation schemes [2]. Motivated by this fact, our objective in this work is to formulate the problem of achieving different trade-offs between the total throughput and fairness, study its complexity and design a distributed resource allocation algorithm to solve it.

Most of the proposed subchannel allocation algorithms in the literature provide a particular level of fairness *e.g.* either proportional ($\alpha = 1$) or max-min ($\alpha = \infty$) fairness. However, no results were provided to study the trade-off between the total throughput and fairness achieved by the proposed resource allocation scheme. In contrast, in this work, we provide a resource allocation scheme that can be used to achieve arbitrary trade-offs between the total throughput and level of fairness. Note that we consider all the values of α in $[0, \infty)$, which correspond to different trade-offs between the total throughput and fairness. To the best of our knowledge, *our work is the first to formulate the ICIC with fixed transmit power problem with the goal of achieving arbitrary trade-offs between the total throughput and fairness*.

III. Publications

The work summarised in this abstract has been published in [8] and [9].

References

[1] A. Adouane, L. Rodier, K. Khawam, J. Cohen, S. Tohme, "Game Theoretic Framework for Inter-Cell Interference Coordination", *Proc. of IEEE WCNC*, pp. 57-62, 2014.

[2] A. Bin Sediq, R. H. Gohary, H. Yanikomeroglu, "Optimal tradeoff between efficiency and Jains fairness index in resource allocation", *Proc. 2012 IEEE PIMRC*, pp. 577583, Nov. 2012.

[3] A. Bin Sediq, R. Schoenen, H. Yanikomeroglu, G. Senarath, "Optimized Distributed Inter-Cell Interference Coordination (ICIC) Scheme Using Projected Subgradient and Network Flow Optimization", *IEEE Transactions on Communications*, Vol. 63, No. 1, pp. 107-124, 2015.

[4] X. Chen, J. Huang, "Database-Assisted Distributed Spectrum Sharing", *IEEE Journal on Selected Areas in Communications*, Vol. 31, No. 11, pp. 2349-2361, Nov. 2013.

[5] R.Y. Chang, Z. Tao, J. Zhang, C.-C.J. Kuo, "Multicell OFDMA Downlink Resource Allocation Using a Graphic Framework", *IEEE Transactions on Vehicular Technology*, Vol. 58, No. 7, pp. 3494-3507, 2009.

[6] H. T. Cheng and W. Zhuang, "An optimization framework for balancing throughput and fairness in wireless networks with QoS support", *IEEE Trans. of Wireless Commun.*, Vol. 7, No. 2, pp. 584-593, Feb. 2008.

[7] J. Ellenbeck, C. Hartmann, L. Berlemann, "Decentralized Inter-Cell Interference Coordination by Autonomous Spectral Reuse Decisions", *Proc. of European Wireless Conference*, 2008.

[8] V. K. Gupta, A. Nambiar and G. S. Kasbekar, "Complexity Analysis, Potential Game Characterization and Algorithms for the Inter Cell Interference Coordination with Fixed Transmit Power Problem", *IEEE Transactions on Vehicular Technology*, Vol. 67, No. 4, pp. 3054 - 3068, Nov. 2017.

[9] V. K. Gupta, and G. S. Kasbekar, "Achieving Arbitrary Throughput-Fairness Trade-offs in the Inter Cell Interference Coordination with Fixed Transmit Power Problem", *Proc. of NETGCOOP*, New York, Nov. 2018.

[10] G. Huang, and J. Li, "Interference mitigation for femtocell networks via adaptive frequency reuse", *IEEE Transactions on Vehicular Technology*, vol. 65, No. 4, pp. 2413-2423, 2016.

[11] R. Jain, D. Chiu, and W. Hawe "A quantitative measure of fairness and discrimination for resource allocation in shared systems", *Digital Equipment Corporation, Tech. Rep. DEC-TR-301*, Sep. 1984.

[12] C. Kosta, B. Hunt, A.U. Quddus, R. Tafazolli, "A Low-Complexity Distributed Inter-Cell Interference Coordination (ICIC) Scheme for Emerging Multi-Cell HetNets", *Proc. of IEEE VTC*, 2012.

[13] G. Li, H. Liu, "Downlink Radio Resource Allocation for Multi-Cell OFDMA System", *IEEE Transactions on Wireless Communications*, Vol. 5, No. 12, pp. 3451-3459, 2006.

[14] D. Lopez-Perez, I. Guvenc, G. De la Roche, M. Kountouris, T.Q.S. Quek, J. Zhang, "Enhanced Intercell Interference Coordination Challenges in Heterogeneous Networks", *IEEE Wireless Communications*, Vol. 18, No. 3, pp. 22-30, 2011.

[15] M. Moretti, A. Todini, A. Baiocchi, G. Dainelli, "A Layered Architecture for Fair Resource Allocation in Multicellular Multicarrier Systems", *IEEE Transactions on Vehicular Technology*, Vol. 60, No. 4, pp. 1788-1798, 2011.

[16] M. Rahman, H. Yanikomeroglu, "Enhancing Cell-edge Performance: A Downlink Dynamic Interference Avoidance Scheme with Inter-cell Coordination", *IEEE Transactions on Wireless Communications*, Vol. 9, No. 4, pp. 1414-1425, 2010.

[17] S. Sheikh, R. Wolhuter, and H. A. Engelbrecht, "An Adaptive Congestion Control and Fairness Scheduling Strategy for Wireless Mesh Networks", *Proc. IEEE SSCI*, pp. 1174-1181, Dec. 2015.

[18] A.L. Stolyar, H. Viswanathan, "Self-Organizing Dynamic Fractional Frequency Reuse in OFDMA Systems", *Proc. of INFOCOM*, 2008.

[19] M. Yassin, "Inter-Cell Interference Coordination in Wireless Networks", Thesis May 2016 [Online]. Available: *https://tel.archives-ouvertes.fr/tel-01316543*.

[20] J. Zheng, Y. Cai, A. Anpalagan, "A Stochastic Game-Theoretic Approach for Interference Mitigation in Small Cell Networks", *IEEE Communications Letters*, Vol. 19, No. 2, pp. 251-254, 2015.

[21] J. Zheng, Y. Cai, N. Lu, Y. Xu, B. Duan, X. Shen, "Optimal Power Allocation and User Scheduling in Multicell Networks: Base Station Cooperation Using a Game-Theoretic Approach", *IEEE Transactions on Wireless Communications*, vol. 13, No. 12, pp. 6928-6942, 2014.

[22] J. Zheng, Y. Cai, N. Lu, Y. Xu, X. Shen, "Stochastic Game-Theoretic Spectrum Access in Distributed and Dynamic Environment", *IEEE Transactions on Vehicular Technology*, vol. 64, No. 10, pp. 4807-4820, 2015.