Journal of Information & Communications

ISSN 1859 - 3534

MINISTRY OF INFORMATION AND COMMUNICATIONS

Volume 2019
Number 2
December

Research and Development on Information & Communication Technology

Các công trình nghiên cứu phát triển Công nghệ Thông tin và Truyền thông

Journal of Information & Communications
MINISTRY OF INFORMATION AND COMMUNICATIONS
ISSN: 1859-3534
VOLUME 2019, NUMBER 2, DECEMBER

JOURNAL ADDRESS:

115 Trần Duy Hưng street, Hanoi, Vietnam
Tel: (84)-24-37737136
Fax: (84)-24-37737130, Website http://ictmag.vn
Email: ict.research@mic.gov.vn (editorial) chuyensanbcvt@mic.gov.vn (administrative)
Journal Office Branch
27 Nguyễn Bỉnh Khiêm Street, District 1, Hồ Chí Minh City, Vietnam
Tel/Fax: (84)-28-38992898

JOURNAL MARKETING AND DISTRIBUTION

Phạm Thị Lâm
Email: ptlam@mic.gov.vn, Tel: 84-948503999

License No 365/GP-BTTTT, dated 19 December 2014; Modified license No 233/GP-BTTTT dated 23 May 2017.
Printed by Hanoi Printing Stock Company.
Printed and delivered in Dec 2019.

Price: 40,000 VND

Research and Development on
Information and Communication Technology
http://www.ictmag.vn/ict
Volume 2019, Number 2
December

TABLE OF CONTENTS

SELECTED ARTICLES

REGULAR ARTICLES

2019 LIST OF REVIEWERS

Hồ Phạm Huy Anh	Ho Chi Minh City University of Technology
Lê Hải Châu	Posts and Telecommunications Insitute of Technology
Hoàng Quốc Chính	Vinmec Healthcare System
Nilanjan Dey	Techno India College of Technology, India
Lê Vũ Hà	University of Engineering and Technology, Vietnam National University, Hanoi
Lê Đức Hậu	Vingroup Big Data Institute
Nguyễn Trung Hiếu	Posts and Telecommunications Insitute of Technology
Trần Hùng	Malardalen University, Sweden
Nguyễn Lê Hùng	The University of Danang
Nguyễn Tấn Hưng	University of Technology, The University of Danang
Trần Đăng Hưng	Hanoi University of Education
Phạm Thị Thảo Khương	University of Technology and Education, The University of Danang
Trương Trung Kiên	Posts and Telecommunications Insitute of Technology
Nattapong Kitsuwan	University of Electro-Communications, Japan
Raghvendra Kumar	LNCT Group of Colleges, India
Võ Sỹ Nam	VinTech City
Võ Duy Phúc	University of Technology, The University of Danang
Huỳnh Nguyễn Bảo Phương	Quy Nhon University
Trần Minh Quang	Ho Chi Minh City University of Technology
Nguyễn Quang Như Quỳnh	University of Technology, The University of Danang
Lê Hoàng Sơn	Information Technology Institute, Vietnam National University, Hanoi
Lê Trung Thành	University of Engineering and Technology, Vietnam National University, Hanoi
Lương Văn Thiện	University of Southampton, UK
Nguyễn Hồng Thịnh	University of Engineering and Technology, Vietnam National University, Hanoi
Ngô Thị Minh Thùy	Oregon Health & Science University, US
Chu Đình Tới	Hanoi University of Education
Lê Cung Tường	Ton Duc Thang University
Nguyễn Linh Trung	University of Engineering and Technology, Vietnam National University, Hanoi
Sudeep Tanwar	Nirma University, India
Nguyễn Duy Nhật Viễn	University of Technology, The University of Danang
Trịnh Anh Vũ	University of Engineering and Technology, Vietnam National University, Hanoi

Introduction to the Special Issue on Bioinformatics and Computational Biology

Le Duc Hau[1], Le Hoang Son[2]

[1] Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam

[2] Information Technology Institute, Vietnam National University, Hanoi, Vietnam

Correspondence: Le Duc Hau, v.hauld1@vintech.net.vn

Communication: received 31 December 2019, revised 31 December 2019, accepted 31 December 2019

Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.917

The Editor coordinating the review of this article and deciding to accept it was Prof. Nguyen Linh Trung

Abstract: The Special Issue on Bioinformatics and Computational Biology in the Journal of Research and Development on Information and Communication Technology (ICT Research) aims at bringing together researchers in Vietnam for exchange of new developments in all areas of bioinformatics and computational biology.

Keywords: *Bioinformatics, computational biology.*

In the last ten years, since the Human Genome Project [1] and the development of the next-generation sequencing technologies, the cost for whole-genome sequencing has reduced substantially. A number of large-scale human genome projects have been launched to characterize genetic variation at both global and specific population levels, finally created large-scale genetic databases. The produced databases from these massive sequencing projects have been used to implement association mapping studies of complex disorders and other phenotypic traits. The availability of genetic databases has facilitated the integration of genetic factors into risk prediction models. A polygenic risk score (PRS) combines the effect of many single nucleotide polymorphisms (SNP) into a single score, and has lately been shown to have a clinically predictive value in various common diseases.

In this special issue, the first study of Nguyen Tran The Hung and Le Duc Hau [2] aims to articulate the evidence supporting the clinical usage of PRS, the concepts behind the validity of PRS, approaches to implement PRS in Vietnamese population and cautions in selecting methods and thresholds in developing an appropriate PRS. In addition, the genetic databases have paved the way for the study of evolutionary relationships among populations by phylogenetic trees. Computational methods for building phylogenetic trees from gene/protein sequences have been developed for decades and come of age.

The next study of Le Sy Vinh [3] analyses widely-used methods to construct large phylogenetic trees, and available methods to build phylogenomic trees from whole-genome datasets. It also makes a recommendation for best practices when performing phylogenetic and phylogenomic analyses.

Finally, the next-generation sequencing technologies have also facilitated the study of liquid biopsies for tracking the genomic evolution of tumors over time with application in cancer early detection by analyzing cell-free DNA. The liquid biopsy technology helps avoid the need to conduct repeated biopsies, thus reduce the risk for patients. The last study of Nguyen Ngoc Tran [4] summarizes major advances in liquid biopsy assay technologies and discusses the types of cancers that most likely benefit from early detection.

Through these researches, a picture of advanced applications of next-generation sequencing technologies especially in Vietnam has been depicted and summarized.

References

[1] J. C. Venter, M. D. Adams *et al.*, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[2] N. T. T. Hung and L. D. Hau, "Development and implementation of polygenic risk score in vietnamese population," *Research and Development on Information and Communication Technology*, vol. 2019, no. 2, pp. 75–83, 2019.

[3] L. S. Vinh, "Phylogenetic and phylogenomic analyses for large datasets," *Research and Development on Information and Communication Technology*, vol. 2019, no. 2, pp. 84–92, 2019.

[4] N. N. Tran, "Liquid biopsy to characterize cell-free DNA in cancer detecting and monitoring," *Research and Development on Information and Communication Technology*, vol. 2019, no. 2, pp. 93–98, 2019.

Le Duc Hau obtained his PhD degree in Bioinformatics from the University of Ulsan, Republic of Korea in 2012. He is now leading the Department of Computational Biomedicine at Vingroup Big Data Institute, Vietnam. He has been focusing on proposing computational methods for disease- and drug-related problems in personalized medicine, especially on identification of disease-associated biomarkers, prediction of drug targets and response. In parallel, he has been developing bioinformatics tools. So far, he has published more than fifty papers in well-recognized journals and conferences, nearly a half of those are in ISI-indexed journals. In addition, he has been a member of program committees and reviewer of several international conferences/journals. Moreover, he is a principal investigator and a key member of some national/ministry-level projects. Specially, he is the principal investigator of the biggest genome project in Vietnam (i.e., building databases of genomic variants for Vietnamese population). Finally, he has been collaborating with some well-recognized international research institutes.

Le Hoang Son obtained the Ph.D. degree in mathematics and informatics from the University of Science of Vietnam National University, Hanoi (VNU), in conjunction with the Politecnico di Milano University, Italy, in 2013. From 2007 to 2018, he was a Senior Researcher and Vice-Director of the Center for High Performance Computing at VNU University of Science. Since, he has been an Associate Professor of information technology in Vietnam. Since August 2018, he has been a Senior Researcher of the Department of Multimedia and Virtual Reality at the Information Technology Institute of VNU. His research interests include artificial intelligence, data mining, soft computing, fuzzy computing, fuzzy recommender systems, and geographic information systems. He is a member of Vietnam Journalists Association, the International Association of Computer Science and Information Technology, Vietnam Society for Applications of Mathematics, the Key Laboratory of Geotechnical Engineering, and Artificial Intelligence at the University of Transport Technology. He is an Associate Editor of the Journal of Intelligent and Fuzzy Systems (SCIE), IEEE ACCESS (SCIE), Data Technologies and Applications (SCIE), International Journal of Data Warehousing and Mining (SCIE), Neutrosophic Sets and Systems (ESCI), Vietnam Research and Development on Information and Communication Technology, VNU Journal of Science: Computer Science and Communication Engineering, and Frontiers in Artificial Intelligence. He serves as an Editorial Board for Applied Soft Computing (SCIE), Plos One (SCIE), International Journal of Web and Grid Services (SCIE), International Journal of Ambient Computing and Intelligence (ESCI), and Vietnam Journal of Computer Science and Cybernetics. He is a Guest Editor of several Special Issues at International Journal of Uncertainty, Fuzziness and Knowledge Based Systems (SCIE) and Journal of Ambient Intelligence and Humanized Computing.

Development and Implementation of Polygenic Risk Score in Vietnamese Population

Nguyen Tran The Hung, Le Duc Hau

Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam

Correspondence: Le Duc Hau, v.hauld1@vintech.net.vn

Communication: received 18 October 2019, revised 23 December 2019, accepted 29 December 2019

Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.893

The Editor coordinating the review of this article and deciding to accept it was Prof. Le Hoang Son

Abstract: Recent technological advancements and availability of genetic databases have facilitated the integration of genetic factors into risk prediction models. A Polygenic Risk Score (PRS) combines the effect of many Single Nucleotide Polymorphisms (SNP) into a single score. This score has lately been shown to have a clinically predictive value in various common diseases. Some clinical interpretations of PRS are summarized in this review for coronary artery disease, breast cancer, prostate cancer, diabetes mellitus, and Alzheimer's disease. While these findings gave support to the implementation of PRS in clinical settings, the populations of interest were derived mainly from European ancestry. Therefore, applying these findings to non-European ancestry (Vietnamese in this context) requires many efforts and cautions. This review aims to articulate the evidence supporting the clinical use of PRS, the concepts behind the validity of PRS, approach to implement PRS in Vietnamese population, and cautions in selecting methods and thresholds to develop an appropriate PRS.

Keywords: *Genetic, clinical, single nucleotide polymoyphism (SNP), polygenic risk score (PRS).*

I. Renewed Interest in Polygenic Risk Score

1. Definition

A Polygenic Risk Score (PRS) in the context of genetic studies is a mathematical aggregation of risk effects conferred by many Single Nucleotide Polymorphisms (SNP). Each SNP contributes a small effect to the development of a disease or a complex trait of interest. In the early days of Genome-Wide Association Study (GWAS), researchers expected to find genetic variants that have a large effect on disease risk [1]. While the sample size of GWASs has already surpassed hundreds of thousand of individuals, they failed to capture the genetic variants that can explain the heritability of common diseases, such as breast cancer, prostate cancer, coronary artery disease, diabetes mellitus, Alzheimer disease [2]. Therefore, there is a growing interest in combining all the small SNP effects into a single score that has significant and applicable values [3, 4].

2. PRS Calculation

In its simplest form, the PRS of an individual can be calculated as the sum of all effect sizes of the effective alleles observed in its genotype. The formula to calculate the PRS is given as follows:

$$\mathrm{PRS}_i = \sum_{j=1}^{m} x_{ij} \times \widehat{B}_j,$$

where PRS_i is the risk score for the i^{th} individual, m is the number of SNPs included in the calculation, x_{ij} is the genotype of the i^{th} individual for the j^{th} SNP (can be 0, 1, or 2 depending on the inheritance model), and $\widehat{B}_j$ is the effect size of the j^{th} SNP, usually obtained from GWAS summary statistics [4].

Although the concept of PRS is as old as the finding of genetic materials (DNA), modern technology allows the integration of more genetic variants and more precise effect sizes. Therefore, there are numerous considerations and thresholds related to developing and validating the formula of PRS that are still controversial [5].

3. Advancements in the Field of Genetics

At this junction, there are many developments of technology and findings of new studies that facilitate the development of PRS. The availability of various populations' reference human genomes can be accessed publicly in the 1000 genomes project [6]. Thousands of GWASs comprise of up to millions of samples. These GWAS summary statistics data can be easily accessed through the

"GWAS Catalog" [7], which is an online database with more than 4000 published studies.

New analysis methods for developing the PRS without relying solely on genome-wide significant hits continue to appear, such as Clumping + Thresholding [8], Penalized Regression [9]. The access to large genotype and phenotype data of large longitudinal cohorts becomes easier through online databases such as "dbGAP" [10] and "UK biobank" [11].

II. Clinical Usage of PRS

While making clinical decisions, doctors often have to classify the susceptibility of a patient based on known risk factors. This disease risk classification is very important in providing an appropriate recommendation for the patient. A group of individuals having certain risk factors could have higher relative risk than the general population to guarantee different clinical management. If existing medical intervention can provide more benefits than adverse effects, with reasonable costs, this group of high-risk individuals would receive more benefits from it. Recent studies have suggested that genetic profiling using the PRS can provide some clinical utilities [12].

PRS analysis and its interpretation revolved around some situations: risk prediction performance of PRS independently or in combination with other non-genetic risk factors and estimation of lifetime risk trajectories. Some recent studies have proposed some clinical interpretations of PRS that can modify therapeutic intervention, disease screening and life planning [13]. This review highlights some recent findings of PRS in certain common diseases: coronary artery disease, diabetes mellitus, breast cancer, prostate cancer and Alzheimer's disease.

1. Coronary Artery Disease

Clinical risk scores like the Framingham risk score is a traditional tool in evaluating 10-year coronary artery disease (CAD) risk [14]. This score uses clinical risk factors to score each individual and infer his chance of developing CAD in the next 10 years. Abraham *et al.* have proved that integrating the PRS in traditional clinical risk model can better capture the lifetime risk of CAD in patients [15]. This argument was supported by better C-statistic (measure of goodness-of-fit) in the combined model as compared to the clinical one. More importantly, men in the top quintile of PRS had 10% cumulative CAD risk around 15 years earlier than men in the bottom quintile.

In the primary prevention setting, statin can be used to treat atherosclerosis and reduce the risk of cardiovascular events [16]. The PRS can identify a group of patients having high genetic risk for CAD, who can receive more benefits from primary prevention with statin therapy [17]. Patients having the top quintile of PRS have higher risk of subclinical atherosclerosis and receive greater absolute risk reduction of CAD event from statin therapy (Figure 1).

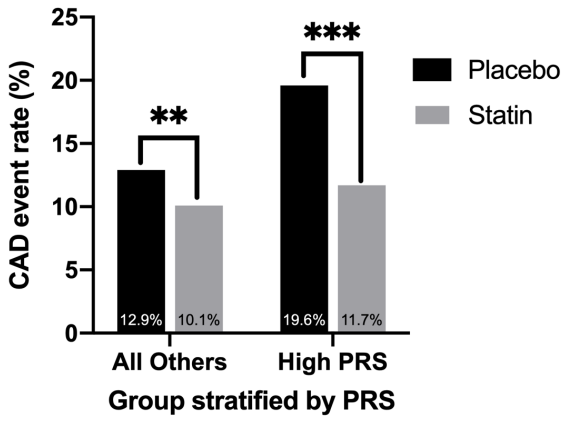


Figure 1. CAD incident by PRS group in statin primary prevention trial. Adapted from a study by Natarajan et al. (2017) [17]. **: chi-square test with p-value < 0.01. ***: chi-square test with p-value < 0.001.

2. Breast Cancer

Breast cancer screening has been recommended for women older than 50 without major risk factors for a long time [18]. The reasoning behind screening for breast cancer in women older than 50 is reducing disease mortality and decreasing false positive diagnosis. A study by Pashayan *et al.* argued that a well-defined risk-stratified screening strategy would improve the quality of life of women and save resources [19]. Based on this risk/benefit threshold, a risk prediction model combining clinical risk factors and the PRS could identify a subgroup of women who had relative risk of developing breast cancer higher than that of 50-year-old women [20]. These high-risk individuals could benefit from earlier screening tests and assertive lifestyle change to reduce certain risk factors. The women in the top quintile of PRS could have the same relative risk as average 50-year-old women around 5-10 years earlier (Figure 2).

3. Prostate Cancer

Current medical guideline suggests that the age to consider prostate cancer screening is 50 years for average-risk men as long as life expectancy is at least 10 years [22]. A study by Seibert *et al.* [23] argued that the PRS was a significant predictor for the age of prostate cancer onset. It was also a relatively inexpensive evaluation of individual's benefits from prostate cancer screening.

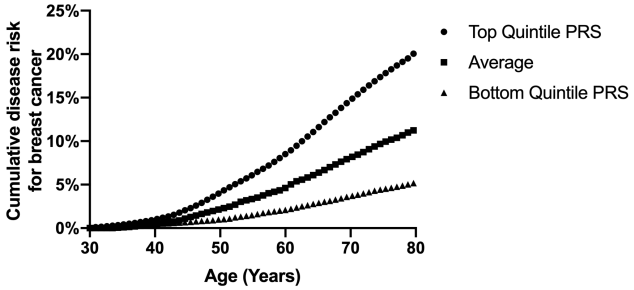


Figure 2. Cumulative breast cancer risk stratified by PRS quintile. Adapted from a study by Maas et al. (2016) [21].

4. Diabetes Mellitus

Early detection of individuals with high risk of type 1 diabetes (T1D) allows better monitoring and prevention of disease progression. Redondo *et al.* evaluated the performance of PRS on T1D patients' relatives without diabetes and with one or more positive autoantibodies [24]. Progression to T1D was best predicted by a combined model of PRS, number of positive auto-antibodies, DPT-1 Risk Score [25], and age. Individuals at high risk of developing T1D can benefit from monitoring and prevention trials.

A prospective cohort study by Lall *et al.* [26] used a PRS that had the strongest association with type 2 diabetes (T2D) in a population-based cohort and evaluated its performance on a prospective individual risk assessment. The hazard for incident T2D was more than 3 times higher in the top quintile of PRS, as compared to others.

5. Alzheimer's Disease

One of the most common causes of dementia is Alzheimer's disease. Desikan *et al.* studied the PRS performance in stratifying Alzheimer's disease risk [27]. This study argued that the PRS could be integrated into screening for individuals with age-specific high genetic risk for Alzheimer's disease. This finding has not been found its use in clinical settings yet, but may prove to be useful for therapeutic trials.

III. Validity of PRS

1. Construct

When the PRS was constructed, it was assumed that SNPs had additive effects on the disease. Because of the large number of SNPs and its unexplained characteristics with the disease of interest, GWAS typically chose the additive model for statistical analysis [28]. However, the biological reality is assuredly more complicated than that. The mode of inheritance of a SNP could be additive, multiplicative, recessive or dominant [29]. Performing only additive model tests could avoid multiple comparisons but overlooked the other inheritance models. Besides, when the gene-gene interaction and gene-environment exposure were taken into account, the model became much more complicated and current statistical methods could not keep up with this complexity [30, 31].

2. Content

In the context of implementation, the PRS was used to predict the genetic disease risk of common diseases. The content of the PRS needed to capture all of the genetic variations with the purpose of reflecting the genetic liability of the disease. However, for many common diseases, genetic variation only accounted for a small portion of the disease phenotype [2].

The diagram in Figure 3 illustrates the contributing factors to T2D development and the way some T2D-risk prediction model were constructed. Conceptually, the SNPs having effect on a complex trait such as T2D consist of SNPs that modified intermediate phenotypes (blood pressure, BMI), which eventually contribute to the risk of T2D. These intermediate phenotypes present themselves as clinical risk factors. T2D is also affected by factors independent with genetics such as age, lifestyle (Figure 3).

The conventional risk prediction model used for T2D in clinical settings only included clinical risk factors [32]. Recent findings in the field of genetics suggested that combining clinical risk factors and the PRS could improve the current risk prediction model and the cost-benefit metrics [33]. The caveat of this approach was that many clinical risk factors are not independent with genetics. The combined prediction model might have the effect of genetic factors and clinical factors of the same pathogenic mechanism counted simultaneously (*e.g.*, the effect of SNPs associated with blood pressure and the effect of clinical high blood pressure were both included in the prediction model in Figure 3). Although the combined model showed improved C-statistic, a single mechanism (blood pressure/BMI) was counted twice: one in the genetic feature and the other in the clinical one. The outcome would be biased toward that mechanism. Another approach was to evaluate the prediction model only with the PRS, stratified by age [26]. This approach highlighted the independent nature of the PRS and visualized the cumulative risk of the disease of the high-risk group compared to the general population.

3. Criterion

Whether the PRS has valid predictive power depends on the specific disease and population. A review by Duncan *et al.* [34] found out that the majority of PRS studies included

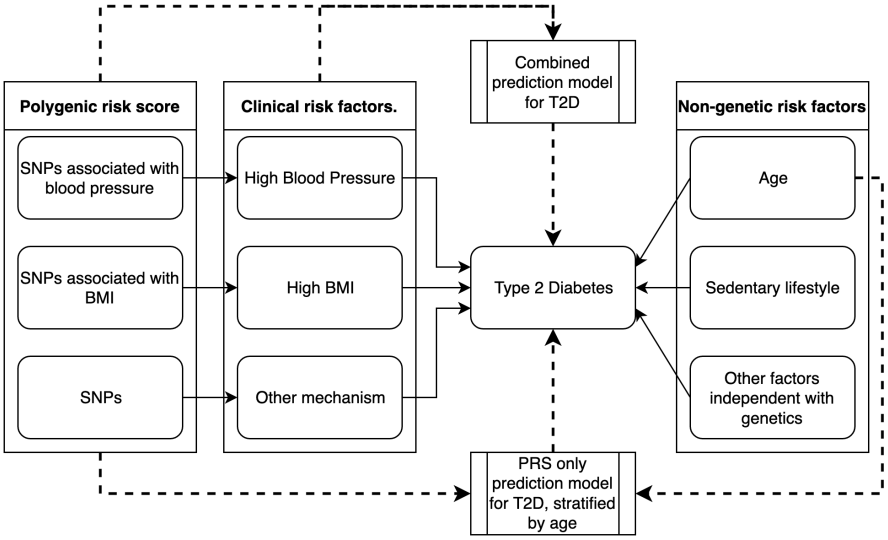


Figure 3. Risk factors of T2D and content of the prediction model for T2D.

European ancestry and East Asian ancestry participants. A PRS derived from the European population had lower performance in the non-European population. As a result, if we wanted to apply PRS utilities in the Vietnamese population, we had to improve methodological choice and threshold to accommodate the difference in linkage disequilibrium and variant frequency between Vietnamese and European/East-Asian. The clinical utility of the PRS needed to be validated in a prospective cohort study [35].

IV. Implementing PRS in Vietnamese Population

1. Method to Read DNA

Genotyping is a method of determining which genetic variants an individual possesses. SNP microarrays allow detection of hundreds of thousands of pre-determined SNPs. In genetic research, SNP arrays are most frequently used for GWASs. A commercial genotyping array included around 500, 000 SNPs and cost around 100 USD [37].

Sequencing is the process of determining the DNA sequence, which is the exact order of DNA's bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Whole-genome sequencing and whole-exome sequencing are more expensive but they allow a more precise detection of genetic variations (*e.g.*, structural variants). Depending on the region, a given stretch of sequence may include some DNA that varies between individuals, in addition to the constant region. Thus, sequencing can be used to determine the genotype of an individual for known variants, as well as identify variants that may be unique to that person [38].

The cost of genotyping array is lower and, thus, more suitable for large scale research in the population. However, the commercial genotyping array used in GWAS has a strong ascertainment bias because SNPs are chosen from European individuals [39]. The best way to overcome this problem is to design an SNP-array suitable for the Vietnamese population.

2. Available Databases

The 1000 genome project, which shared reference genome from diverse populations, contained over 88 million variants from 2504 individuals [6]. It provided a broad representation of human genomes in different populations and ethnicities. This database contained 101 Vietnamese individuals (Kinh ethnic from Ho Chi Minh city). The availability of the Vietnamese reference genome is very important for researchers who are interested in conducting genetic research in Vietnam. The more references there are, the better it can represent Vietnamese human genomes. As a result, risk prediction based on genetic factors will be more reliable and accurate. To this end, a Vietnamese genetic database is being built [40]. As data grow larger, the prospect of implementing genetic findings in Vietnamese clinical settings become more imminent.

Of course, this is only the first step in genetic research in Vietnam. In order to catch up with international research and development, Vietnam still has a long way to go. There are some established databases of human genotype-phenotype like dbGaP [41] and PheGenI [42]. These databases allow researchers to share their datasets and to perform large scale and complex analysis on various diseases. A genotype-phenotype database of the Vietnamese

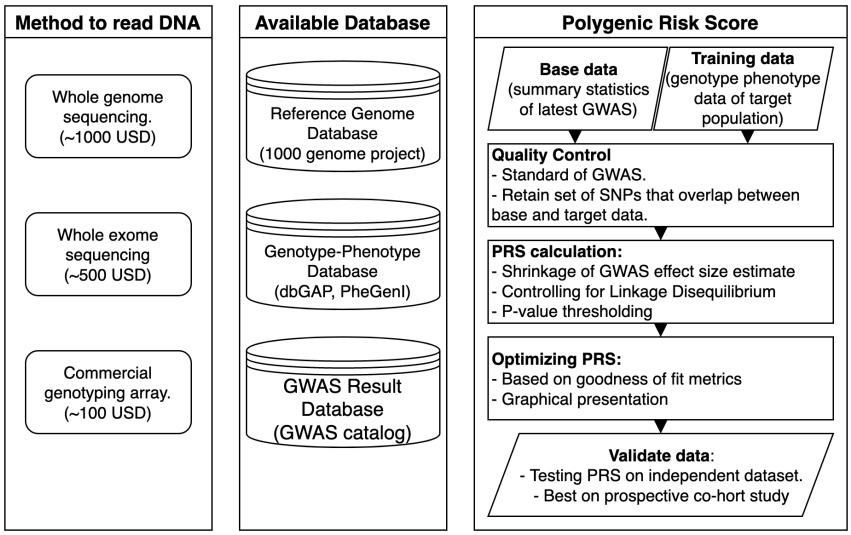


Figure 4. Resources of genetic studies and PRS analysis. Adapted from a tutorial by Choi et al. (2018) [36].

population for replicating published genetic findings is sorely needed. The availability of such a database will facilitate the replication of PRS research with high applicability and will optimize the predictive performance in the Vietnamese population. Besides, replication of GWAS with a large Vietnamese sample size will also provide valuable information for genetic research.

V. Polygenic Risk Score Analysis

Polygenic risk score calculation can be characterized by two input data sets: the base data (the summary statistics of the latest GWAS concerning the disease of interest) and the training data (the individual genotype-phenotype from the population of interest) as illustrated in Figure 4.

1. Quality Control of Input Data

Since the base data come from GWAS, input data must be quality-controlled (QC) according to the standard of GWAS. According to a tutorial on conducting GWAS by Marees *et al.* [28], QC steps for a successful GWAS are:

1) Exclude SNPs that are missing in greater than 2% of the subjects;
2) Exclude individuals with genotyping rate greater than 2%;
3) Exclude individuals with sex discrepancies between recorded data and their X chromosome's heterozygosity;
4) Include SNPs with minor allele frequency (MAF) above the threshold (based on the sample size of the study, larger sample size can use lower threshold);
5) Exclude SNPs that deviate from Hardy-Weinberg equilibrium (for binary traits HWE p-value is less than 10^{-10}, for quantitative traits HWE p-value is less than 10^{-6});
6) Exclude individuals with heterozygosity rate deviated more than 3 standard deviations from the population mean;
7) Perform principal component analysis on the training data and using the first 10 eigenvectors as covariates.

All these steps can be done with plink 1.90 software [45].

Because base data and training data come from different sources, some QC steps need to be taken according to Choi *et al.* [36]:

1) Check the integrity of the transferred file with a software like md5sum [46];
2) Ensure that input data have genomic position of the same genome build with LiftOver program [47];
3) Define the effect allele and the reference allele from the base data since some GWAS summaries categorize the allele as risk allele/non-risk allele and major/minor allele;
4) Process the ambiguous SNPs due to unknown chromosome strand (sense/antisense) from different DNA read platform; removing duplicated SNPs;
5) Ensure the independence of base data and training data by removing overlapping samples and closely related individuals (1st/2nd degree relatives);
6) Check chip-heritability from GWAS summary statistics by using LD score regression [48] to avoid reaching misleading conclusion (because the predictive power of the PRS cannot surpass the power and predictive accuracy of the base GWAS [49]).

TABLE I
DIFFERENT METHODS OF COMPUTING PRS. ADAPTED FROM A STUDY BY CHOI ET AL. (2018) [36]

	Clumping + thresholding	**Penalized Regression**	**Bayesian Shrinkage**
Shrinkage of SNPs' Effect Size	P-value threshold	LASSO, penalty parameters, Elastic Net	Fraction of causal SNPs
Handling of Linkage Disequilibrium	Clumping	LD matrix integral to the algorithm	Shrink effect sizes with respect to LD
Software	PRSice [8]	Lassosum [43] bigstatr [9]	LDpred [44]

2. PRS Calculation

After QC has been performed, PRSs are calculated for all individuals in the training data. There are many methods to calculate the polygenic risk score (see Table I), but a good method has to take into account the following factors. The first factor is the effect sizes of SNPs taken from GWAS summary statistics. These effect sizes consist of true association with the disease and some degree of unknown random variations which produce "winner's curse" effect among the most significant SNPs. The second one is the number of SNPs included in the calculation. And the third one is the correlation among SNPs, *i.e.*, linkage disequilibrium. LD difference between base and training data can make the PRS misestimate the risk of disease.

Clumping-and-thresholding is the most common method to compute the PRS. Clumping in this context means selecting the most significant SNP in a block of related SNPs based on LD (usually $r^2 > 0.2$). The p-value threshold is usually genome-wide significant 5×10^{-8} or the threshold that maximizes the AUC is selected. The clumping-and-thresholding method is fast and easy to use. It can capture the independent effect of the SNPs. However, a compelling criticism of this method is the removal of SNPs in LD with arbitrary r^2 value. Clumping-and-thresholding is automated in PRSice software [8]. This is one of the most common tools for computing the PRS. The processes of calculating, evaluating and visualizing the result are automated in the software.

Penalized regression is a method to integrate LD information of the population of interest into GWAS summary statistics. This technique allows for decreasing the variance at the cost of introducing some bias. The selection of tuning parameters λ and s in the elastic net technique results in a model's best fit. This model is complicated to calculate but performs better than the clumping-and-thresholding method in a small sample size [9].

Bayesian shrinkage is a method that infers the mean effect size of SNP based on a reference LD panel and assumption of causal SNP fraction in the genotype. Simulation and empirical data prove that this method outperforms the clumping-and-thresholding method, particularly in a large sample size [44, 50].

Penalized regression and Bayesian shrinkage lead to millions of SNPs included in the PRS calculation. Most of these SNPs have a negligible effect on individual risk score. Their effect sizes are so small and would have been zero if the effect size would be rounded up to 3 or 4 digits after the decimal point. This approach aims to maximize the C-statistics even if the majority of the included SNPs have minimal effect on individual risk score [50]. This approach, however, convolutes the PRS formula beyond the biological and medical interpretation.

3. Optimizing PRS and Result Presentation

The association between the PRS and the phenotype can be quantified with goodness-of-fit metrics such as the estimate of effect size (beta for quantitative phenotype and OR for binary phenotype), p-value of the null hypothesis test of no association, explained variance (R^2) and area under the curve (AUC).

Explained variance is a statistical measure to quantify the discrepancy between the PRS model and the observed phenotype. A higher percentage of explained variance means a better association of the model, which also means that the PRS has better predictive performance. In linear regression analysis for quantitative phenotype, explained variance is the coefficient of determination. In logistic regression analysis for binary phenotype, explained variance is the pseudo-R^2. Various types of pseudo-R^2 metrics are used in epidemiology field. Among them, Nagelkerke R^2 is perhaps the most popular [51].

The receiver operating characteristic curve illustrates the true positive rate (sensitivity) versus false-positive rate (specificity). The AUC represents the accuracy of the test. A test with AUC between 0.9–1 is an excellent test. A test with AUC between 0.5–0.6 is a worthless test [52].

Result presentation of the PRS analysis should emphasize its relevant usage which is a prognosis of disease liability. The main goal of the PRS is the identification of a group of individuals with high-risk of disease based on genetic factors. Therefore, the result should stratify the population into distinct tiers of risk based on the percentile rank cut-off value, usually top 1 percentile, top 10 percentile, top 20

percentile, etc., *e.g.*, as presented in Figure 1. Generally, the medical community has already defined the appropriate level of risk versus benefits that justifies certain medical interventions [53, 54]. When an individual is determined to have high-risk for a disease, we can infer the relative risk of said individual compared to the reference. The higher the relative risk of said individual is, the more justified the medical intervention becomes.

PRS analysis can also be represented based on age (Figure 2). The cumulative risk of disease stratified by the PRS can guide the decision at which age an individual can benefit the most from a screening test [55]. This age-based criterion can spotlight the balance of average risk of breast cancer and the risk of harm due to the false-positive result.

4. Validating PRS Performance

A common concern in PRS analysis is whether the most optimized PRS overfits the training data [56]. As a result, applying said PRS to the general population can lead to inflated results and false conclusions. The best strategy to prevent overfitting of the PRS-based prediction model is to validate its accuracy on an independent data set. In the absence of an independent data set, the training data can be divided into 2 separated data sets, one for optimizing the PRS and the other for performing out-of-sample prediction [57].

VI. Conclusion

The cost of reading DNA is becoming more and more affordable through advancement of genotyping and sequencing technologies. Alongside the development of data storage, new computing methods and abundance of disease databases, the PRS has provided better accuracy to existing models of risk prediction for common diseases. Consequently, individual clinical management (*e.g.*, disease screening and therapeutic intervention) can be personalized based on individual genetic information. This genetic information can be obtained at any point in life with a minimally invasive procedure (*e.g.*, blood draw or saliva sample) and a single genotype data can be analyzed to provide estimations for many diseases simultaneously. Although the medical community still has doubt and hesitation regarding implementation of the PRS, it will continue to improve and have larger impact in the near future.

References

[1] T. A. Manolio, "Genomewide association studies and assessment of the risk of disease," *New England Journal of Medicine*, vol. 363, no. 2, pp. 166–176, 2010.

[2] T. A. Manolio, F. S. Collins *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[3] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge *et al.*, "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies," *Nature Genetics*, vol. 45, no. 4, pp. 400–405, 2013.

[4] J. N. Cooke Bailey and R. P. Igo Jr, "Genetic Risk Scores," *Current Protocols in Human Genetics*, vol. 91, no. 1, pp. 1.29.1–1.29.9, 2016.

[5] A. C. J. Janssens, "Validity of Polygenic Risk Scores: Are we measuring what we think we are?" *Human Molecular Genetics*, vol. 28, no. R2, pp. R143–R150, 2019.

[6] 1000 Genomes Project Consortium and others, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[7] J. MacArthur, E. Bowler *et al.*, "The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog)," *Nucleic Acids Research*, vol. 45, no. D1, pp. D896–D901, 2017.

[8] J. Euesden, C. M. Lewis, and P. F. O'Reilly, "PRSice: Polygenic risk score software," *Bioinformatics*, vol. 31, no. 9, pp. 1466–1468, 2015.

[9] F. Privé, H. Aschard, and M. G. Blum, "Efficient implementation of penalized regression for genetic risk prediction," *Genetics*, vol. 212, no. 1, pp. 65–74, 2019.

[10] G. Versmée, L. Versmée, M. Dusenne, N. Jalali, and P. Avillach, "dbgap2x: An R package to explore and extract data from the database of Genotypes and Phenotypes (dbGaP)," *Bioinformatics*, vol. 36, no. 4, pp. 1305–1306, 2020.

[11] C. Bycroft, C. Freeman *et al.*, "The UK biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[12] A. Torkamani, N. E. Wineinger, and E. J. Topol, "The personal and clinical utility of polygenic risk scores," *Nature Reviews Genetics*, vol. 19, no. 9, pp. 581–590, 2018.

[13] S. A. Lambert, G. Abraham, and M. Inouye, "Towards clinical utility of polygenic risk scores," *Human Molecular Genetics*, vol. 28, no. R2, pp. R133–R142, 2019.

[14] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[15] G. Abraham, A. S. Havulinna *et al.*, "Genomic prediction of coronary heart disease," *European Heart Journal*, vol. 37, no. 43, pp. 3267–3278, 2016.

[16] R. S. Rosenson and C. C. Tangney, "Antiatherothrombotic properties of statins: Implications for cardiovascular event reduction," *JAMA*, vol. 279, no. 20, pp. 1643–1650, 1998.

[17] P. Natarajan, R. Young *et al.*, "Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting," *Circulation*, vol. 135, no. 22, pp. 2091–2101, 2017.

[18] J. G. Elmore, "Screening for breast cancer: Strategies and recommendations," *Retrieved from the Up to Date website*, 2019. [Online]. Available: http://www.uptodate.com/contents/screening-for-breast-cancer

[19] N. Pashayan, S. Morris, F. J. Gilbert, and P. D. Pharoah, "Cost-effectiveness and benefit-to-harm ratio of risk-stratified screening for breast cancer: A life-table model," *JAMA Oncology*, vol. 4, no. 11, pp. 1504–1510, 2018.

[20] P. Maas, M. Barrdahl *et al.*, "Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States," *JAMA Oncology*, vol. 2, no. 10, pp. 1295–1302, 2016.

[21] A. Lee, N. Mavaddat *et al.*, "BOADICEA: A comprehensive breast cancer risk prediction model incorporating genetic and non-genetic risk factors," *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, vol. 21, no. 8, pp. 1708–1718, 2019.

[22] R. M. Hoffman, "Screening for prostate cancer," 2019. [Online]. Available: www.uptodate.com/contents/screening-for-prostate-cancer

[23] T. M. Seibert, C. C. Fan *et al.*, "Polygenic hazard score to guide screening for aggressive prostate cancer: Development and validation in large scale cohorts," *BMJ*, vol. 360, 2018.

[24] M. J. Redondo, S. Geyer *et al.*, "A type 1 diabetes genetic risk score predicts progression of islet autoimmunity and development of type 1 diabetes in individuals at risk," *Diabetes Care*, vol. 41, no. 9, pp. 1887–1894, 2018.

[25] J. M. Sosenko, J. P. Krischer *et al.*, "A risk score for type 1 diabetes derived from autoantibody-positive participants in the diabetes prevention trial–type 1," *Diabetes Care*, vol. 31, no. 3, pp. 528–533, 2008.

[26] K. Lall, R. Magi, A. Morris, A. Metspalu, and K. Fischer, "Personalized risk prediction for type 2 diabetes: The potential of genetic risk scores," *Genetics in Medicine*, vol. 19, no. 3, pp. 322–329, 2017.

[27] R. S. Desikan, C. C. Fan *et al.*, "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score," *PLoS Medicine*, vol. 14, no. 3, p. e1002258, 2017.

[28] A. T. Marees, H. de Kluiver *et al.*, "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis," *International Journal of Methods in Psychiatric Research*, vol. 27, no. 2, p. e1608, 2018.

[29] P. G. Bagos, "Genetic model selection in genome-wide association studies: Robust methods and the use of meta-analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 12, no. 3, pp. 285–308, 2013.

[30] D. Thomas, "Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies," *Annual Review of Public Health*, vol. 31, no. 1, pp. 21–36, 2010.

[31] J. H. Moore, "Computational analysis of gene-gene interactions using multifactor dimensionality reduction," *Expert Review of Molecular Diagnostics*, vol. 4, no. 6, pp. 795–803, 2004.

[32] P. W. Wilson, J. B. Meigs, L. Sullivan, C. S. Fox, D. M. Nathan, and S. D'Agostino, R. B., "Prediction of incident diabetes mellitus in middle-aged adults: The framingham offspring study," *Archives Internal Medicine*, vol. 167, no. 10, pp. 1068–1074, 2007.

[33] B. J. Keating, "Advances in risk prediction of type 2 diabetes: Integrating genetic scores with Framingham risk models," *Diabetes*, vol. 64, no. 5, pp. 1495–1497, 2015.

[34] L. Duncan, H. Shen *et al.*, "Analysis of polygenic risk score usage and performance in diverse human populations," *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[35] M. Khoury, "Is it time to integrate polygenic risk scores into clinical practice? Let's do the science first and follow the evidence wherever it takes us," *Centers for Disease Control and Prevention*, 2019. [Online]. Available: https://blogs.cdc.gov/genomics/2019/06/03/is-it-time/

[36] S. W. Choi, T. S. H. Mak, and P. O'reilly, "A guide to performing Polygenic Risk Score analyses," *BioRxiv*, 2018.

[37] K. Wetterstrand, "The cost of sequencing a human genome," *National Human Genome Research Institute*, 2019. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

[38] NHGRI, "DNA sequencing fact sheet," *National Human Genome Research Institute*, 2015. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet.

[39] M. Francisco and C. D. Bustamante, "Polygenic risk scores: A biased prediction?" *Genome Medicine*, vol. 10, no. 1, pp. 1–3, 2018.

[40] V. S. Le, K. T. Tran *et al.*, "A Vietnamese human genetic variation database," *Human Mutation*, vol. 40, no. 10, pp. 1664–1675, 2019.

[41] M. D. Mailman, M. Feolo *et al.*, "The NCBI dbGaP database of genotypes and phenotypes," *Nature Genetics*, vol. 39, no. 10, pp. 1181–1186, 2007.

[42] E. M. Ramos, D. Hoffman *et al.*, "Phenotype–Genotype Integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources," *European Journal of Human Genetics*, vol. 22, no. 1, pp. 144–147, 2014.

[43] T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham, "Polygenic scores via penalized regression on summary statistics," *Genetic Epidemiology*, vol. 41, no. 6, pp. 469–480, 2017.

[44] B. J. Vilhjálmsson, J. Yang *et al.*, "Modeling linkage disequilibrium increases accuracy of polygenic risk scores," *The American Journal of Human Genetics*, vol. 97, no. 4, pp. 576–592, 2015.

[45] S. Purcell, B. Neale *et al.*, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[46] U. Drepper, S. Miller, and D. Madore, "md5sum: Verify compact digital fingerprint of a file (GNU GPL version 3 or later)," *Free Software Foundation*, 2010. [Online]. Available: linux.die.net/man/1/md5sum

[47] R. M. Kuhn, D. Haussler, and W. J. Kent, "The UCSC genome browser and associated tools," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 144–161, 2013.

[48] B. K. Bulik-Sullivan, P.-R. Loh *et al.*, "LD score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nature Genetics*, vol. 47, no. 3, p. 291, 2015.

[49] F. Dudbridge, "Power and predictive accuracy of polygenic risk scores," *PLoS Genetics*, vol. 9, no. 3, 2013.

[50] A. V. Khera, M. Chaffin *et al.*, "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations," *Nature Genetics*, vol. 50, no. 9, pp. 1219–1224, 2018.

[51] S. H. Lee, M. E. Goddard, N. R. Wray, and P. M. Visscher, "A better coefficient of determination for genetic profile analysis," *Genetic Epidemiology*, vol. 36, no. 3, pp. 214–224, 2012.

[52] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[53] P. M. Ridker, J. G. MacFadyen *et al.*, "Rosuvastatin for primary prevention among individuals with elevated high-sensitivity C-reactive protein and 5% to 10% and 10% to 20% 10-year risk," *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, no. 5, pp. 447–452, 2010.

[54] P. C. Gøtzsche and O. Olsen, "Is screening for breast cancer with mammography justifiable?" *The Lancet*, vol. 355, no. 9198, pp. 129–134, January 2000.

[55] G. A. Colditz and B. Rosner, "Cumulative risk of breast cancer to age 70 years according to risk factor status: Data from the Nurses' Health Study," *American Journal of Epidemiology*, vol. 152, no. 10, pp. 950–964, 2000.

[56] B. A. Goldstein, L. Yang, E. Salfati, and T. L. Assimes, "Contemporary considerations for constructing a genetic risk score: An empirical approach," *Genetic Epidemiology*, vol. 39, no. 6, pp. 439–445, 2015.

[57] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: A multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

Nguyen Tran The Hung received his doctor of medicine degree from Universities of Medicine and Pharmacy of Ho Chi Minh city (Viet Nam) in 2016. He then got a master degree in biomedical science from China Medical Universities (Taichung, Taiwan) in 2019. His research field is human genetic and diabetes mellitus. He worked briefly as a pediatrician before pursuing his career in academia as a research scientist at Vingroup Big Data Institute from 2019 until now. His thesis on type 2 diabetic nephropathy and the application of polygenic risk score made him believe in the potential impact that genetic research can make in healthcare.

Le Duc Hau obtained his PhD degree in Bioinformatics from University of Ulsan, Republic of Korea in 2012. He is now leading the Department of Computational Biomedicine, Vingroup Big Data Institute, Vietnam. He has been focusing on proposing computational methods for disease- and drug-related problems in personalized medicine, especially on identification of disease-associated biomarkers, prediction of drug targets and response. In parallel, he has been developed bioinformatics tools. So far, he has been published more than fifty papers in well-recognized journals and conferences, nearly a half of those are in ISI-indexed journals. In addition, he has been a member of program committees and reviewer of several international conferences/journals. Moreover, he is a principal investigator and a key member of some national/ministry-level projects. Specially, he is the principal investigator of the biggest genome project in Vietnam (i.e., building databases of genomic variants for Vietnamese population). Finally, he has been collaborating with some well-recognized international research institutes.

Phylogenetic and Phylogenomic Analyses for Large Datasets

Le Sy Vinh

University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

Email: vinhls@vnu.edu.vn

Communication: received 28 October 2019, revised 30 December 2019, accepted 30 December 2019

Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.898

The Editor coordinating the review of this article and deciding to accept it was Prof. Le Duc Hau

Abstract: The phylogenetic tree is a main tool to study the evolutionary relationships among species. Computational methods for building phylogenetic trees from gene/protein sequences have been developed for decades and come of age. Efficient approaches, including distance-based methods, maximum likelihood methods, or classical maximum parsimony methods, are now able to analyze datasets with thousands of sequences. The advanced sequencing technologies have resulted in a huge amount of data including whole genomes. A number of methods have been proposed to analyze the whole-genome datasets, however, numerous challenges need to be addressed and solved to translate phylogenomic inferences into practices. In this paper, we will analyze widely-used methods to construct large phylogenetic trees, and available methods to build phylogenomic trees from whole-genome datasets. We will also give recommendations for best practices when performing phylogenetic and phylogenomic analyses. The paper will enable researchers to comprehend the state-of-the-art methods and available software to efficiently study the evolutionary relationships among species from large datasets.

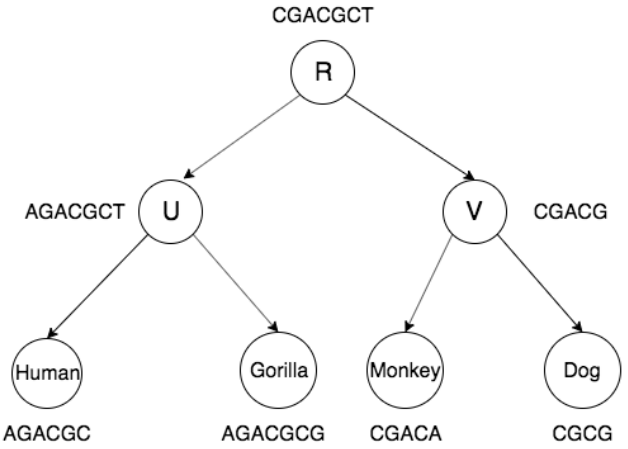


Figure 1. A phylogenetic tree represents the evolutionary relationships among species.

Keywords: *DNA sequence, evolutionary relationship, senome, phylogenetics, phylogenomics, large dataset, protein sequence.*

I. Introduction

Phylogenetic reconstruction is one of the most essential means to study the relationships among species. Analyzing the relationships among species shed light on the evolution of species. The phylogenetic information also enables us to study the structures and functions of DNA/protein sequences. Phylogenetic inference is an active research field for decades, and phylogenetic tree reconstruction methods based on molecular data have been developed and employed to study the evolution of species in tens of thousands of studies.

The evolutionary relationships among species are typically described by a binary tree where external nodes represent for present species; internal nodes represent for ancestors of species, and branches reflex connections and genetic changes between species (see Figure 1). Computational approaches have been proposed and implemented to construct phylogenetic trees based on several types of data, including morphological characters and molecular data, *i.e.*, DNA sequences, protein sequences, or even whole genomes. In this paper, we focus on phylogenetic inferences from molecular data.

The first class of phylogenetic tree construction approaches is distance-based methods. The phylogenetic tree is built based on genetic distances between sequences, *i.e.*, species with small distance should be placed closely in the tree. A number of distance-based methods have been proposed to reconstruct large phylogenetic trees [1–3]. Among them, the Neighbor-joining method and its improved modification [1, 2] have been used to build phylogenetic trees for tens of thousands of evolutionary studies.

The maximum parsimony approach searches the best phylogenetic tree that minimizes the number of changes to explain the difference among species. The maximum parsimony methods are simple, intuitive, and widely used to build phylogenetic trees [4–7]. However, they suffer a systematic limitation that they cannot describe complex changes occurred along branches of the tree. Thus, the maximum parsimony methods are suitable for analyzing datasets of closely related species.

The maximum likelihood approach has been developed to search for the tree that maximizes the probability of data [8–11]. The maximum likelihood methods are normally better than distanced-based and maximum parsimony methods, however, they are computationally expensive for large datasets. Heuristic approaches have been proposed to search maximum likelihood trees for datasets with thousands of species. Currently, maximum likelihood methods are most widely used for phylogenetic inferences.

The current sequencing technologies enable us to obtain whole genomes of thousands of species. Building phylogenomic trees based on whole-genome data becomes a common practice. Numerous challenges must be considered when analyzing whole genomes, including computational burden, the heterogeneity of evolutionary models for different genes, or gene structural variation in the genomes. Both computational methods, evolutionary models, and efficient software must be expanded to handle all the requirements when analyzing large genome datasets.

II. Data and Models

1. Data

Nowadays, molecular data, *i.e.*, DNA and protein sequences are the most common data type used to study evolutionary relationships among species. Each species is represented by one or multiple sequences, even by its whole genome. A DNA sequence (or genome) is coded as a string of 4 different nucleotides, *i.e.*, A, C, G, T; while a protein sequence is described as a string of 20 different amino acid types. Each nucleotide/amino acid in a DNA/protein sequence is called a character. Two characters in two genomes are called homologous if they have derived from the same common ancestor character. Generally, two DNA/protein sequences are called homologous if they originated from the same ancestor sequence. Note that we are only able to obtain molecular sequences from current species. In other words, we are not able to collect molecular data from common ancestor species for our studies. The data from ancestor species are normally inferred from the data of their descendants.

Three common kinds of character changes during the evolution are substitutions, insertions, and deletions. The substitutions will change the content of sequences. As a result, two homologous sequences originated from the same ancestor might be different. The insertion/deletion events (indels for short) during the evolution resulted in homologous sequences with different lengths. The first task in phylogenetic analysis is to align homologous sequences to create a multiple sequence alignment such that characters at the same column are assumed to be homologous (see Table I). The difference between homologous characters is phylogenetic signals reflexing the evolutionary relationships among species. A number of alignment methods have been proposed to align DNA/protein sequences, notably ClustalW [12] and Muscle [13]. They apply a progressive aligning strategy to build a multiple sequence alignment that minimizes the number of changes among sequences. Multiple sequence alignments are the input for phylogenetic or phylogenomic inferences.

TABLE I
A multiple sequence alignment of four sequences where '-' character represents for indels

	1	2	3	4	5	6	7
Human	A	G	A	C	G	C	-
Gorilla	A	G	A	C	G	C	G
Monkey	C	G	A	C	A	-	-
Dog	C	G	-	C	G	-	-

2. Models

The substitution process of characters during the evolution is complex. It can be typically simplified and modeled by a Markov process with following assumptions [14]:

- The rate of change from a character i to another nucleotide j is independent of the history of nucleotide i (Markov property);
- The substitution rate is time-homogeneous, *i.e.*, constant over the time course;
- The substitution between characters is time-continuous, *i.e.*, occurring at any time during the evolution process;
- The frequencies of characters are at equilibrium (stationarity).

The substitution process is represented by an instantaneous substitution rate matrix $Q = \{Q_{ij}\}$ where Q_{ij} is the number of substitutions from character i to character j per time unit. Note that matrix Q is a 4×4 matrix for nucleotide model or 20×20 matrix for amino acid model.

As usual, the substitution process in phylogenetic analyses is also assumed to be time-reversible, *i.e.*, the relative substitution rates between nucleotide i and nucleotide j are

TABLE II
THE SUBSTITUTION MODEL FOR NUCLEOTIDES

$Q =$		A	C	G	T
	A	$-r_{AC}\pi_C$ $-r_{AG}\pi_G$ $-r_{AT}\pi_T$	$r_{AC}\pi_C$	$r_{AG}\pi_G$	$r_{AT}\pi_T$
	C	$r_{CA}\pi_A$	$-r_{CA}\pi_A$ $-r_{CG}\pi_G$ $-r_{CT}\pi_T$	$r_{CG}\pi_G$	$r_{CT}\pi_T$
	G	$r_{GA}\pi_A$	$r_{GC}\pi_C$	$-r_{GA}\pi_A$ $-r_{GC}\pi_C$ $-r_{GT}\pi_T$	$r_{GT}\pi_T$
	T	$r_{TA}\pi_A$	$r_{TC}\pi_C$	$r_{TG}\pi_G$	$-r_{TA}\pi_A$ $-r_{TC}\pi_C$ $-r_{TG}\pi_G$

the same in both directions. As a result, the instantaneous substitution rate matrix Q can be decomposed into a symmetric relative substitution rate matrix $R = \{r_{ij}\}$ and character frequency vector π. Technically, $Q_{ij} = \pi_j r_{ij}$ if $i \neq j$, otherwise, $Q_{ij} = -\sum_{x \neq i} Q_{ix}$ (see Table II).

It is well known that the substitution rates are heterogeneous among character sites, *i.e.*, the substitution rates are normally fast at unimportant functional sites and slow at important functional sites. The rate models have been proposed to describe the rate heterogeneity. The widely-used rate model in phylogenetic analysis is the combination of a gamma rate model G and an invariant site rate model I [15]. The gamma rate model G assumes that the substitution rates at sites follow a gamma distribution that can be approximated by a discrete gamma distribution with several categories. The invariant rate model I assumes that a certain portion of sites in the alignment is invariant.

The parameters of substitution models and/or rate models can be directly estimated from the dataset under the study, except the amino acid substitution models. The amino acid substitution models consist of more than 200 parameters that cannot be properly estimated from small or medium datasets under the study. Normally, the parameters of amino acid substitution models are empirically estimated from large datasets [16–19].

III. PHYLOGENETIC ANALYSIS

Given $D = \{d_1, d_2, \ldots, d_n\}$ is a multiple sequence alignment of n sequences with length l, where n sequences representing n species, the phylogenetic tree reconstruction method will build a binary tree T with n leaves representing n sequences, and internal nodes representing ancestors. As we do not have data from ancestors, it is hard to determine the root of a phylogenetic tree. Therefore, the phylogenetic tree is typically represented by a binary unrooted tree. The number of binary unrooted tree structures with n leaves is $\prod_{i=3}^{n}(2i-5)$ that increases exponentially with the number of leaves. Searching the best tree for large n is computationally expensive. A number of heuristic approaches have been proposed to construct phylogenetic trees based on different criteria. In this section, we will discuss three widely-used approaches, including distance-based approach, maximum parsimony approach, and maximum likelihood approach, to build large phylogenetic trees.

1. Distance-based Approach

Analyzing the evolutionary relationships among species based on genetic distances among species have been introduced for more than fifty years [20]. The distance-based approach builds a phylogenetic tree such that closely related sequences with small distances should be placed nearby on the tree. The distance-based method comprises two steps: estimating the genetic distance matrix between sequences and constructing a tree based on the distance matrix. The genetic distance between two sequences can be efficiently estimated using the maximum likelihood method. We note that the accuracy of the estimated distance between two sequences increases with the length of sequences.

Let $D = \{d_{ij}\}$ be a distance matrix where d_{ij} is the genetic distance estimated between two sequences i and j. Let p_{ij} be the length of the path connecting two sequences i and j on the tree. The distance-based approach builds a tree such that the path length p_{ij} on the tree reflexes the distance d_{ij} on the distance matrix D for all pairs of sequences. Minimizing the total square difference is the best statistically justified distance-based objective [14]. Technically, the distance discrepancy $\Delta(T) = \sum_{u=1}^{n}\sum_{v=1}^{n}(d_{uv} - p_{uv})^2$ is the objective function of the least square method. The least square method will construct a tree T to minimize the distance discrepancy $\Delta(T)$. Given a tree structure, the first task of the least square method is the estimation of branch lengths to minimize the function $\Delta(T)$. This task can be efficiently solved by algebraic analysis [21]. As the number of tree structures increases exponentially, searching the least square tree is an NP-complete problem [22]. A number of heuristic distance-based methods have been proposed to solve the problem.

Among distance-based methods, neighbor joining method (NJ) or its modified version BioNJ is the most popular and widely used to construct large phylogenetic trees [1, 23]. The NJ algorithm starts from a star tree of n leaves, *i.e.*, each leaf is directly connected to the root and considered as a subtree. The NJ algorithm iteratively joins subtrees to build a whole binary unrooted tree (see

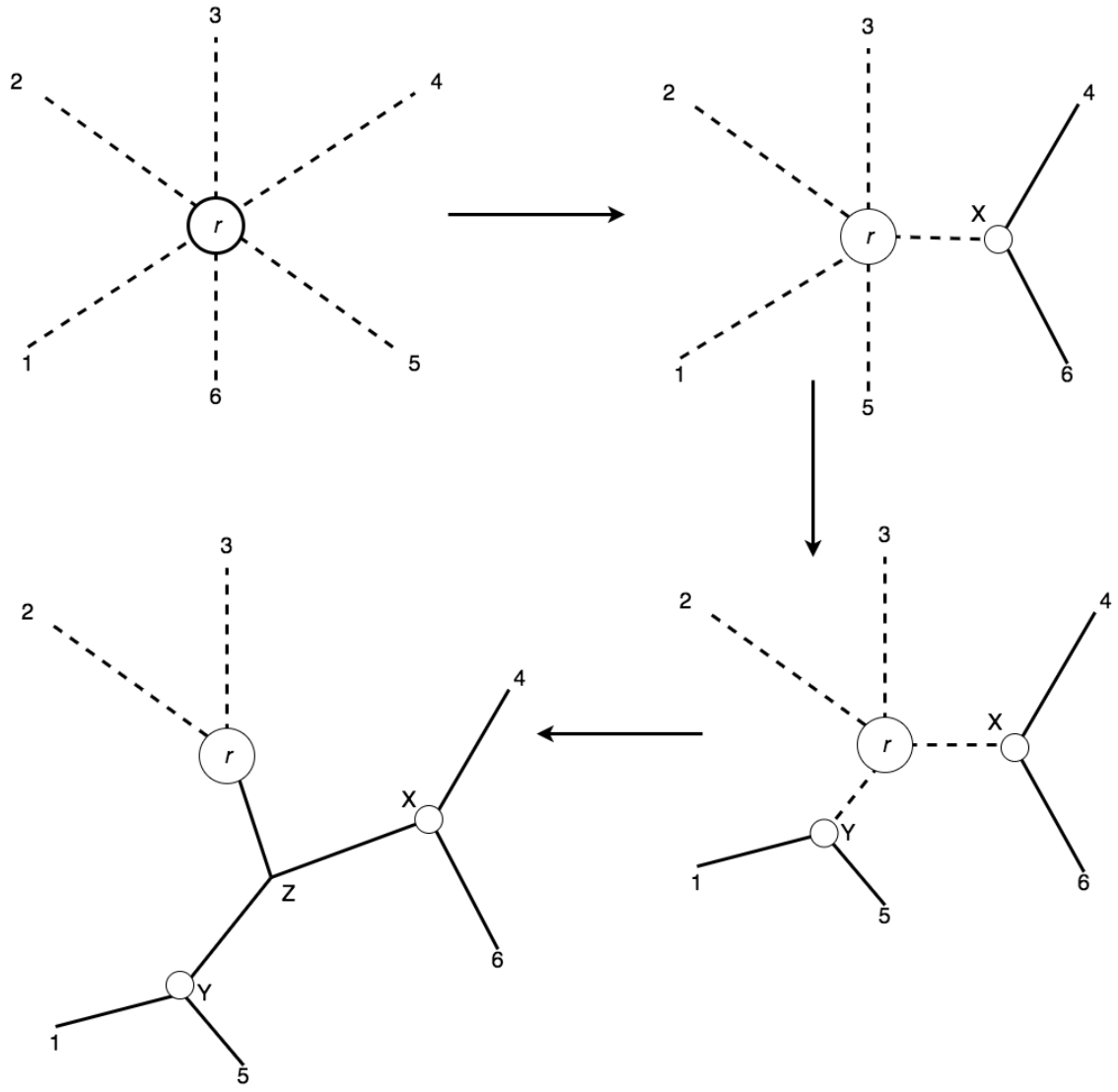


Figure 2. The neighbor joining algorithm to build a phylogenetic tree based on the distance matrix between sequences.

Figure 2). The total distance discrepancy based on the star tree is considerably large, therefore, the NJ algorithm step-by-step joins subtrees together to create a new subtree consisting of the two subtrees. The NJ algorithm iteratively joins two subtrees that helps to reduce the distance discrepancy at most. The complexity of the NJ algorithm is $O(n^3)$, thus, it is suitable to build distance-based trees for datasets containing several hundred sequences.

Other faster distance-based methods have been developed to build trees with larger datasets. Vinh and colleagues have proposed the shortest triplet clustering (STC) algorithm with the complexity of $O(n^2)$ to build distance-based phylogenetic trees for thousands of sequences (Vinh and von Haeseler 2005 [3]). Consider two leaves x and y, let r_{xy} be the most common ancestor of x and y. The key idea of the STC algorithm is the observation that if r_{xy} is the farthest internal node to the root of the tree, two leaves x and y should be neighbors on the tree. The STC algorithm uses the condition to iteratively group sequences to build a whole tree. Experiments on a wide range of simulated datasets showed that the STC algorithm was faster and more accurate than the NJ method. In practice, the NJ algorithm is still the most trusted and widely used distance-based method.

2. Maximum Parsimony Approach

Maximum parsimony approach is a simple and intuitive approach to construct phylogenetic trees from alignments. The approach searches for the tree that requires minimum number of changes along the tree (called the tree length) to explain the difference between sequences [24]. Given a tree with sequences at leaves, determining sequences at internal

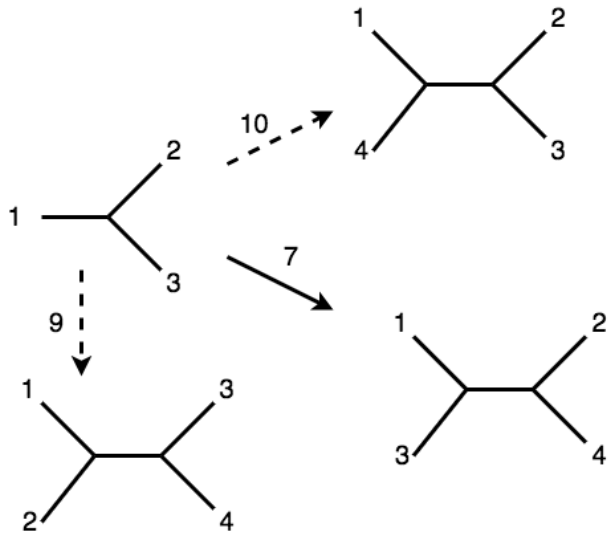


Figure 3. The stepwise addition algorithm to build a phylogenetic tree.

nodes to minimize the tree length can be efficiently solved by dynamic programming algorithm [25]. However, the number of tree structures increases exponentially with the number of sequences, searching the maximum parsimony tree is an NP-complete problem [26].

Several heuristic methods have been proposed to search the maximum parsimony trees. The key search strategy combines a stepwise addition tree construction method and a hill-climbing optimization. The stepwise addition tree construction method starts from a unique tree of three sequences, then iteratively adds new sequences into the current tree to construct a full tree of all sequences. A new sequence will be added into a branch of the current tree such that the length of the new tree is minimal (see Figure 3). As the stepwise addition tree construction method is a heuristic method, the constructed tree is normally still far away from the best tree. The constructed tree is further optimized by a hill-climbing method based on subtree rearrangements such as nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR), or tree bisection and reconnection (TBR).

The combination of stepwise addition tree construction algorithm and hill-climbing optimization results in reasonable trees, however, they are still local optimal trees. The tree constructed by the stepwise addition tree construction algorithm depends on the order of sequences adding to the tree, *i.e.*, different sequence orders result in different trees. In an attempt to find the best maximum parsimony tree, we can perform the search process several times. The tree with the minimum length found will be considered as the best tree. Several software has been developed to build maximum parsimony trees from DNA/protein sequences such as PAUP* [4], TNT [5], or MPBoot [6]. The TNT and MPBoot have been designed to build maximum parsimony trees with large datasets including thousands of sequences. Note that, there might exist multiple most parsimonious phylogenies for the same set of sequences. To solve the problem, TNT and MPBoot software are also able to quickly build bootstrap trees to assess the reliability of constructed trees.

The maximum parsimony methods do not assume any substitution model to represent the substitution process of characters. As a result, the main drawback of maximum parsimony methods is the inability of reflexing complex changes of characters along branches of the tree, *i.e.*, branches are not able to present parallel substitutions, back substitutions, or multiple substitutions. Thus, the length of maximum parsimony tree might underestimate the true number of character changes when analyzing datasets with largely diverse sequences. In other words, the maximum parsimony methods are suitable for analyzing datasets consisting of closely related species.

3. Maximum Likelihood Approach

Maximum likelihood approach has been developed to overcome the limitation of maximum parsimony approach. Given a multiple sequence alignment $A = \{a_1, a_2, \ldots, a_n\}$ of n sequences with length l; and a substitution model M, the maximum likelihood (ML) method determines an unrooted binary tree T to maximize the probability of alignment A based on tree T and model M. Specifically, determining T and M such that the likelihood value $L(M,T;A) = P(A|M,T)$ is maximum where $P(A|M,T)$ is the conditional probability of alignment A given tree T and model M.

As usual, we assume that the substitution processes among sites are independent, and follow the same tree T and model M. Let a^j be nucleotides at position j^{th} in the alignment A, the likelihood value $L(M,T;A)$ can be calculated from the likelihood values of all sites as follows:

$$L(M,T;A) = \prod_{j=1}^{l} L(M,T;a^j), \tag{1}$$

where $L(M,T;a^j) = P(a^j|M,T)$, the conditional probability of alignment a^j given tree T and model M.

The rate heterogeneity among sites should be taken into account when calculating the likelihood of a tree T. We recall that the site rates can be described by a rate model V combining an invariant rate model and a discrete gamma distribution with C classes whose rates are $r_1, r_2, \ldots, r_C$,

respectively [27, Yang (1994)]. Technically, the likelihood value of $L(M, V, T; A)$ is calculated as follows:

$$L(M, V, T; A) = \delta \prod_{j=1}^{l} L(inv; a^j) + (1 - \delta) \prod_{j=1}^{l} \frac{1}{C} \sum_{c} L(M, r_c T; a^j), \quad (2)$$

where $L(inv; a^j) = P(a^j | inv)$, the conditional probability of a^j given that all characters at the site are identical (invariant), and $L(M, r_c T; a^j) = P(a^j | M, r_c T)$, the conditional probability of a^j where $r_c T$ is the tree T whose lengths are multiplied with rate r_c.

Given a tree structure with sequences at leaves, a critical task is determining branch lengths of the tree to maximize its likelihood value. The task can be efficiently solved by numerical analyses such as Brent or Newton-Raphson's methods. Searching the maximum likelihood tree is an NP-Hard problem [28]. Numerous heuristic methods have been proposed to build maximum-likelihood trees for large datasets with thousands of sequences.

Maximum likelihood methods that use the combination of stepwise addition tree construction algorithm and hill-climbing optimizations have been developed for searching maximum likelihood trees such as RAxML [10]. Similarly, Guidon and Gascuel proposed the PhyML method for building large maximum likelihood trees. The key idea of the PhyML method is a quick hill-climbing method based on NNI operations. The PhyML algorithm tries to perform multiple independent good swaps of nearest neighbor subtrees simultaneously instead of performing one swap per time. The strategy allows PhyML to handle datasets with thousands of sequences, and result in reasonably good maximum-likelihood trees.

The quartet-based approach is another strategy to build phylogenetic trees. For each quartet of four sequences, there are only three possible binary unrooted tree structures (or quartet trees, see Figure 4). Thus, it is easy to determine the maximum likelihood quartet tree. The quartet-based approach starts from a unique tree of three sequences, and iteratively adds new sequences into the current tree to build a whole tree. The key idea of quartet-based methods for inserting new sequences is the condition that if $T(a, b|c, y)$ is the best quartet tree of four sequences including a new sequence y and three leaves a, b, c of the current tree, the new sequence y should not be placed on the path connecting two leaves a and b. The quartet-based methods insert new sequences into the current tree such that the condition is hold for most of the available quartets.

Quartet puzzling method is the first quartet-based method to build a maximum likelihood tree [29]. The method evaluates all n^4 possible quartets to build a phylogenetic tree for n sequences. It is able to handle datasets with a few hundred sequences. To overcome the high complexity, the Important Quartet Puzzling (IQP) method has been proposed to build trees with thousands of sequences. The key idea of the IQP method is the definition of the so-called "important quartet" that consists of closely related sequences [9]. The IQP method selects only n^2 important quartets to build a phylogenetic tree. More importantly, the IQP algorithm is combined with the quick hill-climbing optimization based on the nearest neighbor interchange operations to further optimize the quartet-based constructed tree. Experiments on both real and simulated datasets showed that the combined method, IQPNNI, found better maximum likelihood trees than the PhyML method. Note that the IQPNNI method was implemented to run in parallel mode to handle large datasets [30].

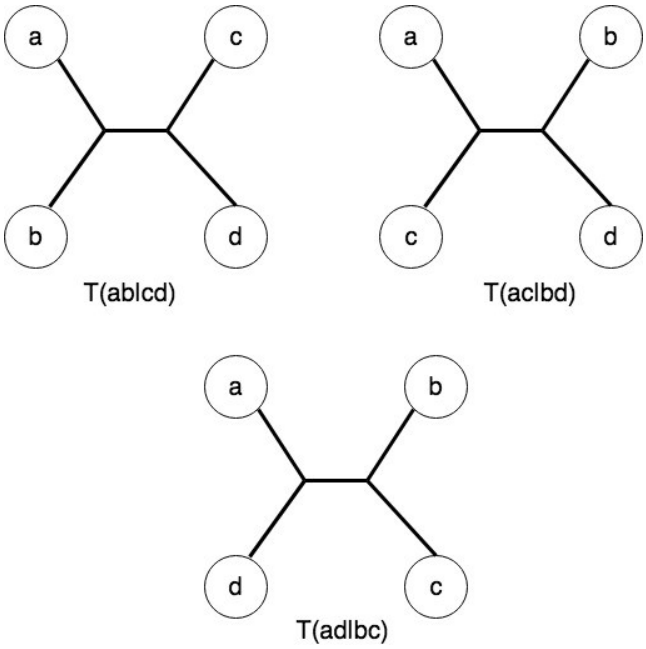


Figure 4. There different quartet trees topologies of a quartet.

The IQPNNI algorithm was improved into an efficient stochastic search algorithm for building large maximum likelihood trees [11]. They implemented IQ-TREE software that combined quick hill-climbing optimizations and a stochastic perturbation method to search maximum likelihood trees. The IQ-TREE perturbs current local optimal trees to escape from the current local optimal points and subsequently optimizes the perturbed trees by quick hill-climbing optimizations to search for the global optimal tree. The searching process is repeated several times and the best local optimal tree found will be considered as the best tree. Experiments showed that IQ-TREE was better than both

TABLE III
THE STRUCTURAL VARIANTS IN THE GENOMES.

	1	2	3	4	5	6	7	8	9
Human	G1	G2	G3	G4	G5	G6	G7	G8	G9
Gorilla	G1	G7	**G6**	**G5**	**G4**	**G3**	G7	G8	G9
Monkey	-	G5	G6	G7	G8	G9	**G2**	**G3**	**G4**
Dog	-	G5	G6	G7	G8	G9	**G4**	**G3**	**G2**

PhyML and RAxML methods in a majority number of cases tested. The IQ-TREE software is user-friendly and widely used by biologists for studying the evolution of species from molecular data.

IV. PHYLOGENOMIC ANALYSIS

Analyzing whole genomes to investigate the relationships among species is a comprehensive and challenging problem. In phylogenetic analysis, a genome is separated into a list of loci each corresponds to a gene or a region of interest in the genome. Homologous loci are aligned to create multiple sequence alignments. As a result, the input for phylogenomic analysis is a list of multiple sequence alignments. Phylogenetic tree construction approaches have been expanded to build phylogenomic trees from a list of multiple sequence alignments.

The first challenge in analyzing whole genomes is the occurrence of structural variants in the genomes. Besides variants occurring inside genes, other common types of structural changes are gene insertions/deletions, gene inversions (inverting the order of genes in the genome), gene transpositions (moving a number of genes in the genome from one position to another position in the genome), and inverted transpositions (combining both gene inversion and gene transposition events into one event). Table III demonstrates an example of structural variants in the genomes, *e.g.*, there is a gene inversion between the human genome and gorilla genome (genes G3, G4, G5, G6 in the human genome were inverted into G6, G5, G4, G3 in the gorilla genome).

The structural changes resulted in genomes with different structures. The structural difference between genomes can be used as phylogenetic signals for studying the evolution of species, *i.e.*, the number of structural changes to explain the structural difference between two genomes can be considered as the genetic distance between the genomes to evaluate their relationships. Overall, the genetic distance between two genomes is a combination of character changes inside genes and structural changes. Weighting and combining these changes properly for estimating the overall genetic distance between genomes enable us to build better phylogenetic trees using distance-based methods [31].

The second challenge when analyzing the genomes is model selection. The evolutionary models including rate models and substitution models might vary among loci. One model cannot properly reflex the evolutionary process of all loci. The phylogenomic analyses should use model selection methods to assign a proper model for each locus. As estimating model parameters is not strongly affected by the tree structure, model selection methods normally include two steps: building an initial tree and estimating model parameters based on the initial tree and the alignment. Building an initial tree can be done efficiently by distance-based methods such as NJ method. Provided that the initial tree is reasonably close to the best tree, it is good enough to estimate model parameters. Estimating model parameters for an alignment based on the initial tree can be solved by numerical optimization methods such as Brent's algorithm or more efficient Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Software like IQ-TREE provides us options to automatically determine the best model for each locus when analyzing whole-genome datasets. Note that normally we do not optimize parameters of amino acid models from each alignment because each alignment does not contain sufficient data to estimate a large number of parameters.

Finally, the computational expense is a critical burden of phylogenomic inferences. A genome dataset contains several dozens to thousands of genomes with the length up to several hundred million nucleotides. To overcome the problem, more efficient heuristic algorithms in terms of both running time and memory requirement should be developed to handle whole-genome datasets. As genomes are separated into loci, parallel computing is a promising approach to perform phylogenomic inference on individual alignments simultaneously. Most of the current widely-used phylogenetic software such as RAxML or IQ-TREE provide options to conduct phylogenetic inferences in parallel.

V. DISCUSSIONS

Phylogenetic inference is a core study in molecular biology. It is an active research field for several decades and the main focus of prominent researcher groups. Phylogenetic reconstruction for single or several genes perhaps come to age. The distance-based methods are able to build large reasonable phylogenies that could be used as starting points to search maximum parsimony or maximum likelihood trees. Nowadays, maximum likelihood methods such as RAxML, PhyML or IQ-TREE are able to efficiently construct trees with thousands of sequences.

All popular phylogenetic tree reconstruction software are based on heuristic search methods, therefore, the results from different software, or even from different runs of the same software, might not completely congruence. It is

especially true when analyzing datasets whose phylogenetic signals support polytomy tree structures. We recommend researchers to perform bootstrap analyses to assess the reliability of branches in the constructed tree. Although phylogenetic bootstrapping is computationally expensive, ultrafast bootstrap methods such as UFBoot2 [32] are able to build large bootstrap trees in an acceptable time.

Determining proper evolutionary models (*i.e.*, site rate models and/or substitution models) for datasets under the study is very critical in phylogenetic inferences. Using wrong evolutionary models in analyzing data will lead to inaccurate results [32]. The evolutionary models are normally selected from a list of existing models, and the model parameters can be directly estimated from the input data.

The advance of sequencing technologies has produced large genome datasets consisting of dozens to thousands of genomes with the length up to billion nucleotides. The large genome datasets provide us an unprecedented opportunity to study the relationships among species from their whole genomes. However, new efficient computational methods should be developed for phylogenomic inferences. The relationships among genomes should be analyzed at both levels: point level (*i.e.*, nucleotide/amino acid changes) and structural level (gene insertions/deletions as well as gene rearrangements). Combining changes at both levels to have a comprehensive evaluation is a new challenge for researchers in phylogenomic analyses. Another challenge in analyzing whole genomes is the heterogeneity of evolutionary processes between loci. Thus, determining proper evolutionary models is very critical when analyzing multiple genes or whole genomes. Currently, several software such as IQ-TREE are able to perform phylogenomic inferences for large genome datasets.

REFERENCES

[1] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[2] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data." *Molecular Biology and Evolution*, vol. 14, no. 7, pp. 685–695, 1997.

[3] L. S. Vinh and A. von Haeseler, "Shortest triplet clustering: reconstructing large phylogenies using representative sets," *BMC Bioinformatics*, vol. 6, no. 1, p. 92, 2005.

[4] J. C. Wilgenbusch and D. Swofford, "Inferring evolutionary trees with PAUP*," *Current Protocols in Bioinformatics*, no. 1, pp. 6–4, 2003.

[5] P. A. Goloboff, J. S. Farris, and K. C. Nixon, "TNT, a free program for phylogenetic analysis," *Cladistics*, vol. 24, no. 5, pp. 774–786, 2008.

[6] D. T. Hoang, L. S. Vinh, T. Flouri, A. Stamatakis, A. von Haeseler, and B. Q. Minh, "MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation," *BMC Evolutionary Biology*, vol. 18, no. 1, p. 11, 2018.

[7] A. Varón, L. S. Vinh, and W. C. Wheeler, "POY version 4: phylogenetic analysis using dynamic homologies," *Cladistics*, vol. 26, no. 1, pp. 72–85, 2010.

[8] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[9] L. S. Vinh and A. von Haeseler, "IQPNNI: moving fast through tree space and stopping in time," *Molecular Biology and Evolution*, vol. 21, no. 8, pp. 1565–1571, 2004.

[10] A. Stamatakis, "Using RAxML to infer phylogenies," *Current Protocols in Bioinformatics*, vol. 51, no. 1, pp. 6–14, 2015.

[11] L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," *Molecular Biology and Evolution*, vol. 32, no. 1, pp. 268–274, 2015.

[12] J. Thompson, D. Higgins, and T. Gibson, "ClustalW," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[13] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[14] F. Joseph, *Inferring Phytogenies*. Sunderland, MA, USA: Sinauer Associates, 2003.

[15] Z. Yang, "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites." *Molecular Biology and Evolution*, vol. 10, no. 6, pp. 1396–1401, 1993.

[16] S. Whelan and N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach," *Molecular Biology and Evolution*, vol. 18, no. 5, pp. 691–699, 2001.

[17] S. Q. Le and O. Gascuel, "An improved general amino acid replacement matrix," *Molecular Biology and Evolution*, vol. 25, no. 7, pp. 1307–1320, 2008.

[18] C. C. Dang, Q. S. Le, O. Gascuel, and V. S. Le, "FLU, an amino acid substitution model for influenza proteins," *BMC Evolutionary Biology*, vol. 10, no. 1, p. 99, 2010.

[19] D. T. Jones, W. R. Taylor, and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences," *Computational Applied Bioinformatics*, vol. 8, no. 3, pp. 275–282, 1992.

[20] L. L. Cavalli-Sforza and A. W. Edwards, "Phylogenetic analysis: models and estimation procedures," *American Journal of Human Genetics*, vol. 19, pp. 233–257, 1967.

[21] A. Rzhetsky and M. Nei, "Theoretical foundation of the minimum-evolution method of phylogenetic inference." *Molecular Biology and Evolution*, vol. 10, no. 5, pp. 1073–1095, 1993.

[22] W. H. Day and D. Sankoff, "Computational complexity of inferring phylogenies by compatibility," *Systematic Biology*, vol. 35, no. 2, pp. 224–229, 1986.

[23] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data." *Molecular Biology and Evolution*, vol. 14, no. 7, pp. 685–695, 1997.

[24] A. W. Edwards and Cavalli-Sforza, "The Reconstruction of Evolution," *Annals of Human Genetics*, vol. 27, pp. 105–106, 1963.

[25] W. M. Fitch, "Toward defining the course of evolution: minimum change for a specific tree topology," *Systematic Biology*, vol. 20, no. 4, pp. 406–416, 1971.

[26] R. Graham and L. Foulds, "Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time," *Mathematical Biosciences*, vol. 60, no. 2, pp. 133–142, 1982.

[27] Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate

methods," *Journal of Molecular Evolution*, vol. 39, no. 3, pp. 306–314, 1994.

[28] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees is hard," in *Annual International Conference on Research in Computational Molecular Biology*, 2005, pp. 296–310.

[29] K. Strimmer and A. Von Haeseler, "Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies," *Molecular Biology and Evolution*, vol. 13, no. 7, pp. 964–969, 1996.

[30] B. Q. Minh, L. S. Vinh, A. Von Haeseler, and H. A. Schmidt, "PIQPNNI: parallel reconstruction of large maximum likelihood phylogenies," *Bioinformatics*, vol. 21, no. 19, pp. 3794–3796, 2005.

[31] L. S. Vinh, A. Varón, and W. C. Wheeler, "Pairwise alignment with rearrangements," *Genome Informatics*, vol. 17, no. 2, pp. 141–151, 2006.

[32] D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, and L. S. Vinh, "UFBoot2: improving the ultrafast bootstrap approximation," *Molecular Biology and Evolution*, vol. 35, no. 2, pp. 518–522, 2018.

Le Sy Vinh obtained PhD in Bioinformatics from Heinrich Heine University, Duesseldorf, Germany 2005, subsequently followed a postdoc fellowship at American Museum of Natural History, NYC from 2005 to 2008. He is currently the Dean of the Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi.

Le Sy Vinh is an expert in phylogenetic analysis, the author of widely-used software such as IQPNNI, POY4, UFBoot2. He is the group leader of many human genome projects in Vietnam including the first Vietnamese human genome, building the comprehensive Vietnamese human genome database, or Autism spectrum disorder in Vietnamese children.

Liquid Biopsy to Characterize Cell-Free DNA in Cancer Detection and Monitoring

Nguyen Ngoc Tran

Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam

Email: v.trannn3@vintech.net.vn

Communication: received 21 October 2019, revised 30 December 2019, accepted 30 December 2019

Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.895

The Editor coordinating the review of this article and deciding to accept it was Prof. Le Duc Hau

Abstract: Liquid biopsy, a concept introduced approximately a decade ago, refers to noninvasive approaches that have become the focus of biomedical research. In clinical oncology and research, the term liquid biopsy is used in a broad sense as the sampling and analysis of analytes including cell-free DNA from various accessible biological fluids for diagnosis, prognosis and prediction of a therapeutic response. This technology has the potential to be used in tracking the genomic evolution of tumors over time. It may also have therapeutic implications in terms of its ability to detect actionable events or resistant subclonal populations while avoiding the need to conduct repeated biopsies. This paper briefly reviews the major advances in liquid biopsy assay technologies and discusses the types of cancers that most likely benefit from early detection.

Keywords: *Cell-free DNA, liquid biopsy, cancer detection.*

I. Introduction

Cancer has a significant impact on public health worldwide. One approach to lower its burden is through cancer screening and early detection. It is well established that patients have a higher cure rate and 5-year survival if diagnosed at early stages [1].

Tissue biopsy is the most widely-used tool for cancer detection, staging, and prognosis, but sometimes tumor tissues can be challenging to acquire. Moreover, it is impractical to use tissue biopsy for cancer screening and early diagnosis when the tumors are non-existent. Currently, several methods have been proved useful for cancer prevention such as mammography, Pap test and colonoscopy [2]. However, these screening methods have limited sensitivity and specificity, as well as limited applicability to certain cancer types. In order to perform large-scale cancer screening among healthy individuals, a more general and cost-effective approach is crucial.

In recent years, the rapid development of next-generation sequencing (NGS) technologies has led to a significant reduction in sequencing cost with improved accuracy. In the area of liquid biopsy, NGS has been applied to sequence circulating analytes. Analytes in the blood include cell-free DNA (cfDNA), which in cancer patients contains circulating tumor DNA (ctDNA), circulating extracellular vesicles (*e.g.*, exosomes, proteins, and metabolites) [3]. Collectively, these analytes can provide information about features of primary tumors or metastases for screening and early diagnosis of cancer (Figure 1).

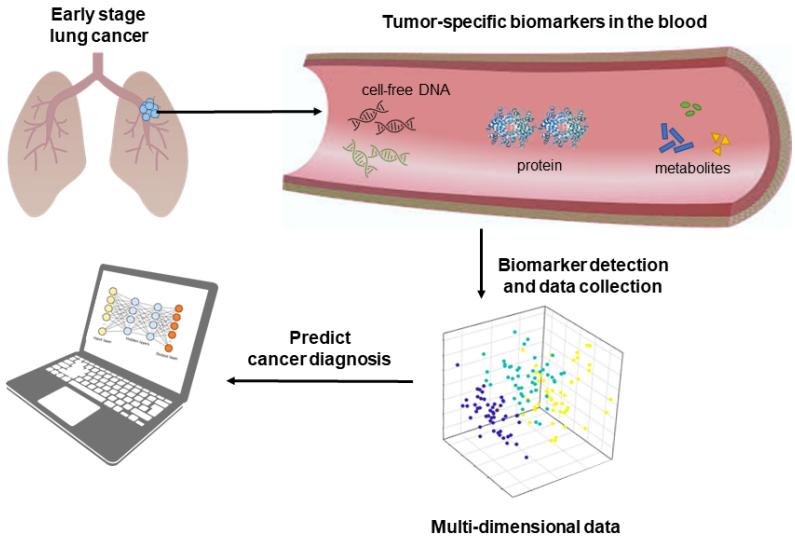
In addition to the information about genomic mutations and copy number alterations that is usually obtained from cfDNA, liquid biopsies are increasingly being used to generate information about the transcriptome, the epigenome, the proteome, and the metabolome. Among all analytes, cfDNA has the potential to revolutionize detection and monitoring of tumors. However, the methods reported to date have been limited by modest sensitivity, applicability to only a minority of patients, and the need for patient-specific optimization. To overcome these limitations, several new strategies have been developed for the analysis of cfDNA.

II. Cancer Types that Benefit from Early Detection

1. Pancreatic Cancer

Pancreatic cancer patients will most likely benefit from early detection by liquid biopsy. It is estimated that in 2019, about 57,000 people will be diagnosed with pancreatic cancer, which has fewer than 9% of patients surviving 5 years after diagnosis [4]. With no biomarkers for early detection, the deadly nature of pancreatic cancer is largely the result of late onset of symptoms when the cancer has already metastasized. Chemotherapy for pancreatic cancer involves conventional cytotoxic drug combinations but only extends life by clinically unimpressive margin [5].

The prevalence of pancreatic cancer is low in asymptomatic adults. Thus, an effective screening test for pancreatic cancer would have to be highly specific to achieve a



Credit: Adapted from Nat. Rev. Genet. 2018, DOI: 10.1038/S41576-018-0071-5

Figure 1. Liquid biopsies track individual's plasma for biomarkers such as DNA alterations, proteins, and metabolites released by tumor cells. Once these biomarkers are analyzed, the data are used to train machine-learning algorithms to distinguish between cancerous and noncancerous blood samples.

reasonable positive predictive value and to minimize the incorrect identification of healthy individuals. Currently used circulating biomarkers for pancreatic cancer, such as CA19-9, lack adequate sensitivity and specificity for early diagnosis when used alone. Additional biomarkers, including CA125 and LAMC2 have been investigated *in silico* as combinatory diagnostic strategies to CA19-9 with varying degrees of success [6]. The diagnosis and molecular analysis of pancreatic cancer may involve fine-needle aspiration or core needle biopsy. Yet due to the location of the pancreas and expense, repeated biopsies for longitudinal analysis are generally avoided [7]. Therefore, using tissue biopsies to longitudinally monitor pancreatic cancer is infeasible, whereas liquid biopsies provide a new opportunity for molecular profiling of the genetic landscapes of pancreatic cancer throughout disease progression.

The study of cfDNA in pancreatic cancer started in the 1980s, when a high level of cfDNA (> 100 ng/ml) was detected in 90% of pancreatic cancer patients [8]. Later on, KRAS mutations in the cfDNA of plasma from three pancreatic ductal adenocarcinoma (PDAC) patients were detected by PCR [9]. More recently, several groups have attempted to validate the clinical utility of mutation detection in the cfDNA of pancreatic cancer patients. For instance, Kinugasa *et al.* [10], detected KRAS mutations in 62.5% of serum samples from pancreatic cancer patients and found that these mutations were correlated with worse overall survival (OS). By ddPCR, Sausen *et al.* [11] detected KRAS mutations in the plasma cfDNA of 43% patients with early-stage pancreatic cancer. Moreover, using matched plasma samples before and after operations from 59 pancreatic cancer patients, Lee *et al.* [12] was able to detect KRAS mutations from 38 out of 42 (90.5%) patients. The ctDNA detected prior to operation was correlated with worse median recurrence free survival (RFS), 10.3 months versus not reached and after operation was 5.4 months versus 17.1 months. These data suggest that cfDNA is a promising prognostic biomarker for early diagnosis of pancreatic cancer.

2. Lung Cancer

Lung cancer is the most common cancer worldwide, accounting for 2.1 million new cases and 1.8 million deaths in 2018 [13]. More than half of people with lung cancer die within one year of being diagnosed [14]. When the cancer is detected at a localized stage (I–II), the 5-year survival rate is about 56% [4]. However, only 16% of lung cancer cases are diagnosed at an early stage. For advanced and metastatic tumors (stage IV), the 5-year survival rate is only 5% [15]. Therefore, it is imperative to identify diagnostic methods for early detection of lung cancer, enabling a timely treatment plan while potentially reducing healthcare costs [16].

One of the earliest and best examples is the use of cfDNA testing to identify the emergence of the epidermal

growth factor receptor (EGFR) T790M gatekeeper mutation after EGFR inhibitor therapy in *EGFR* mutant non–small-cell lung cancer. EGFR tyrosine kinase domain mutation status was strongly associated with the responsiveness to epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKI). Several studies have shown that *EGFR* mutations were detectable in serum samples obtained from patients with NSCLC. The *EGFR* mutation status in serum detected by different methods was correlated statistically with responsiveness to, and the progression-free survival of EGFR-TKI treatment [17, 18]. Indeed, the cobas *EGFR* Mutation Test can identify T790M and other *EGFR* driver mutations in plasma and has been approved by the Food and Drug Administration as a companion diagnosis for choosing specific *EGFR* therapies [19].

KRAS is an important molecule in the downstream signaling network of EGFR. The mutation rate of *KRAS* is about 15% to 30% in NSCLC. However, the occurrences of *KRAS* and *EGFR* mutations are often mutually exclusive and patients with *KRAS* mutations are found to be nonresponsive to EGFR-TKIs [20–22]. It has been determined that lung cancer with *KRAS* mutation may be resistant to EGFR-TKI and that *KRAS* mutation in plasma DNA correlates with the mutation status in the matched tumor tissues of patients with NSCLC [23, 24]. However, at present, no mature technology exists for early cancer detection. Such an approach would require a highly sensitive method in order to detect trace amounts of cfDNA released by preneoplastic lesions. It would also require high specificity to minimize false positive results in the large unaffected population undergoing screening. Also, even though we have made considerable progress in finding therapeutic biomarkers, less progress has been made to identify patients who are likely to have relapse after surgical resection.

One key study has shown that the postoperative detection of tumor-specific mutations in cfDNA can predict residual disease and tumor relapse in lung cancer [25]. This approach has the potential to become a critical tool in the postoperative management of the care of patients with cancer and is currently being tested in prospective clinical trials that will assess the usefulness of residual postoperative cfDNA detection to guide adjuvant chemotherapy.

3. Liver Cancer

Liver cancer ranks the fourth leading cause of cancer-related death worldwide. There are more than 841,000 patients diagnosed with liver cancer globally with the five-year survival rate of 18% [26]. Currently, surgical resection or liver transplantation is the primary therapy for liver cancer patients. Unfortunately, the majority of patients at the time of first liver cancer diagnosis, have already reached an advanced cancer stage, and less than 30% of the patients are eligible for surgical intervention [27]. The fact that the only clinically accepted molecular assay for detection of liver cancer is serum alpha-fetoprotein levels makes it critical to find a novel method to detect early liver cancer [27].

In liver cancer, the molecular pathogenesis is tremendously complex and heterogeneous. Thus, analysis of therapeutic targets and drug resistance-conferring gene mutations from ctDNA released into the circulation contributes to a better understanding and clinical management of drug resistance in cancer patients. Studies have found that cfDNA concentration in serum or plasma of liver cancer patients is 3-4 times higher than in patients with chronic hepatitis and is up to 20 times higher than in healthy individuals [28, 29]. Ng *et al.* demonstrated that a number of somatic variants can be reliably detected in plasma but can only be found by interrogation in the biopsy counterparts, supporting the notion that genetic analysis of a single diagnostic biopsy may not be representative of the disease [30].

In a small number of liver cancer patients, ultra-deep sequencing of DNA from plasma was able to detect variants on commonly mutated genes in HCC including *TERT*, *JAK1*, *CTNNB1*, *BRAF* and *TP53* [31]. Another study in advanced liver cancer demonstrated that serial assessment of alterations in cancer-related driver genes using cfDNA NGS can reveal genomic changes with time. The concordance levels between genomic alterations found in plasma and tissue were high for the three most commonly altered genes, *TP53*, *CTNNB1*, and *ARID1A*, indicating that liquid biopsy is relatively accurate compared to traditional biopsy in monitoring of advanced liver cancer patients [32]. Cai *et al.* demonstrated that both single nucleotide variants (SNVs) and copy number variants (CNVs) of cfDNA found in plasma samples can quantifiably reflect the tumor size of liver cancer patients and may be prognostic [33]. However, even though SNVs are a more sensitive measurement than CNVs, they failed to detect tumor burden when the tumor fraction of cfDNA is relatively low [33].

III. Advances in the Detection of cfDNA

Recent research has positioned the detection of genetic alterations in the cfDNA at the frontline of cancer management. In this part of the article, we discuss contemporary NGS approaches to cfDNA analysis (Table I). Here, we explain the technical and analytical challenges to low-frequency mutation detection in the context of early cancer detection. An overview of the effects of biological noise on the detection of low frequency mutations in plasma cfDNA is provided.

TABLE I
ASSAYS DEVELOPED FOR THE EARLY DETECTION OF CANCERS

Assay	**CAPP-Seq**	**TEC-Seq**	**TRACERx**	**CancerSEEK**
Technique	Deep sequencing	targeted sequencing	multiplex PCR	multiplex PCR
Application	cfDNA	ctDNA	ctDNA	ctDNA and protein
Panel size	128 genes	58 genes	61 ROI (16 genes)	Median of 18 patient specific SNVs
Technology	NimbleGen SeqCap	Agilent SureSelect	Signatera	Safe-Seq
Year published	2014	2017	2014	2018
Citation	[34]	[35]	[36]	[37]

TABLE II
ABBREVIATIONS FOR THE TERMS USED IN TABLE I

Abbreviation	**Definition**
CAPP-Seq	Cancer Personalized Profiling by deep Sequencing
ctDNA	circulating tumour DNA
ROI	Regions Of Interest
Safe-Seq	Safe-Sequencing system
SNV	Single-Nucleotide Variant
TEC-Seq	Targeted Error Correction Sequencing
TRACERx	Tracking Cancer Evolution Through Therapy

1. CancerSEEK

The recent paper by Cohen *et al.* [37] addresses the possibility of detecting solid tumors at an early stage with a combined assay for genetic alterations and protein biomarkers (Figure 1). The approach is simple and cost-effective, potentially resulting in significant time savings for diagnosis, and improved patient survival. Although solid outcome has not yet been achieved, this manuscript advances the field of early cancer detection.

In the original study, the patient cohort included 1,005 participants diagnosed with non-metastatic forms of one of eight cancer types (ovarian, liver, stomach, pancreatic, oesophageal, colorectal, lung and breast). CancerSEEK had a median sensitivity of 70% among the eight cancer types with a specificity of 99% [38]. To achieve this, the test was developed with rigorous statistical methods to ensure the accuracy. The most important attribute to the algorithm was the presence of cfDNA mutation followed by the elevation of cancer antigens (including CA-125 and HGF) [38]. Since the driver gene mutations are usually not tissue specific, the majority of localization information was derived from protein markers. The accuracy of prediction was highest for colorectal cancers and lowest for lung cancers [37].

The study design is perhaps the weakest point of the paper. First, the study analyzed the plasma of a comparatively small number of healthy controls without matched controls for inflammatory lesions that could imitate the disease process. Second, among eight chosen cancer types, the detection capability of the test was 100% for ovarian cancer but only reasonable for other common cancers, such as breast and lung cancers, which led to another main caveat of the test. The predictive value of the test which relies on the prevalence of the disease within the tested population. The prevalence of the chosen cancers in healthy individuals over age 64 is approximately 1% [37]. Thus, even if CancerSEEK could achieve a 99% sensitivity and 99% specificity, the resulting positive predictive value would be only 50% (50% of test positives would be false positive). An established challenge in early cancer detection is that patients at increased risk for cancer may also have precancerous conditions resulting in elevation of serum protein biomarkers, a factor that was not well addressed in the healthy control population of the CancerSEEK test [38].

2. CAPP-Seq

The cancer personalized profiling by deep sequencing (CAPP-Seq) method assesses plasma cfDNA for alterations in 128 genes that are recurrently mutated in NSCLC. When being assessed for disease monitoring and minimal residual disease detection, the method achieved the sensitivity of 50% for stage-I tumors, and 100% among those with stage-II to -IV tumors, with a specificity for both groups of 96%. Notably, it was able to detect a specific *TP53* hotspot variant at a median frequency of approximately 0.18% across all plasma DNA samples. When CAPP-Seq was combined with a computational error correction approach called integrated digital error suppression (iDES), the specificity in detecting EGFR hotspot variants was 100% [34].

The main caveat of CAPP-Seq is that it used cfDNA levels to detect tumor burden since cfDNA can only predict residual tumor rather than the tissue of origin. This means cfDNA should be used in combination with another method such as imaging for disease burden. Note that CAPP-Seq is currently unable to acknowledge whether cfDNA rates from primary lesions and metastatic disease are released at the same. Potentially, the different rates in which tumors or clones release their DNA could cause problems with determining tumor burden and clonal evolution [34].

3. TEC-Seq

Targeted error correction sequencing or TEC-Seq was developed to allow the profiling of tiny amounts of ctDNA present in early-stage cancer using massively parallel sequencing. In brief, DNA fragments were isolated from a patient's blood, the researchers labeled each fragment with a unique barcode. These barcodes allowed the tracking of each fragment as they sequenced it to approximately 30,000x in depth. By cross-checking sequences at each position to each other and to a reference genome, the researchers could confirm whether a detected mutation was genuine or a false positive [35].

TEC-Seq detects 58 cancer-related genes encompassing 81 kb that are often altered in certain tumors and can detect tumor-specific mutations. They used the technique to analyze plasma samples from 44 healthy individuals and 200 individuals with breast, colorectal, lung, or ovarian cancers in various stages. TEC-Seq correctly identified 56%, 83%, 62%, and 71% of these individuals, respectively, identifying gene alterations for each case. TEC-Seq was sensitive enough to detect early-stage tumors. From a pool of 138 stage-I and -II cancers, TEC-Seq identified 86, or more than 60% of these diseases. Overall, the test was most successful at detecting colorectal and ovarian cancers [35].

The key for any early detection assay is to determine whether the mutations present in the blood are tumor-derived. Here, it was found that in 82% of cases, a mutation detected in the blood was also present in the tumor when analyzing matched tumor and plasma samples. In a further test, the researchers applied TEC-Seq to blood samples from 44 healthy individuals. They did not detect any alterations among any of the investigated 80,000 bases, leading to a very high specifivity [35].

TEC-Seq may also be useful for forecasting cancer relapse. The study applied the method to plasma samples from 38 individuals with colorectal cancer who had undergone removal. They found that patients with large abundant of ctDNA at diagnosis had shorter survival with less ctDNA [35].

IV. Future Perspectives

We have discussed technical and biological challenges associated with cfDNA detection in patients in the context of liquid biopsy. Liquid biopsy is increasingly being adopted for an extensive variety of applications in oncology. Unquestionably, using cfDNA detected from liquid biopsy for effective screening of early cancers as well as monitoring developed diseases represents the best hope for reducing cancer mortality.

The conceptual improvements and the practicability of the discussed assays are important milestones toward the application of early cancer detection. The ongoing development of cost effective and sensitive blood-based cancer screenings is set to revolutionize cancer management. Nevertheless, the use of these promising biomarkers requires a fundamental understanding of the mechanisms underlying liquid analytes and the resolution of important technology. Additional preclinical studies addressing the biology of liquid biopsy analytes are needed. Most existing assays have focused on a single analyte; however, sensitivity and accuracy could be improved by adopting multiparametric assays that integrate data from the multiple analytes present in a single sample. Innovative assay technology must be accompanied by novel statistical and machine learning tools that make use of high dimensional and large amounts of data. Thus, future liquid biopsy developments are expected to be multi-disciplinary. Critically, it is only after clinical validity and clinical utility have been demonstrated that liquid biopsies will reach their full potential and have the expected impact on clinical oncology and the clinical management of patients.

References

[1] World Health Organization, "Cancer." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Center for Disease Control and Prevention, "How to prevent cancer or find it early: screening tests." [Online]. Available: https://www.cdc.gov/cancer/dcpc/prevention/screening.htm

[3] E. Heitzer, I. S. Haque, C. E. Roberts, and M. R. Speicher, "Current and future perspectives of liquid biopsies in genomics-driven oncology," *Nature Reviews Genetics*, vol. 20, no. 2, pp. 71–88, 2019.

[4] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: A Cancer Journal for Clinicians*, vol. 66, no. 1, pp. 7–30, 2016.

[5] M. J. Moore, D. Goldstein *et al.*, "Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: A phase III trial of the National Cancer Institute of Canada Clinical Trials Group," *Journal of Clinical Oncology*, vol. 25, no. 15, pp. 1960–1966, 2007.

[6] A. Chan, I. Prassas *et al.*, "Validation of biomarkers that complement CA19. 9 in detecting early pancreatic cancer," *Clinical Cancer Research*, vol. 20, no. 22, pp. 5787–5795, 2014.

[7] T. Kamisawa, L. D. Wood, T. Itoi, and K. Takaori, "Pancreatic cancer," *The Lancet*, vol. 388, no. 10039, pp. 73–85, 2016.

[8] B. Shapiro, M. Chakrabarty, E. M. Cohn, and S. A. Leon, "Determination of circulating DNA levels in patients with benign or malignant gastrointestinal disease," *Cancer*, vol. 51, no. 11, pp. 2116–2120, 1983.

[9] G. D. Sorenson, D. M. Pribish, F. H. Valone, V. A. Memoli, D. J. Bzik, and S.-L. Yao, "Soluble normal and mutated DNA sequences from single-copy genes in human blood." *Cancer Epidemiology and Prevention Biomarkers*, vol. 3, no. 1, pp. 67–71, 1994.

[10] H. Kinugasa, K. Nouso *et al.*, "Detection of K-ras gene mutation by liquid biopsy in patients with pancreatic cancer," *Cancer*, vol. 121, no. 13, pp. 2271–2280, 2015.

[11] M. Sausen, J. Phallen *et al.*, "Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients," *Nature Communications*, vol. 6, no. 1, pp. 1–6, 2015.

[12] C. K. Lee, J. Man *et al.*, "Clinical and molecular characteristics associated with survival among patients treated with checkpoint inhibitors for advanced non–small cell lung carcinoma: A systematic review and meta-analysis," *JAMA Oncology*, vol. 4, no. 2, pp. 210–216, 2018.

[13] G. Boloker, C. Wang, and J. Zhang, "Updated statistics of lung and bronchus cancer in United States (2018)," *Journal of Thoracic Disease*, vol. 10, no. 3, pp. 1158–1161, 2018.

[14] C. Fitzmaurice, T. F. Akinyemiju *et al.*, "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: A systematic analysis for the global burden of disease study," *JAMA Oncology*, vol. 4, no. 11, pp. 1553–1568, 2018.

[15] E. F. Patz Jr, M. J. Campa, E. B. Gottlin, I. Kusmartseva, X. R. Guan, and J. E. Herndon, "Panel of serum biomarkers for the diagnosis of lung cancer," *Journal of Clinical Oncology*, vol. 25, no. 35, pp. 5578–5583, 2007.

[16] A. E. Revelo, A. Martin *et al.*, "Liquid biopsy for lung cancers: an update on recent developments," *Annals of Translational Medicine*, vol. 7, no. 15, p. 349, 2019.

[17] H. Kimura, K. Kasahara *et al.*, "Detection of epidermal growth factor receptor mutations in serum as a predictor of the response to gefitinib in patients with non–small-cell lung cancer," *Clinical Cancer Research*, vol. 12, no. 13, pp. 3915–3921, 2006.

[18] R. Rosell, T. Moran *et al.*, "Screening for epidermal growth factor receptor mutations in lung cancer," *New England Journal of Medicine*, vol. 361, no. 10, pp. 958–967, 2009.

[19] U. Malapelle, R. Sirera *et al.*, "Profile of the roche cobas® EGFR mutation test v2 for non-small cell lung cancer," *Expert Review of Molecular Diagnostics*, vol. 17, no. 3, pp. 209–215, 2017.

[20] Y. Suzuki, M. Orita, M. Shiraishi, K. Hayashi, and T. Sekiya, "Detection of ras gene mutations in human lung cancers by single-strand conformation polymorphism analysis of polymerase chain reaction products." *Oncogene*, vol. 5, no. 7, pp. 1037–1043, 1990.

[21] S.-W. Han, T.-Y. Kim *et al.*, "Optimization of patient selection for gefitinib in non–small cell lung cancer by combined analysis of epidermal growth factor receptor mutation, K-ras mutation, and Akt phosphorylation," *Clinical Cancer Research*, vol. 12, no. 8, pp. 2538–2544, 2006.

[22] C. Camps, R. Sirera *et al.*, "Is there a prognostic role of K-ras point mutations in the serum of patients with advanced non-small cell lung cancer?" *Lung Cancer*, vol. 50, no. 3, pp. 339–346, 2005.

[23] W. Pao, T. Y. Wang *et al.*, "KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib," *PLoS Medicine*, vol. 2, no. 1, 2005.

[24] S. Wang, T. An *et al.*, "Potential clinical significance of a plasma-based kras mutation analysis in patients with advanced non–small cell lung cancer," *Clinical Cancer Research*, vol. 16, no. 4, pp. 1324–1330, 2010.

[25] A. A. Chaudhuri, J. J. Chabon *et al.*, "Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling," *Cancer Discovery*, vol. 7, no. 12, pp. 1394–1403, 2017.

[26] A. Villanueva, "Hepatocellular carcinoma," *New England Journal of Medicine*, vol. 380, no. 15, pp. 1450–1462, 2019.

[27] Q. Ye, S. Ling, S. Zheng, and X. Xu, "Liquid biopsy in hepatocellular carcinoma: circulating tumor cells and circulating tumor DNA," *Mol. Cancer*, vol. 18, p. 114, 2019.

[28] Y. Tokuhisa, N. Iizuka *et al.*, "Circulating cell-free DNA as a predictive marker for distant metastasis of hepatitis C virus-related hepatocellular carcinoma," *British Journal of Cancer*, vol. 97, no. 10, pp. 1399–1403, 2007.

[29] Z. Huang, D. Hua *et al.*, "Quantitation of plasma circulating DNA using quantitative PCR for the detection of hepatocellular carcinoma," *Pathology & Oncology Research*, vol. 18, no. 2, pp. 271–276, 2012.

[30] C. Ng, G. Di Costanzo *et al.*, "Genetic profiling using plasma-derived cell-free DNA in therapy-naïve hepatocellular carcinoma patients: A pilot study," *Annals of Oncology*, vol. 29, no. 5, pp. 1286–1291, 2018.

[31] I. Labgaa, C. Villacorta-Martin *et al.*, "A pilot study of ultra-deep targeted sequencing of plasma DNA identifies driver mutations in hepatocellular carcinoma," *Oncogene*, vol. 37, no. 27, pp. 3740–3752, 2018.

[32] S. Ikeda, J. S. Lim, and R. Kurzrock, "Analysis of tissue and circulating tumor DNA by next-generation sequencing of hepatocellular carcinoma: Implications for targeted therapeutics," *Molecular Cancer Therapeutics*, vol. 17, no. 5, pp. 1114–1122, 2018.

[33] Z. Cai, G. Chen *et al.*, "Comprehensive liquid profiling of circulating tumor DNA and protein biomarkers in long-term follow-up patients with hepatocellular carcinoma," *Clinical Cancer Research*, vol. 25, no. 17, pp. 5284–5294, 2019.

[34] A. M. Newman, S. V. Bratman *et al.*, "An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage," *Nature Medicine*, vol. 20, no. 5, pp. 548–554, 2014.

[35] J. Phallen, M. Sausen *et al.*, "Direct detection of early-stage cancers using circulating tumor DNA," *Science Translational Medicine*, vol. 9, no. 403, 2017.

[36] M. Jamal-Hanjani, A. Hackshaw *et al.*, "Tracking genomic cancer evolution for precision medicine: The lung TRACERx study," *PLoS Biology*, vol. 12, no. 7, pp. e1 001 906–e1 001 906, 2014.

[37] J. D. Cohen, L. Li *et al.*, "Detection and localization of surgically resectable cancers with a multi-analyte blood test," *Science*, vol. 359, no. 6378, pp. 926–930, 2018.

[38] M. Kalinich and D. A. Haber, "Cancer detection: Seeking signals in blood," *Science*, vol. 359, no. 6378, pp. 866–867, 2018.

Nguyen Ngoc Tran was born in Ho Chi Minh City in 1989. She received the Bachelor in Pharmacy from The University of Medicine and Pharmacy in Ho Chi Minh City, Ho Chi Minh, Vietnam in 2011. She went on to receive a Ph.D. in Cancer Biology from The University of Texas at MD Anderson Cancer Center, Houston, Texas in 2018. Her thesis work focused on studying the mechanism of skin cancer development. She spent two years working at Moffitt Cancer Center, Tampa, Florida as a Research Associate and Postdoctoral Scholar. Since 2019, she has been with Vingroup Big Data Institute, Hanoi, Vietnam, where she is currently a Research Scientist. Her main area of research interest is translational research, including identifying preventative biomarkers and developing diagnostic tests for cancer. Dr. Nguyen Ngoc Tran is a member of the American Association for Cancer Research.

A Microstrip MIMO Antenna with Enhanced Isolation for WiMAX Applications

Nguyen Ngoc Lan, Nguyen Thi Thu Hang, Ho Van Cuu
Faculty of Electronics and Telecommunications, Saigon University, Vietnam
Correspondence: Nguyen Ngoc Lan, nnlan@moet.edu.vn
Communication: received 02 July 2019, revised 09 September 2019, accepted 12 September 2019
Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.869
The Editor coordinating the review of this article and deciding to accept it was Prof. Nguyen Tan Hung

Abstract: In this paper, a multiple-input multiple-output (MIMO) antenna with high isolation is designed using defected ground structure (DGS). The proposed antenna is constructed by two sets of four elements (2×2)**, which are designed at the central frequency of** 3.5 **GHz for Worldwide Interoperability for Microwave Access (WiMAX) applications. The antenna is fabricated on a FR4 substrate with an overall size of** $144 \times 99 \times 1.6$ **mm. Thanks to DGS, the designed MIMO antenna achieves a high isolation of** 30 **dB and a high radiation efficiency of over** 90%**. Besides, this MIMO antenna attains a** 7.5 **dBi gain. There is a good agreement between the simulated S-parameters and the measurement results.**

Keywords: *MIMO antenna, mutual coupling, defected ground structure (DGS), microstrip antenna.*

I. Introduction

Multiple-output multiple-input (MIMO) is one of the prominent solutions to satisfy the high data rate demand of end users in wireless networks. Employing multiple-element antennas, MIMO can improve both the spectral efficiency and reliability of the transmission without increasing transmitting power or bandwidth [1]. However, when the distance between the antenna elements is not large enough, mutual coupling happens. This is an undesired phenomenon because it not only reduces channel capacity [2], but also introduces extra power loss to the system [3].

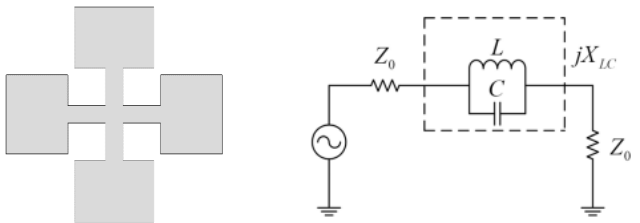
Many solutions have been proposed to reduce mutual coupling between antenna elements using, *e.g.*, shorting pins [4], compact coplanar waveguide (CPW) feeding [5], electromagnetic band gap (EBG) [6], parallel coupled-line resonators [7]. These methods have reached a recognizable improvement in isolation enhancement. The isolation between the antenna elements in [4–6] is around 20 dB. The antenna gains are under 2 dBi in [5] and [7]. In addition, the radiation efficiency of the antenna in [5] is 70%. However, these figures can be further improved. In fact, it is challenging to simultaneously optimize multiple parameters, such as the isolation and radiation efficiency, in designing MIMO antennas.

In this paper, we propose a MIMO antenna with enhanced isolation. The antenna is designed at the central frequency 3.5 GHz. We apply a defected ground structure (DGS) to achieve a high isolation between the antenna elements of over 35 dB although the edge-to-edge distance is only 4.3 mm. Based on a FR4 substrate with a thickness of 1.6 mm, $\varepsilon_r = 4.4$, and $\tan\delta = 0.02$, the antenna has dimension $144 \times 99 \times 1.6$ mm. At the central frequency 3.5 GHz, the MIMO antenna reaches over 7.5 dBi gain. The antenna is designed, simulated, and optimized with the Computer Simulation Technology (CST) Studio software. The simulation results are compared with the measurement ones to verify the performance of the proposed antenna.

II. MIMO Antenna Design

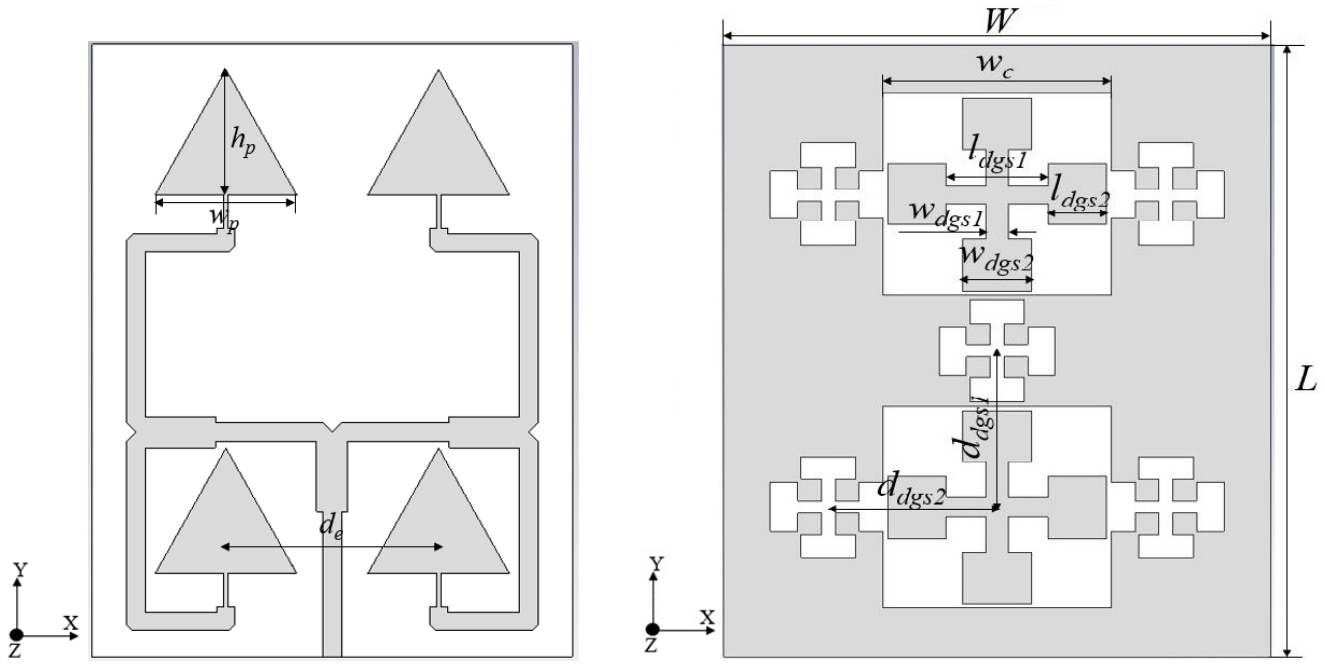
1. Design of the Antenna Array

Our antenna array design is based on the DGS. Figure 1 shows the model of DGS and its equivalent circuit. The DGS is created by connecting two rectangular areas and a microstrip line. The equivalent circuit is made based on the following principle: two rectangles can be considered as an inductance while a microstrip line corresponds to



(a) Model (b) Equivalent circuit

Figure 1. Defected ground structure (DGS) [9].



(a) Top view (b) Bottom view

Figure 2. Configuration of proposed DGS-based antenna array.

a capacitance. Therefore, the capacitance and inductance values are given by [8, 9]

$$C = \frac{\omega_c}{2z_0\left(\omega_0^2 - \omega_c^2\right)}, \tag{1}$$

$$L = \frac{1}{4\pi^2 f_0^2 C}, \tag{2}$$

where Z_0 is the input and output port impedance; ω_0 is the angular frequency of the parallel LC resonator, f_0 is the resonant frequency, and ω_c is the cutoff angular frequency.

Then, the resonant frequency can be calculated as

$$f_r = \frac{1}{2\pi\sqrt{LC}}. \tag{3}$$

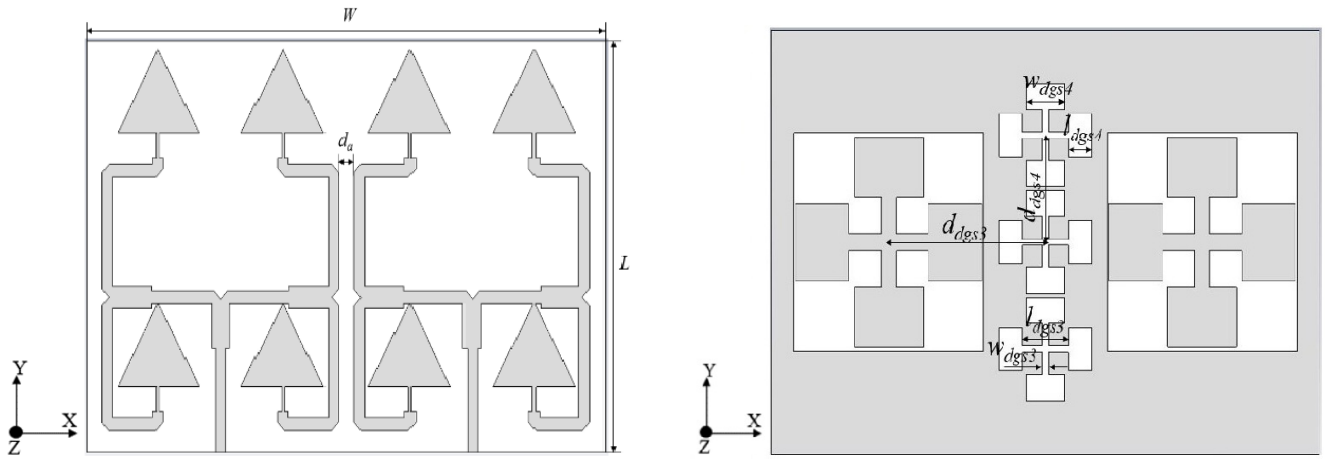
Next, Figure 2 shows the model of our proposed DGS-based antenna array. The antenna is designed to operate at 3.5 GHz and it is printed on a single layer FR4 substrate with a dielectric constant of 4.4, a thickness of 1.6 mm, and a loss tangent of 0.02. The antenna array includes four radiation elements (2×2) and three power dividers on top of the substrate. The ground plane with DGS is placed on the bottom. Using the formula in [10], we can calculate the size of each radiation element. Furthermore, we select equal dividers whose principle can be found in [11]. We use the CST Studio software to optimize and obtain that the dimension of each element is 23.6×20.8 mm and the distance between elements is approximately 0.4λ where λ is the wavelength in free space. The overall size of the antenna array is 80×102 mm. Table I lists the values of some dimension parameters of the designed antenna array. These parameters are defined in Figure 2.

TABLE I
DIMENSION PARAMETERS (DEFINED IN FIGURE 2) OF PROPOSED ANTENNA ARRAY

Parameter	**Value** (mm)	**Parameter**	**Value** (mm)
W	80	L	102
w_p	23.6	l_p	20.8
l_{dgs1}	15	w_{dgs1}	3.2
l_{dgs2}	8.5	w_{dgs2}	10.2
d_e	35.4	d_{dgs1}	26
w_c	33.5	d_{dgs2}	24.7

2. Design of the MIMO Antenna

The MIMO antenna consists of two symmetric antenna arrays placed side by side with a separation of 4.3 mm from edge to edge as illustrated in Figure 3. There are many feeding techniques for antenna such as coaxial feed, coupled, and $\lambda/4$ impedance transformer [10, 12]. We choose the $\lambda/4$ impedance transformer in this paper due to its simplicity in impedance matching. Based on a FR4 substrate with a thickness of 1.6 mm, the MIMO antenna has overall size of $144 \times 99 \times 1.6$ mm while the size of one patch is 23×20.125 mm.



(a) Top view (b) Bottom view

Figure 3. Configuration of proposed MIMO antenna.

TABLE II
DIMENSION PARAMETERS (DEFINED IN FIGURE 3) OF PROPOSED MIMO ANTENNA

Parameter	**Value** (mm)	**Parameter**	**Value** (mm)
W	144	L	99
d_a	4.3	d_{dgs3}	41.14
d_{dgs4}	25	w_{dgs4}	10
l_{dgs4}	6	w_{dgs3}	1.5
l_{dgs3}	12		

Table II lists the value of some parameters of the proposed MIMO antenna. These parameters are defined in Figure 3. To reduce the mutual coupling between the elements of the MIMO antenna, we integrate a DGS on the ground plane. The DGS consists of five cells (two large cells and three small cells). To make the capacitance (C) and inductance (L) variable, we use a compensation structure as shown in Figures 1 and 2. This makes a flexibility in optimization.

III. RESULTS AND DISCUSSIONS

1. Simulation Results

a) Antenna Array

Figure 4 presents the reflection coefficients of the proposed antenna array over the frequency band from 2.5 GHz to 4.5 GHz. As can be seen, the bandwidth of the antenna is 330 MHz, corresponding to 9.4% of the resonant frequency 3.5 GHz. In addition, the antenna has a low return loss of −30 dB at the resonant frequency.

Figure 5 shows the 3D and polar patterns of the proposed antenna array. It can be observed that the the antenna has quite high directivity: it has an angular width (3 dB) of 40.7 degrees. Moreover, the antenna remains a high radiation efficiency of over 90%.

b) MIMO Array

As mentioned in Section II.2, the MIMO antenna is integrated with DGS to reduce the effect of mutual coupling. To better understand the effect of DGS on mutual coupling reduction, we compare the S-parameters of the proposed MIMO antenna with and without DGS. The results are displayed in Figure 6. It can be observed that the isolation between antenna elements is significantly improved in the case of DGS. The mutual coupling is −15 dB without DGS while with DGS, this figure is less than −40 dB. This can be explained as follows. Normally, the current distribution in microstrip antenna is uniform without DGS. When there is DGS, the current is redistributed according to the size and position of DGS. By adjusting the size and position of DGS, we can distribute the current at a desired place while limiting the current at other places. This helps us to achieve a high isolation of 40 dB for the proposed antenna. On top of that, using DGS also improves the bandwidth. We can see that the antenna bandwidth with DGS includes two resonant modes while this value is only one without DGS. As a result, the bandwidth with DGS is larger than without DGS.

Figure 7 shows the pattern of the proposed MIMO antenna. It can be seen that the main lobe direction of the antenna is 190 degrees while the angular width at 3 dB is 40.7 degrees. Normally, the main lobe direction of a microstrip antenna is 0 degree (uniform current distribution). In our case, utilizing DGS causes a distortion in the current distribution, thus the main lobe direction can be changed. This is a common tradeoff when using DGS. However, given the gain in antenna isolation, this main

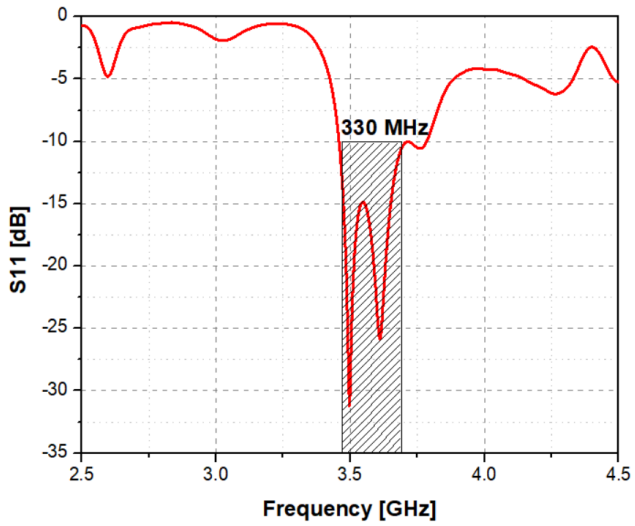


Figure 4. Reflection coefficients of proposed antenna array.

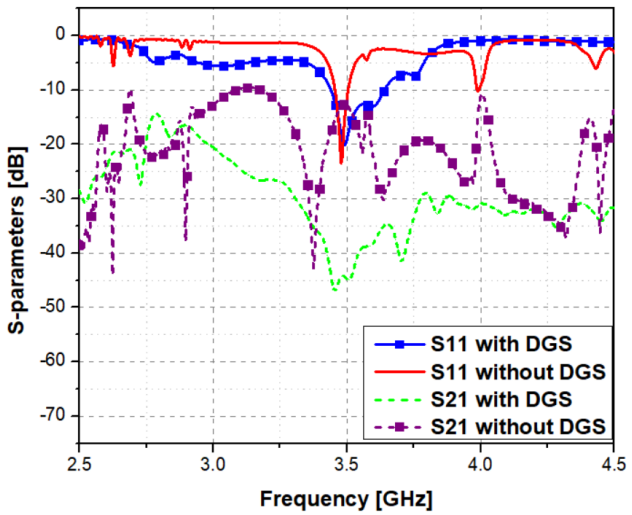
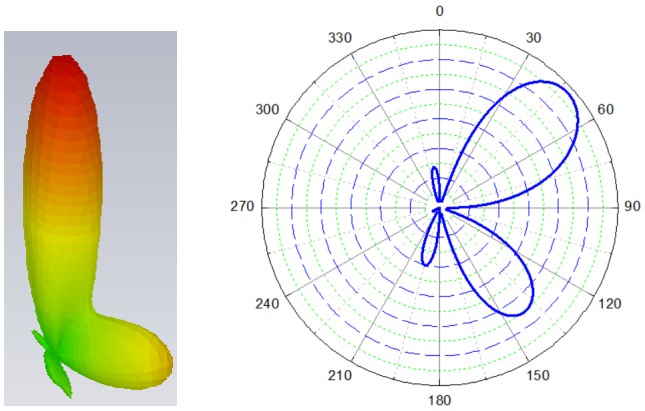
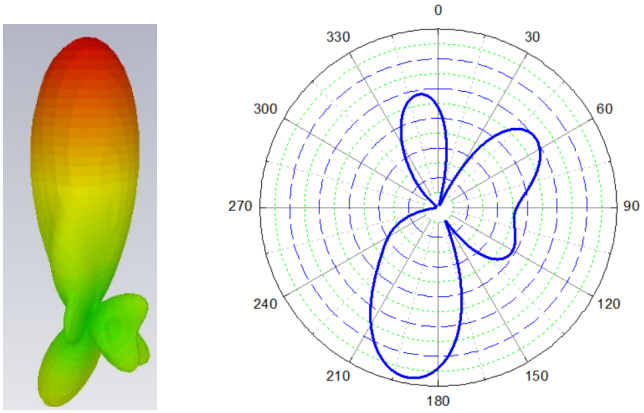


Figure 6. S-parameters of proposed MIMO antenna.



(a) 3D (b) Polar

Figure 5. Pattern of proposed antenna array.



(a) 3D (b) Polar

Figure 7. Pattern of proposed MIMO antenna.

lobe direction change is acceptable. Besides, the gain and radiation of the MIMO antenna reach more than 7.5 dBi and 90%, respectively.

Moreover, in MIMO systems, the independence between radiation patterns of the antennas can be evaluated by the enveloped correlation coefficient (ECC), denoted by ρ_e. ECC can be calculated from the S-parameters as [13]

$$\rho_e = \frac{S_{11}S_{12}^* + S_{21}S_{22}^*}{\sqrt{\left(1-|S_{11}|^2-|S_{21}|^2\right)\left(1-|S_{22}|^2-|S_{12}|^2\right)\eta_1\eta_2}}. \quad (4)$$

It can also be calculated from the radiation patterns as [14]

$$\rho_e = \frac{\iint_{4\pi} E_1(\theta,\phi)E_2^*(\theta,\phi)d\Omega}{\iint_{4\pi} E_1(\theta,\phi)E_1^*(\theta,\phi)d\Omega \iint_{4\pi} E_2(\theta,\phi)E_2^*(\theta,\phi)d\Omega}, \quad (5)$$

where E_1 and E_2 are the far-field radiation patterns, generated from ports 1 and 2 of the antenna while θ and ϕ represents the spherical angles, namely, the elevation and azimuth, respectively. Figure 8 presents the ECC of the proposed MIMO antenna. It is clear that the ECC is smaller than 0.01 from 3.08 GHz to 3.84 GHz, corresponding to a band of 760 MHz. This is suitable for devices with $\rho_e \leq 0.3$ [15].

2. Measurement Results

In order to confirm the simulation results by experimental measurements, the prototype of the proposed antenna as shown in Figure 9 is implemented based on a FR4 sheet (h = 1.6 mm, ε_r = 4.4 and $\tan\delta$ = 0.02) with a size of 80×102×1.6 mm for a single array and 144×99×1.6 mm for the MIMO antenna.

In Figure 10, we compare the simulated results based on the CST Microwave software and the measurement ones. As can be seen, the measured impedance bandwidths at −10 dB of a single array and the MIMO antenna

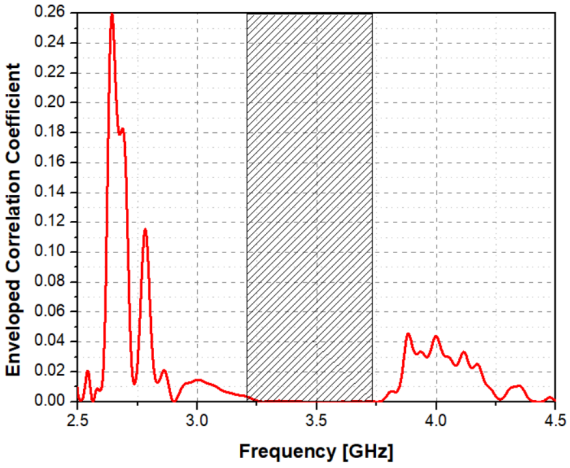


Figure 8. Enveloped correlation coefficient of proposed MIMO antenna.

TABLE III
PERFORMANCE COMPARISON OF PROPOSED ANTENNA AND RECENT ANTENNAS

References	[17]	[18]	[19]	This paper
Frequency [GHz]	2.4	3.5	2.6	**3.5**
Bandwidth [%]	3.3	6.2	14.6	**4.15**
Isolation [dB]	15	22	15	**29**
Efficiency [%]	85	87	x	**90**
Gain [dBi]	x	3.1	4.48	**7.5**

are 120 MHz (3.43%) and 145 MHz (4.15%), respectively, while the corresponding figures from simulation are 330 MHz and 200 MHz, respectively. Moreover, the measured isolation between elements in the MIMO antenna is approximately 30 dB. There is a tolerance between the simulation and measurement results. This can be caused by the instability of parameters in the FR4 substrate [16] and the tolerance in fabrication. However, this tolerance does not affect the operation of antenna and is thus acceptable. To verify the advantage of our proposed antenna, we compare its performance with some previously proposed antenna in the literature in Table III. We can see that the isolation of the antenna in [17] is only 15 dB and the percentage of the impedance bandwidth is not large (3.3%). In [18], although there is a high isolation (22 dB) between array elements, the gain of the antenna is quite low (3.1 dBi). The MIMO antenna in [19] has a large percentage of the impedance bandwidth, but achieves a low gain (4.48 dBi) while the efficiency is not mentioned. In this paper, by using DGS, the proposed antenna achieves a high isolation of approximately 30 dB, and, at the same time, a high efficiency of over 90%. Therefore, despite of the main lobe shift (190 degrees) and a higher complexity when using DGS, our antenna is a promising solution to operate at 3.5 GHz.

IV. CONCLUSION

This paper presents a MIMO antenna with enhanced isolation for WiMAX applications. The antenna is realized at the central frequency of 3.5 GHz and it is built on a FR4 substrate with parameters: $h = 1.6$ mm, $\varepsilon_r = 4.4$, and $\tan\delta = 0.02$. The proposed MIMO antenna contains two sets of 2×2 elements which incorporates DGS to obtain a high isolation between the elements. From measurements, the antenna achieves approximately 30 dB in isolation and 90% in radiation efficiency. Moreover, the antenna has a gain of 7.5 dBi while the measured bandwidth is 145 MHz at -10 dB. The proposed antenna has a compact size, a planar structure, an easy fabrication, and a low cost, thus can be utilized in practice.

ACKNOWLEDGMENT

This work is carried out in the framework of the project entitled "Research on methods for mutual coupling reduction between array antenna elements in multi-antenna wireless communication system" under Grant number CS2019-38.

REFERENCES

[1] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bölcskei, 2004, "An overview of MIMO communications - A key to gigabit wireless," in *Proceedings of the IEEE*, vol. 92, no. 2, pp. 198–217, Feb. 2004.

[2] B. Clerckx, D. Vanhoenacker-Janvier, C. Oestges, and L. Vandendorpe, "Mutual coupling effects on the channel capacity and the space-time processing of MIMO communication systems," in *Proc. IEEE International Conference on Communications*, Anchorage, AK, USA, June 2003, pp. 2638–2642,.

[3] J. P. Linnartz, "Effects of Antenna Mutual Coupling on the Performance of MIMO Systems," in *Proc. 29th Symposium on Information Theory in the Benelux*, no. May, 2008.

[4] M. Abdullah, Q. Li, W. Xue, G. Peng, Y. He, and X. Chen, 2019, "Isolation enhancement of MIMO antennas using shorting pins," *Journal of Electromagnetic Waves and Applications*, vol. 33, no. 10, pp. 1249-1263, 2019.

[5] Duong Thi Thanh Tu, Nguyen Tuan Ngoc, and Vu Van Yem, "Compact Wide-Band and Low Mutual Coupling MIMO Metamaterial Antenna using CPW Feeding for LTE/Wimax Applications," *Research and Development on Information and Communication Technology*, vol. 2, no. 15, 2018.

[6] N. Kumar and U. Kiran Kommuri, "MIMO Antenna Mutual Coupling Reduction for WLAN Using Spiro Meander Line UC-EBG," *Progress In Electromagnetics Research C*, vol. 80, Dec. 2018, pp. 65–77.

[7] K. S. Vishvaksenan, K. Mithra, R. Kalaiarasan, and K. S. Raj, "Mutual coupling reduction in microstrip patch antenna arrays using parallel coupled-line resonators," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 2146–2149, May 2017.

[8] L. H. Weng, Y. C. Guo, X. W. Shi, and X. Q. Chen, "An Overview on Defected Ground Structure," *Progress In Electromagnetics Research B*, vol. 7, pp. 173–189, 2008.

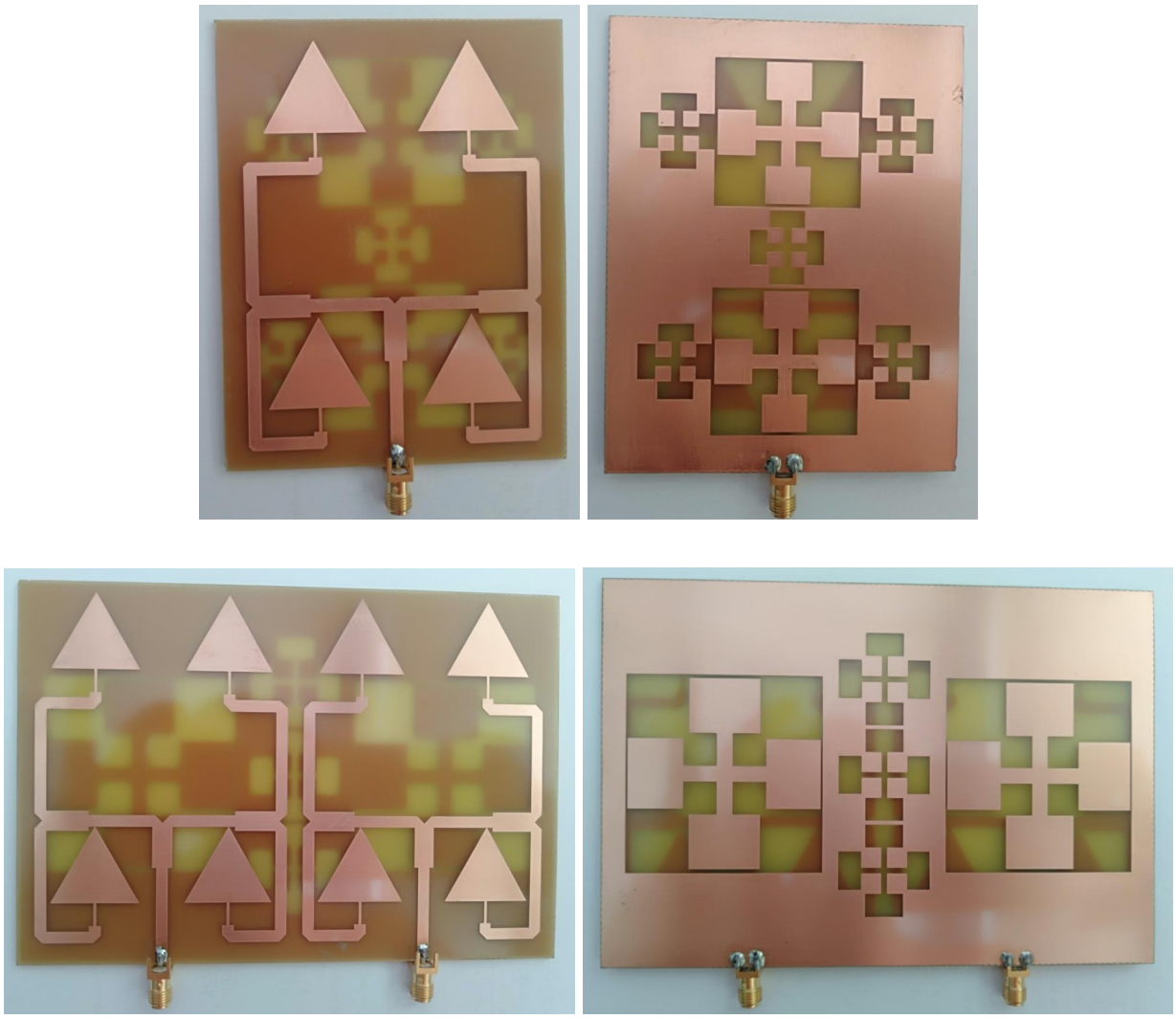
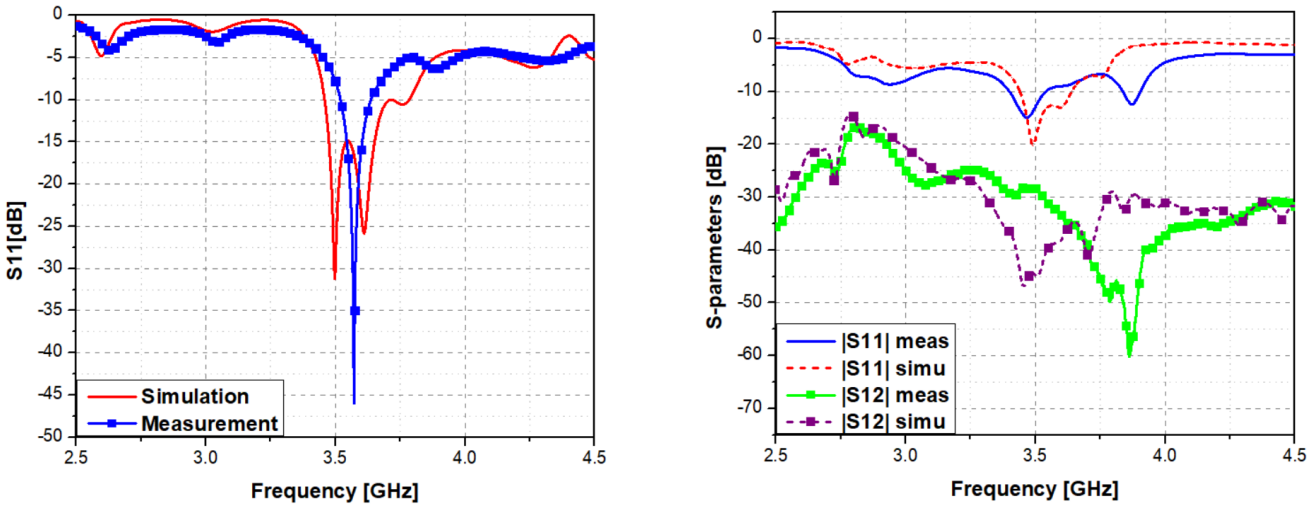


Figure 9. Prototype of fabricated single array (top) and MIMO (bottom) antennas



(a) Single array

(b) MIMO antenna

Figure 10. The measured and simulated S-parameter results of the proposed antenna.

[9] T. I. Dal Ahn, Jun-Seok Park, Chul-Soo Kim, Juno Kim, and Yongxi Qian, "A design of the novel lowpass filter using defected ground structure," *IEEE Transactions on Microwave Theory and Techniques*, vol. 49, no. 1, pp. 86–93, 2001.

[10] Constantine A. Balanis, *Antenna Theory: Analysis and Design*. United States of America: John Wiley & Sons, Inc, 2016.

[11] N. L. Nguyen and V. Y. Vu, "Gain enhancement for MIMO antenna using metamaterial structure," *International Journal of Microwave and Wireless Technologies*, pp. 1–12, no. May, 2019.

[12] A. I. Ramesh Garg, Prakash Bhartia, and Inder Bahl, *Microstrip Antenna Design Handbook*. Norwood, MA: Artech House, Inc, 2001.

[13] I. C. S. Blanch and J.Romey, "Exact representation of antenna system diversity performance from input parameter description," *Electronics Letters*, vol. 39, no. 9, pp. 705–707, 2003.

[14] R. Cornelius, A. Narbudowicz, M. J. Ammann, and D. Heberling, "Calculating the envelope correlation coefficient directly from spherical modes spectrum," in *Proc. 11th European Conference on Antennas and Propagation*, EUCAP 2017, no. 3, 2017, pp. 3003–3006.

[15] V. 3. 3GPP TS 36.101, *EUTRA User Equipment Radio Transmission and Reception*. 2008.

[16] L. Peng and C. L. Ruan, "UWB band-notched monopole antenna design using electromagnetic-bandgap structures," *IEEE Transactions on Microwave Theory and Techniques*, vol. 59, no. 4, PART 2, pp. 1074–1081, 2011.

[17] J. Deng, J. Li, L. Zhao, and L. Guo, "A Dual-Band Inverted-F MIMO Antenna with Enhanced Isolation for WLAN Applications," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 2270–2273, 2017.

[18] R. Karimian and H. Tadayon, "Multiband MIMO Antenna System with Parasitic Elements for WLAN and WiMAX Application," *International Journal of Antennas and Propagation*, vol. 2013, pp. 1–7, 2013.

[19] L. Yang, J. Fang, and T. Li, "Compact dual-band MIMO antenna system for mobile handset application," *IEICE Transactions on Communications*, vol. E98B, no. 12, pp. 2463–2469, 2015.

Nguyen Ngoc Lan received the Master and Ph.D. degrees in School of Electronics and Telecommunications, Hanoi University of Science and Technology, Vietnam, in 2014 and 2019, respectively. Currently, she is a lecturer at the Faculty of Electronics and Telecommunications, Saigon University, Vietnam. Her research interests include microstrip antenna, mutual coupling, MIMO antennas, array antennas, reconfigurable antennas, polarization antennas, metamaterial, and metasurface.

Nguyen Thi Thu Hang received the Bachelor and Master degrees from the Ho Chi Minh City University of Technology, Vietnam, in 1999 and 2002, respectively. Currently, she is a lecturer at the Faculty of Electronics and Telecommunications, Saigon University, Vietnam. Her research interests include digital signal processing, FPGA, and integrated circuit design.

Ho Van Cuu received the Ph.D. degree from the Department Telecommunication, Posts and Telecommunications Institute of Technology, Vietnam, in 2006. Currently, he is a lecturer at the Faculty of Electronics and Telecommunications, Saigon University, Vietnam. His research interests include wireless communications and digital communications.

Three Phase Resolution Transmitarray Element for Electronically Reconfigurable Transmitarrays

Nguyen Minh Thien, Nguyen Huu Minh, Nguyen Binh Duong
International University, Vietnam National University Ho Chi Minh City, Vietnam

Correspondence: Nguyen Binh Duong, nbduong@hcmiu.edu.vn
Communication: received 27 November 2019, revised 30 December 2019, accepted 30 December 2019
Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.905
The Editor coordinating the review of this article and deciding to accept it was Prof. Nguyen Tan Hung

Abstract: Electrical beam scanning is a feature enabling an antenna array to electrically control its main beam toward a desired direction. In this paper, a three-phase state element for electronically reconfigurable transmitarrays is presented. The element is made up of C-patches and modified ring slots loaded rectangular gaps. By controlling the bias state of four p-i-n diodes, three phase states are obtained. The dimension of the element is optimized by using full-wave EM simulation and performance of the element is validated by both simulation and an experimental waveguide system. A transmitarray consisting of 12×12 elements has been simulated to validate the steering capabilities. Experimental results indicate the element has good characteristics and excellent phase change capabilities.

Keywords: *Transmitarray antenna, flat lens antenna, reconfigurable transmitarray antenna.*

I. Introduction

Planar array antennas have been widely used in many applications requiring low profile antenna with high gain and beam control capacity. Examples of radar applications working at X-band are marine rain map radar, airborne weather surveillance and warning system and Synthetic Aperture Radar (SAR) for satellites. Recently, modern tunable electronic components enable high-speed beam forming and vibration elimination for planar array antennas. This is a considerable advantage over the traditional high gain antennas in which a bulky mechanical system is required to physically rotate the antenna direction. Phased array, reflectarray and transmitarray are three types of antennas that can provide high gain. A planar phased array typically uses a microstrip feeding network to excite the array elements in a specified state to form the main beam. On the other hand, reflectarray and transmitarray utilize

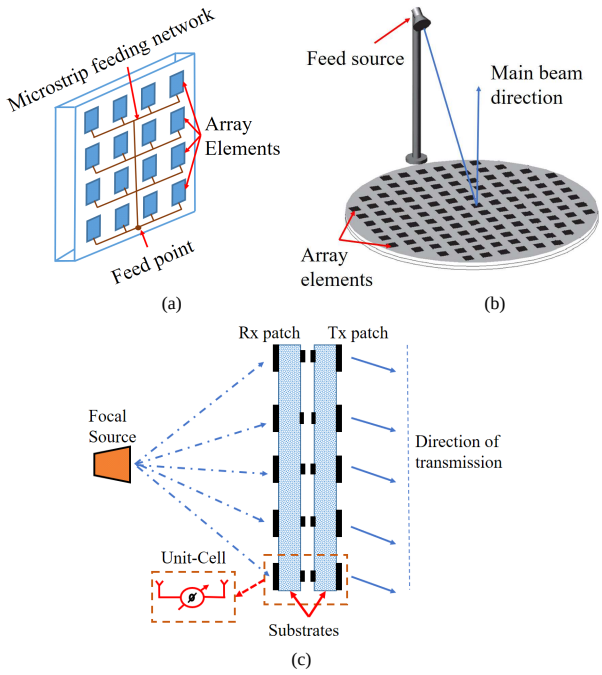


Figure 1. Basic diagram of (a) planar phased array, (b) reflectarray, and (c) transmitarray.

the space-feed mechanism in which an external feed source is placed at a focal point, thus no complex beamforming network is required in the array's surface. A transmitarray, as shown in Figure 1(c), overcomes the disadvantage of reflectarray which is the feed blockage phenomenon due to the fact that the feed source of transmitarray is located in the opposite side of the radiated wave. Recent transmitarray designs are mainly based on multilayer frequency selective surfaces (M-FSS) [1–5]. A typical transmitarray using M-FSS structure composes of multiple microstrip arrays of

antenna elements and a feed source. The first planar layer that is placed in front of the feed source is labeled as receiver (Rx), working in receiving mode. The last layer acts as transmitter (Tx).

A transmitarray transforms the spherical wave from the feed source into a planar wavefront. In order to collimate the energy toward a specific direction, each element is designed to provide a desired phase shift to compensate the phase error due to different path lengths from the feed source. Hence, the performance of the element is the utmost factor in any transmitarray design. The transmitarray element must be able to not only generate desired phase shift but also to achieve minimum energy loss.

To achieve electronically beam-scanning ability, recent works have proposed different designs for transmitarrays in which different tunable components are employed to control the main beam direction. The transmitarrays proposed in [6–8] use varactors. The advantage of a transmitarray element using varactors is the continuous phase control. This is based on the fact that the capacitance of varactors is varied as a function of the DC bias voltage. Different capacitance values could lead to different transmission phase states of an element. However, to achieve enough phase range for implementing a transmitarray, transmitarray elements proposed in [6–8] use a large number of varactors that leads to high losses and increases the complexity of the element geometry.

Besides the electronically reconfigurable transmitarray based on varactors, various transmitarrays using RF switches such as RF-MEM [9], p-i-n diodes [10–13] have been proposed. Although RF switches cannot provide a continuous phase variation, they are much more suitable for transmitarrays operating at high frequencies, especially in millimeter-wave frequency band. Moreover, the RF switches are not much sensitive to DC voltage, this makes the reconfigurable transmitarrays using RF-switches more stable against the fluctuation of DC bias voltage. In [10–12], an 1-bit reconfigurable transmitarray element using two p-i-n diodes and in [13], a 2-bit reconfigurable transmitarray element have been proposed. The disadvantage of the reconfigurable transmitarray using RF switches is the low phase resolution. A low phase resolution leads to large quantization loss, resulting in gain degradation of the transmitarray. To increase phase resolution, higher number of switches is required. However, this may increase the complexity of the biasing network. In design of an electronically reconfigurable transmitarray using RF-switches, there is a trade-off between the number of phase states and the complexity of the transmitarray element.

In this paper, a reconfigurable transmitarray with three-phase states is presented. The three phase states are obtained by using four p-i-n diodes. The proposed element is an extended version of our previous work [12] for the purpose of increasing the number of phase states of the transmitarray element. Four additional p-i-n diodes are biased in pairs, which makes the biasing network simple. The three-phase state element is designed to operate at the center frequency of 11.5 GHz. A passive prototype with the use of an ideal switch has been fabricated. Measurements on the prototype have been conducted to demonstrate the performances of the proposed element. A transmitarray of 12 × 12 elements was designed and simulated to validate the performance and the capability of main beam reconfigurability.

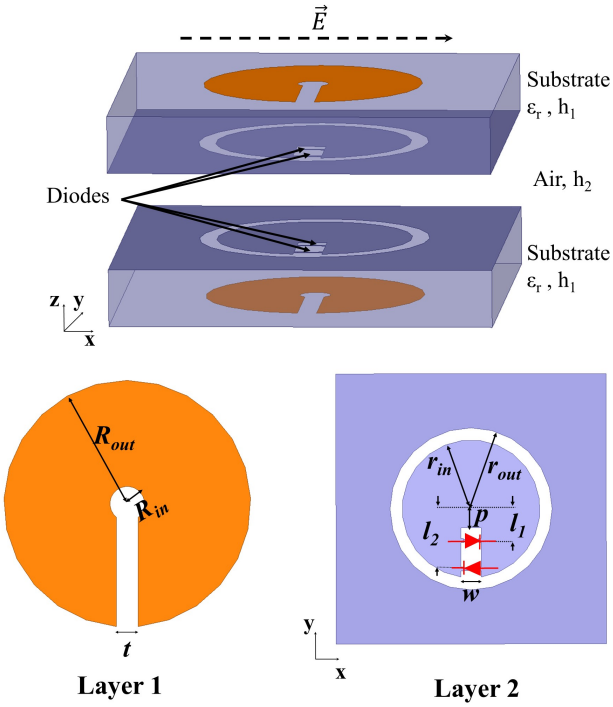


Figure 2. Geometry of the proposed transmitarray element.

II. Transmitarray Element Design

1. Element Design

Figure 2 shows the proposed transmitarray element that is implemented using four metallic layers printed on two identical Roger RO5870 substrates. The Roger RO5870 substrate has a thickness of 1.575 mm and dielectric constant ε_r = 2.33. Two substrates are separated by 1.5 mm. On each substrate, a C-patch and a modified ring slot loaded by a rectangular gap are printed. The C-patch is on the top of the substrate and the modified ring slot loaded by a rectangular gap is on the bottom of the substrate. Two p-i-n diodes are inserted on the rectangular gap of each ring slot. The p-i-n diodes are used to control the transmission

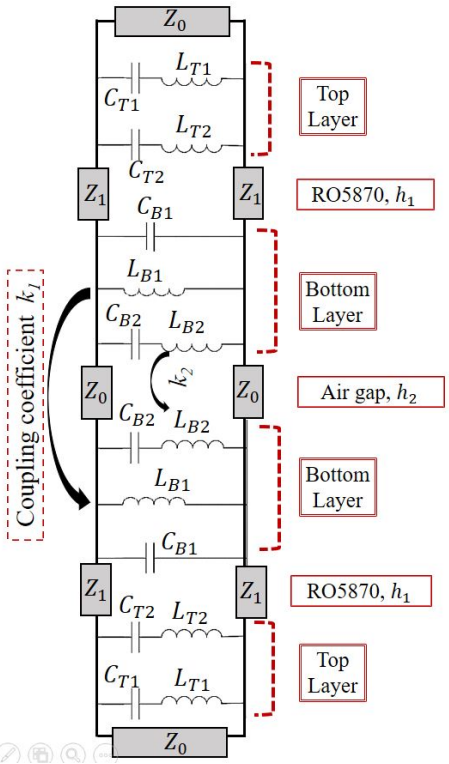


Figure 3. Equivalent circuit of our proposed structure.

phase shift. The element is designed to operate at X band with the center frequency of 11.5 GHz. It operates with linear polarization where E-field of the incident wave is perpendicular to the gap of the C-patch. The cell size is 14 mm, equivalent to $0.5\lambda_o$ where λ_o is the wavelength in the free space.

The proposed element is developed from the transmitarray element that was investigated in [12]. In this work, two more diodes are added to provide the third phase state. Increasing the phase states is based on the ability of changing the transmission phase by modifying the length of the rectangular gap of the ring slot. Therefore, we can increase the number of phase states by increasing the number of switches on the rectangular gap to control the length of the rectangular gap. To understand the working principle of the proposed structure as well as the capability of changing the phase according the length of the rectangular gap of the ring slot, it is helpful to analyze the equivalent circuit of the structure. As presented in [14], the equivalent circuit of the element can be represented as shown in Figure 3.

For the equivalent circuit, the two C-patches placed on the top of two substrates are modeled as a parallel circuit containing two series LC tanks (C_{T1}, L_{T1}, C_{T2}, L_{T2}). The ring slot loaded with a rectangular gap is represented by a parallel LC tank (C_{B1}, C_{B1}) placed in parallel with a series LC tank (C_{B2}, L_{B2}). The substrate with a thickness

TABLE I
DIMENSIONS OF THE PROPOSED ELEMENT

Element	Dimensions (mm)
Layer 1	$R_{\text{in}} = 0.8$; $R_{\text{out}} = 5.7$; $t = 1$.
Layer 2	$R_{\text{in}} = 5.1$; $R_{\text{out}} = 6$; $w = 1.5$; $p = 1.4$; $l_1 = 2.5$; $l_2 = 4.3$.
Substrates	$h_1 = 1.575$
Air gap	$h_2 = 1.5$

of h_1 is modeled as a transmission line with a length of h_1 and a characteristic impedance of $Z_1 = Z_0/\sqrt{\varepsilon_r}$, where ε_r is the relative permittivity of the substrate and $Z_0 = 377\Omega$ is the free space impedance. As shown in Figure 3, the element structure has multi-resonances. For a transmitarray element structure, multi-resonances can allow a large phase range [1, 2]. The modification of a ring slot to become the ring slot loaded with a rectangular gap makes the structure having two resonances. The original ring slot creates a single resonance that can be modelled as a parallel LC tank (C_{B1}, L_{B1}). Adding a rectangular gap creates the second resonance at a high frequency. It is modeled by a series LC tank (C_{B2}, L_{B2}). The frequency of the second resonance is a function of the length of the rectangular gap. The second resonant frequency is shifted towards to the lower frequency when the length of the rectangular gap is increased. Variation of the second resonance of the ring slot leads to a variation of the transmission phase of our structure. Therefore, three phase states can be achieved by using four p-i-n diodes. These four p-i-n diodes are placed on the rectangular gaps to control the length of the rectangular gaps. We use the same biasing circuit as presented in [12] to supply the DC power to p-i-n diodes. Four diodes work in pairs. Each pair is biased in reverse compared to the other. The position of four diodes is optimized to obtain three phase states with a step of 120° at 11.5 GHz according to ON/OFF state of diodes. The first phase state is obtained when all diodes are OFF. The second phase state and the third phase state are obtained when the first pair and second pair of diodes are turned ON, respectively. The parameters of the element are shown in Table I.

2. Frequency Response of The Transmitarray Element

The performance of the element should be validated before implementing a transmitarray. ANSYS HFSS software version 13 is used to simulate and to optimize the proposed element. To obtain the transmission phase and magnitude, a method is to use the waveguide simulator. Since the center

frequency of the element is 11.5 GHz, the WR-90 standard waveguide is suitable to be used as waveguide simulator. In the simulation, the element is placed in the open-end of two WR-90 standard waveguides. Two excitation ports are assigned at the other ends of two waveguides to measure the transmission coefficients. Before the final version of an electronically reconfigurable transmitarray is implemented, the performance of the element is first evaluated using ideal RF-switches. In this case, metallic strips are used as ideal p-i-n diodes. For the ON state of a diode, the metallic strips are inserted on the gaps. For the OFF state, the metallic strips are removed.

Figure 4 shows the simulated transmission coefficients of the proposed element for three phase states. As it can be seen, the transmission magnitude of three phase states at 11.5 GHz is greater than −1 dB. The common −3 dB transmission bandwidth of three phase states is 16.5% from 10.6 GHz to 12.5 GHz. As shown in Figure 4(b), the transmission phase curve successfully changes when we change the state of four diodes. Three phase curves have a step of 120° at 11.5 GHz. However, the step of 120° is not maintained for frequencies far from 11.5 GHz, due to the non-linearity of the phase curves.

III. Experimental Validation of the Element

A prototype of the element is implemented to validate the performance of the proposed element. The element is fabricated by standard PCB fabrication technique. A small metallic strip that acts as an ideal switch in the ON state is soldered across the rectangular gap of ring slot layer. That metallic strip is removed for the OFF state of the switch, as shown in Figure 5.

The method to measure the frequency response of the element prototype is also to use waveguide simulators. This technique requires two WR-90 standard waveguides whose open-end size is 22.86×10.16 mm^2. Since the element's shape is a square while the aperture of the waveguide is rectangular, two rectangular-to-square transitions are implemented and they are used as an adaptor to put the element in the middle of two waveguides. A metallic plate with a hollow of $14\times14\times1.5$ mm^3 is inserted between two parts of the element to ensure that two substrates are separated by an air gap of 1.5 mm. Figure 6 presents the measurement system. The measurement of the transmission coefficients is performed using Agilent E5071C Vector Network Analyzer. The measurement system has been calibrated at the ends of two straight waveguides, not including the two rectangular-to-square transitions. Figure 7 shows the measured transmission coefficients in comparison with that of the simulation. As shown in Figure 7, the measured results agree well with simulated results.

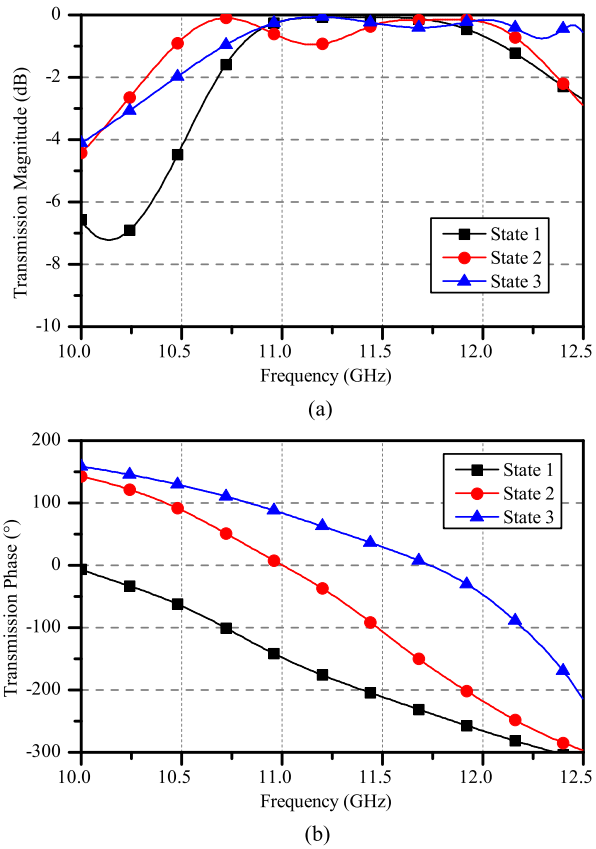


Figure 4. (a) Simulated transmission magnitude and (b) transmission phase of the proposed element.

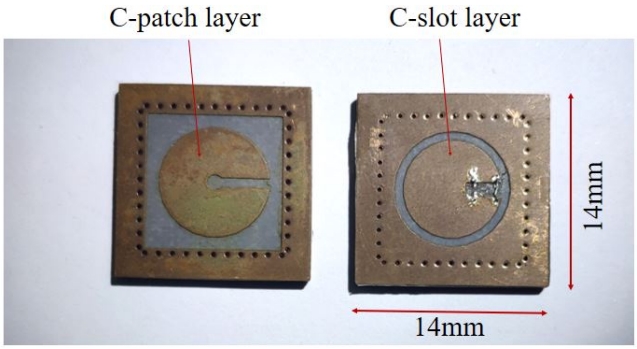


Figure 5. (Simulated transmission magnitude (left) and transmission phase (right) of the proposed element.

IV. Transmitarray Design

A square transmitarray antenna is designed with 12×12 elements to validate the radiation characteristics and beam steering capacity. As the periodicity of each element is 14 mm, the transmitarray size is 168×168 mm^2, corresponding to $6.44\lambda_o \times 6.4\lambda_o$ at 11.5 GHz. A small aperture horn antenna is used as the feed source for the array. Its aperture is 32×23 mm^2 and its directivity is 11 dB. The horn antenna is placed at a focal length of 150 mm corresponding

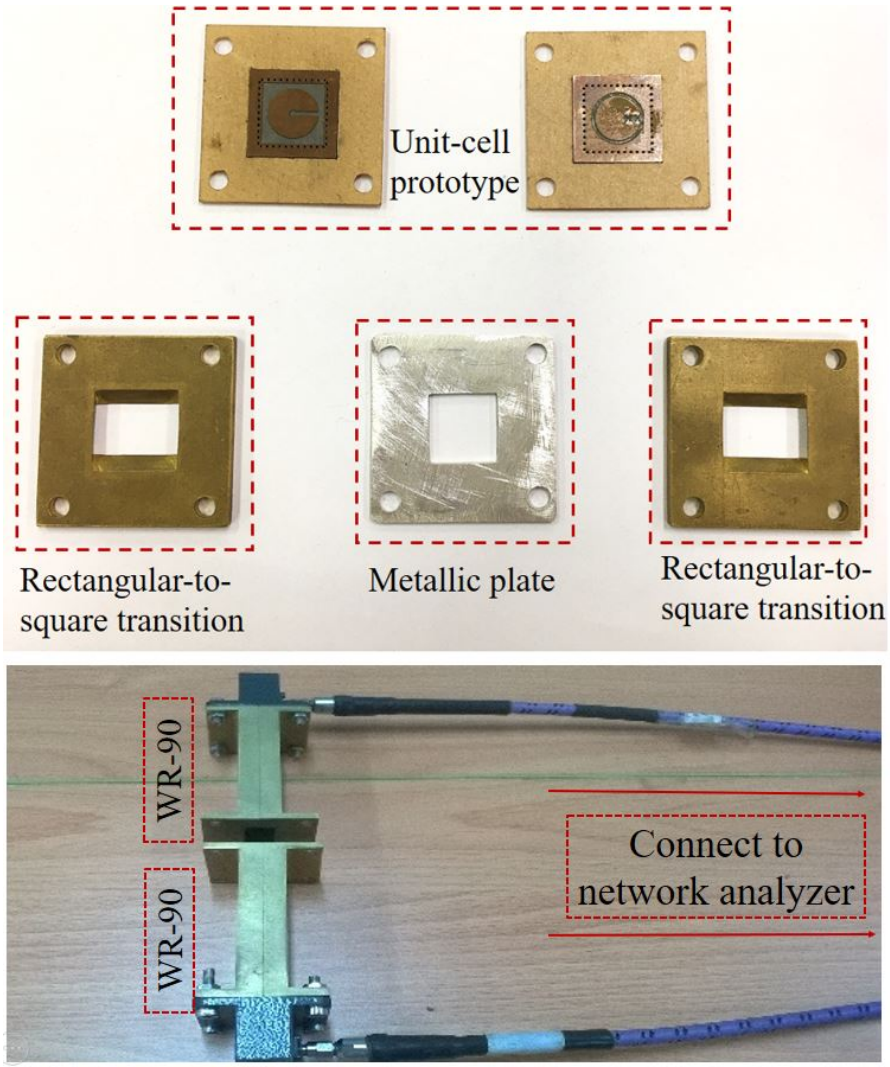


Figure 6. The measurement system.

to an F/D ratio of 0.89. The transmitarray in 3D and the simulation environment are shown in Figure 8.

In the design of a space-fed array antenna, to steer the main beam to direction (θ,ϕ), the transmission phase of each element can be calculated using equations (1) and (2) as follows:

$$\varphi(x_i, y_i) = k_0\big(d_i - \sin\theta(x_i\cos\phi + y_i\sin\phi)\big), \qquad (1)$$

$$d_i = \sqrt{(x_i - x_f)^2 + (y_i - y_f)^2 + (z_i - z_f)^2}, \qquad (2)$$

where (θ, ϕ) is the direction of main beam, x_i, y_i and z_i are the coordinates of the i^{th} element, x_f, y_f and z_f are the coordinates of the feed source, and k_0 is a propagation constant.

According to equation (1), the phase distribution on the transmitarray aperture is depicted in Figure 9. In this figure, the desired main beam direction is $\theta = \phi = 0°$.

Since the transmitarray antenna is based on the element which provides three phase states as discussed above, after calculating the theoretical compensation phase of each element using equations (1) and (2) for a main beam at direction (θ, ϕ), the real phase $\psi(x_i, y_i)$ of the element at the position with the coordinates x_i, y_i on the transmitarray is quantized using equation (3). This corresponds to the three phase states:

$$\psi(x_i, y_i) = \begin{cases} 0°, & -60° < \varphi(x_i, y_i) < -60°, \\ 120°, & 60° < \varphi(x_i, y_i) < 180°, \\ 240°, & 180° < \varphi(x_i, y_i) < 300°, \end{cases} \qquad (3)$$

where $\psi(x_i, y_i)$ is the quantized phase of the i^{th} element at the position with the coordinates x_i, y_i.

In order to evaluate the beam steering capabilities of the transmitarray, various phase distributions obtained by arranging the suitable transmission phase are designed.

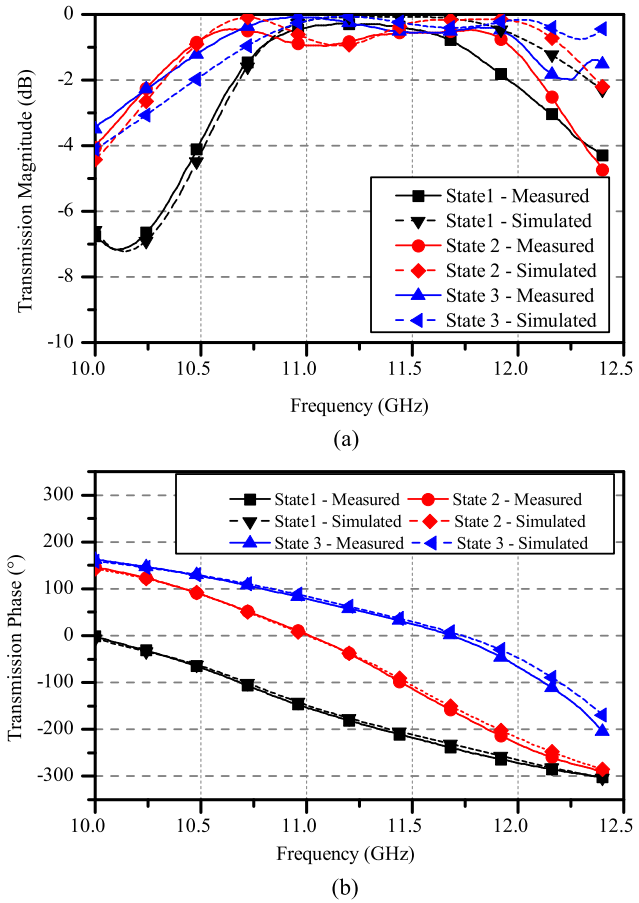


Figure 7. Measured and simulated transmission coefficients of the prototype: (a) transmission magnitude and (b) transmission phase.

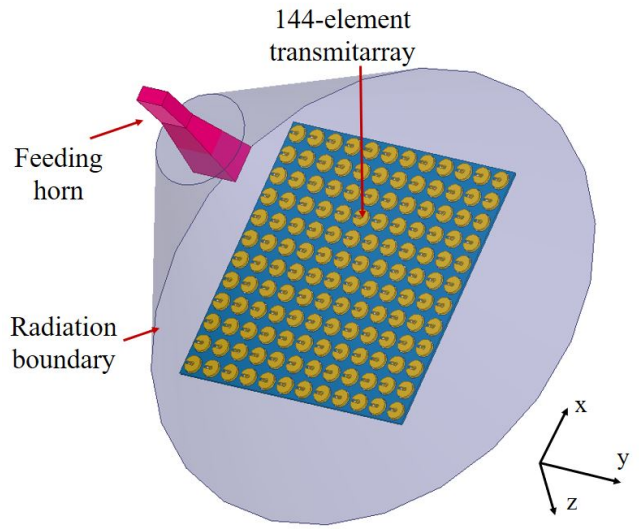


Figure 8. Simulation system for 12×12-element transmitarray.

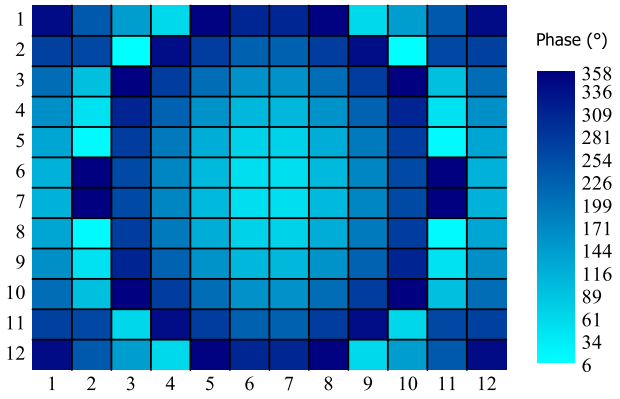


Figure 9. Theoretical compensation phase distribution required in the broadside transmitarray.

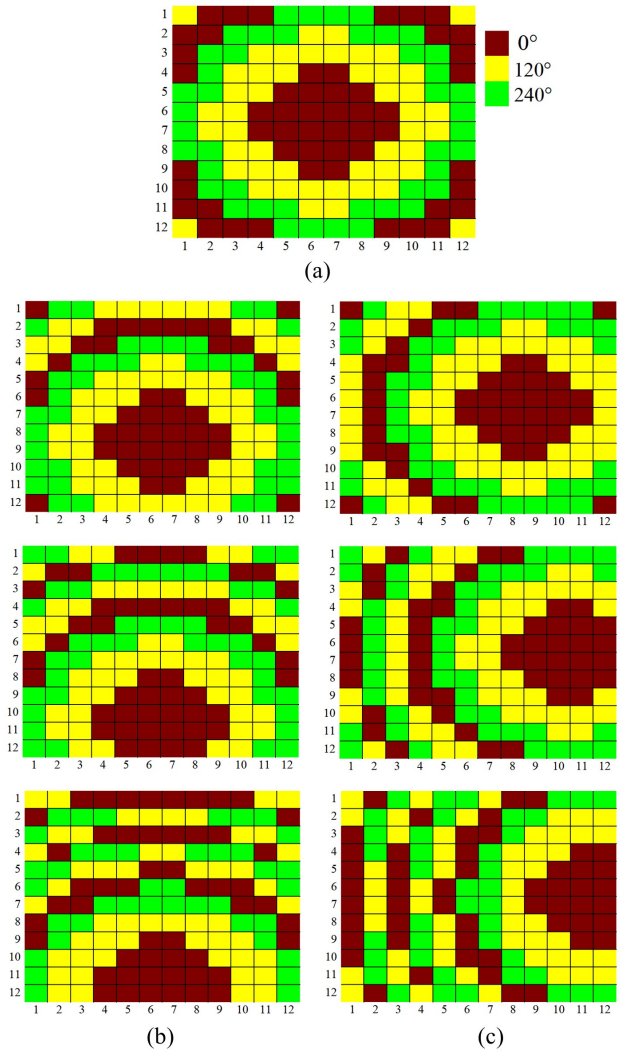


Figure 10. Phase distribution for the main beam pointed at different angles: (a) $\theta = 0°$, $\phi = 0°$ or $90°$, (b) $\phi = 0°$, $\theta = 10°$, $20°$, $30°$, and (c) $\phi = 90°$, $\theta = 10°$, $20°$, $30°$.

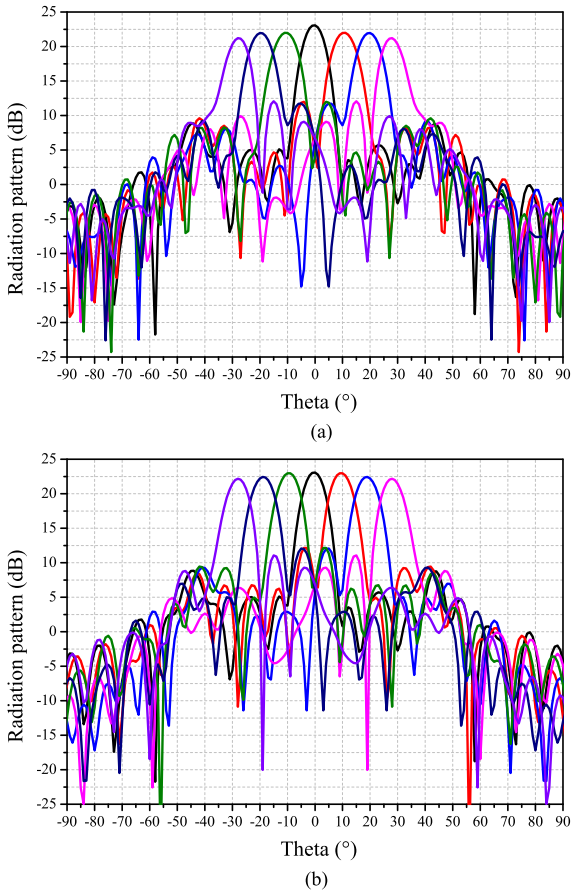


Figure 11. Simulated radiation patterns at 11.8 GHz with main beam tilted from −30° to +30° in (a) E-plane and (b) H-plane.

Figure 10 illustrates the phase distributions to steer the main beam directions along the theta angle in both E-plane ($\phi = 0°$) and H-plane ($\phi = 90°$). All phase distributions are constructed using three phase states. For the transmitarray, the elements with all diodes being OFF are used to provide the phase state 0°. The elements with first pair of diodes in ON state are used to provide the second phase state (120°). The elements with the second pair of diodes being ON provide the third phase states (240°). The scanning performance of the transmitarray has been evaluated by simulations. The radiation patterns at 11.5 GHz for different main beam directions varying from −30° to +30° for both E-plane and H-plane are shown in Figure 11. It can be seen that the main beam changes consistently with phase arrangement and low side lobes are obtained. The maximum directivity is 23.9 dBi when the main beam is at broadside direction while the scan loss reaches 1.6 dB and 2.9 dB for beams tilted at ±30° in H- and E- planes, respectively. In comparison with the 1-bit element for reconfigurable transmitarray in [12] where the maximum directivity is reported with 21 dBi, the transmitarray validated in this paper provides higher efficiency while the size and the number of elements are identical.

V. Conclusion

The three-phase-state element for reconfigurable transmitarray has been presented in this paper. Both simulation and measurement results validated good phase shifting capability and a wide −3 dB transmission bandwidth. While the prototype is still passive, where the ideal metallic strips are used as p-i-n diodes, simulation results indicated that the fully populated reconfigurable transmitarray can provide a wide scan angle with low scan loss. Further study and implementation of real p-i-n diodes will be deployed in an electronically tunable version of the transmitarray.

Acknowledgment

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2016.35.

References

[1] C. G. M. Ryan, M. R. Chaharmir, J. R. B. J. Shaker, J. R. Bray, Y. M. M. Antar, and A. Ittipiboon, "A wideband transmitarray using dual-resonant double square rings," *IEEE Transactions on Antennas and Propagation*, vol. 58, no. 5, pp. 1486–1493, 2010.

[2] G. Liu, H.-j. Wang, J.-s. Jiang, F. Xue, and M. Yi, "A high-efficiency transmitarray antenna using double split ring slot elements," *IEEE Antennas and Wireless Propagation Letters*, vol. 14, pp. 1415–1418, 2015.

[3] C. Tian, Y.-C. Jiao, G. Zhao, and H. Wang, "A wideband transmitarray using triple-layer elements combined with cross slots and double square rings," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 1561–1564, 2017.

[4] B. Rahmati and H. R. Hassani, "High-efficient wideband slot transmitarray antenna," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 11, pp. 5149–5155, 2015.

[5] T. Nguyen, B. D. Nguyen, V.-S. Tran, M. Linh, and L. H. Trinh, "Wideband unit-cell for linearly polarized X-band transmitarray applications," in *2018 IEEE International Conference on Advanced Technologies for Communications (ATC)*, 2018, pp. 125–128.

[6] P. Padilla, A. Munoz-Acevedo, M. Sierra-Castaner, and M. Sierra-Perez, "Electronically reconfigurable transmitarray at Ku band for microwave applications," *IEEE Transactions on Antennas and Propagation*, vol. 58, no. 8, pp. 2571–2579, 2010.

[7] J. Y. Lau and S. V. Hum, "A wideband reconfigurable transmitarray element," *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 3, pp. 1303–1311, 2011.

[8] L. Boccia, I. Russo, G. Amendola, and G. Di Massa, "Multilayer antenna-filter antenna for beam-steering transmit-array applications," *IEEE transactions on microwave theory and techniques*, vol. 60, no. 7, pp. 2287–2300, 2012.

[9] C.-C. Cheng, B. Lakshminarayanan, and A. Abbaspour-Tamijani, "A programmable lens-array antenna with monolithically integrated MEMS switches," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 8, pp. 1874–1884, 2009.

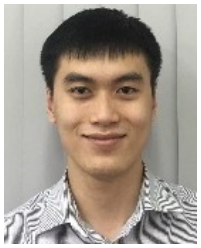
[10] L. Di Palma, A. Clemente, L. Dussopt, R. Sauleau, P. Potier, and P. Pouliguen, "1-bit reconfigurable unit cell for Ka-band transmitarrays," *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 560–563, 2015.

[11] A. Clemente, L. Dussopt, R. Sauleau, P. Potier, and P. Pouliguen, "1-Bit reconfigurable unit cell based on PIN diodes for transmit-array applications in X-band," *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 5, pp. 2260–2269, 2012.

[12] B. D. Nguyen and C. Pichot, "Unit-cell loaded with PIN diodes for 1-bit linearly polarized reconfigurable transmitarrays," *IEEE Antennas and Wireless Propagation Letters*, vol. 18, no. 1, pp. 98–102, 2018.

[13] F. Diaby, A. Clemente, L. Di Palma, L. Dussopt, K. Pham, E. Fourn, and R. Sauleau, "Linearly-polarized electronically reconfigurable transmitarray antenna with 2-bit phase resolution in Ka-band," in *2017 IEEE International Conference on Electromagnetics in Advanced Applications (ICEAA)*, 2017, pp. 1295–1298.

[14] B. D. Nguyen and M. T. Nguyen, "Three-bit unit-cell with low profile for X-band linearly polarized transmitarrays." *Applied Computational Electromagnetics Society Journal*, vol. 38, no. 9, 2019.



Nguyen Minh Thien was born in Vietnam in 1995. He received his Bachelor of Engineering in Electrical Engineering from the International University, Ho Chi Minh City in 2017. He is currently pursuing a Master program in the School of Electrical Engineering, International University. His research interests mainly focus on design high gain antenna array, unit-cell design for passive reflectarray, transmitarrays, electronically reconfigurable transmitarray.

Nguyen Huu Minh was born in Vietnam in 1992. He received his Bachelor of Engineering in Electrical Engineering from the International University, Ho Chi Minh City in 2019. He is currently working as a hardware engineer for Homa Techs Inc. His interests mainly focus on passive and active transmitarrays, PCB antenna design.

Nguyen Binh Duong was born in Vietnam in 1976. He received the B.S. degree in electronic and electrical engineering from Ho Chi Minh University of Technologies, Ho Chi Minh, Vietnam, in 2000 and the M.S. and Ph.D. degrees in electronic engineering from the University of Nice-Sophia Antipolis, France, in 2001 and 2006 respectively. From 2001 to 2006, he was as a Researcher at the Laboratoire d'Electronique d'Antennes et Telecommunication, University of Nice-Sophia Antipolis, France. His research interests focus on millimeter antenna, reflector, reflectarray and FSS.

An Evaluation of Pose Estimation in Video of Traditional Martial Arts Presentation

Nguyen Tuong Thanh[1], Le Van Hung[2], Pham Thanh Cong[1]

[1] School of Electronics and Telecommunications, Hanoi University Science and Technology, Vietnam

[2] Tan Trao University, Vietnam

Correspondence: Le Van Hung, van-hung.le@mica.edu.vn

Communication: received 20 May 2019, revised 8 September 2019, accepted 11 September 2019

Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.864

The Editor coordinating the review of this article and deciding to accept it was Dr. Le Vu Ha

Abstract: Preserving, maintaining, and teaching traditional martial arts are very important activities in social life. That helps individuals preserve national culture, exercise, and practice self-defense. However, traditional martial arts have many different postures as well as varied movements of the body and body parts. The problem of estimating the actions of human body still has many challenges, such as accuracy, obscurity, and so forth. This paper begins with a review of several methods of 2-D human pose estimation on the RGB images, in which the methods of using the Convolutional Neural Network (CNN) models have outstanding advantages in terms of processing time and accuracy. In this work we built a small dataset and used CNN for estimating keypoints and joints of actions in traditional martial arts videos. Next we applied the measurements (length of joints, deviation angle of joints, and deviation of keypoints) for evaluating pose estimation in 2-D and 3-D spaces. The estimator was trained on the classic MSCOCO Keypoints Challenge dataset, the results were evaluated on a well-known dataset of Martial Arts, Dancing, and Sports dataset. The results were quantitatively evaluated and reported in this paper.

Keywords: *Estimation of keypoints, pose estimation, deep learning, skeleton, conserving and teaching traditional martial arts.*

I. Introduction

Estimating and predicting actions of human body are well-studied problems in the robotics and computer vision community [1]. Application domains include social safety, preservation of cultural identity values (conserving and maintaining traditional martial arts and national dance songs), production of toys and games, interaction with intelligent robots, sports analysis (tactical analysis in sports such as football, tennis, badminton, etc.), health protection (detection of falling events in hospital for the elderly), etc. Solving these problems can be based on a set of methods such as analyzing people in the images, locating people in the images, locating keypoints on human bodies, and identifying joints (skeleton) from the points featuring their body.

The problem of estimating the skeleton from the image of a person is usually based on color images, depth images, or the contexts of objects and actions [1]. The above systems often use color image information, depth information [2], or skeleton [3] obtained from different types of sensors. In particular, the Microsoft (MS) Kinect sensor version 1 (v1) is a common and cheap sensor that can collect several types of information such as color, depth, skeleton, and acceleration vector [4].

Color is the most common type of information obtained from cameras/sensors. Changes of appearance and posture of the human body structure in the image create a set of characteristics of deformation part model (DPM). That makes it difficult to estimate the shape and joints of the human body. The transformation of a complex human body is made up of changes in human body parts, which can be common transformations such as translation, rotation, and resizing. Previous studies often train DPM feature sets for detecting, recognizing, and estimating human postures and poses in images [5–7].

Recently, human pose estimation is still very challenging [3] in such terms as processing time and accuracy [8], especially for 3-D human pose estimation performed on datasets which have many occlusions [9]. Currently, with strong development of deep learning for detection, recognition, and human action estimation, it has become a good approach for solving these problems. There are many proposed Convolutional Neural Networks (CNNs) which achieved very good results in detecting and recognizing objects, such as Fast R-CNN [10], Faster R-CNN [11], and YOLO [12, 13]. Recently, there have been many studies of skeleton estimation on images using CNN models, such as [14–16].

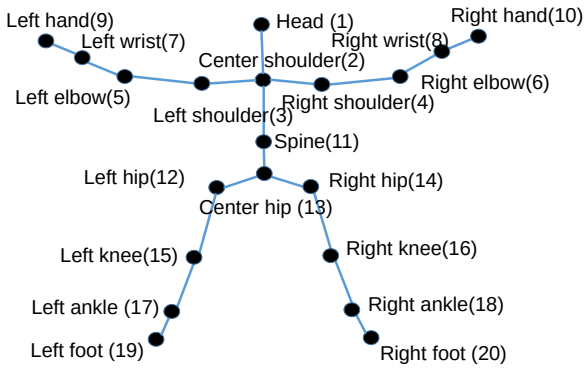


Figure 1. Keypoints on the human body and the labels.

In this work we only use the CNN model which was proposed and trained as [17] to estimate and predict the actions of people in videos of instructors and practitioners performing martial arts. This approach is based on the model that is trained from the confidence maps of feature vectors [18] which are extracted from the training images. The trained model estimates the keypoints on the human skeleton model as seen in Figure 1. In particular, this approach can estimate the posture of people based on the skeleton in case of being obscured.

Data obtained from the MS Kinect sensor includes color images, depth images, and correspondence. The first two types of data are calibrated to a center based on the approach and the intrinsic matrix of the MS Kinect sensor v1 proposed by Nicolas *et al.* [19]. Each frame builds a scene in 3-D environment. The estimated results of keypoints and joints are also transferred to 3-D space. It is then possible to build a martial arts teaching application in a more intuitive way.

The main contributions of this work include: (i) using a CNN model for 2-D human pose estimation on RGB images, achieving outstanding advantages in terms of processing time and accuracy; (ii) a small dataset of traditional martial arts videos and using the CNN-based pose estimation model [17] trained on the COCO [20] dataset to evaluate the skeleton estimations performed on the contributed dataset and the dataset of Zhang *et al.* [21]; and (iii) proposing measurements to evaluate human pose estimations in 2-D and 3-D spaces.

II. Related Works

In Vietnam [22, 23] as well as many countries in the world like China [24], Japan, and Thailand, there are many martial arts postures or martial arts that need to be preserved and passed down to posterity. Conservation and storage in the era of technology can be done in many different ways. An intuitive approach is to save the joints in the skeleton model of a martial arts instructor. Data obtained from MS Kinect sensor v1 usually contains a lot of noise and is lost when being obscured, especially skeleton data of people. The skeleton data is important and presents the human pose in a video action.

In the past, studies often looked into deformation part model features to address the problem of skeleton estimation on images. Felzenszwalb *et al.* [5] proposed a method for training a multiscale deformation part model (DPM) for object detection on images. In a partial deformation model approach, the human body is represented as a star-shaped structure, consisting of a root filter, a set of part detectors, and a partial deformation model. In the DPM model, deformation refers to relative positions of the body parts. An SVM (Support Vector Machine) classifier is trained on extracted features to predict the positions of the human body parts. Sun *et al.* [6] proposed a model based on the Articulated Part-based Model (APM) to detect parts of the human body and estimate the posture of the person. The APM model represents an object as a collection of parts at a level of details in the range from coarse to smooth, in which parts at all levels are connected to a coarser level through a parent-child relationship. Pishchulin *et al.* [25] as well as Andriluka [26] used the method of dividing the human body into parts and training the model on the parts for the estimation of one's body pose. Andriluka *et al.* [26] used AdaBoost for predicting the posture of the person. Umer *et al.* [27] used Regression Forests to estimate the direction of users on depth images obtained from MS Kinect sensor v2. The model estimation was performed on the parts of the labeled person, with 1000 sample position patterns on depth images. However, the highest average accuracy was just 35.77%.

Recently, with the strong development of deep learning, the estimation of keypoints on human body is often done by CNN models. Daniil *et al.* [14] introduced a new CNN model for learning the features on a keypoint dataset: the location of keypoints and the relationship between pairs of points on the human body. This new network is based on the OpenPose toolkit [17] and training can be done without GPU (CPU only). In particular, the CNN model of this study is trained and evaluated on the COCO 2016 key-point challenge dataset [20]. This is a huge dataset of labels containing images of over 150 thousands people with 1.7 million labels of keypoints. Kyle *et al.* [15] used a CNN model to learn from the data of the keypoints of the human body that were marked and extracted from the connection data when projecting two cameras into people. The result was then projected into 3-D space and then the least squared distance algorithm was used to evaluate the

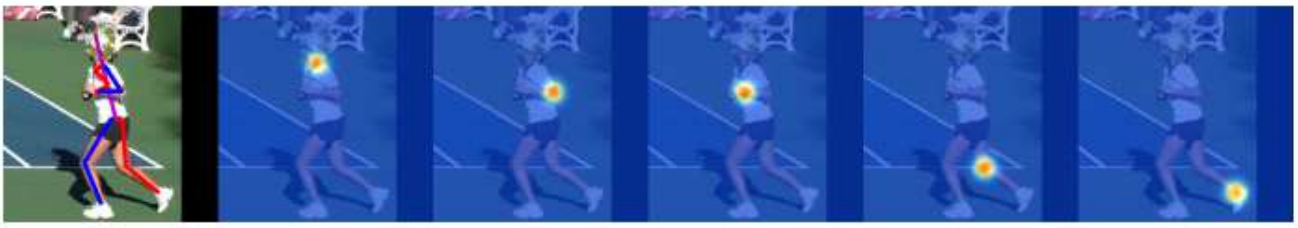


Figure 2. Illustration of heatmaps predicted from human body image. Therein, each heatmap is a candidate prediction of keypoint locations (x, y) [28].

obtained estimates. Cao *et al.* [18] used a CNN model to learn the positions of keypoints on the human body and the geometric transformations of the lines connecting the keypoint pairs with the above connected human body. Evaluations were conducted on two classic datasets, the MPII [29] and the COCO [20]. In particular, the COCO dataset of keypoints [30, 31] has been developed for many years. MPII and COCO datasets contain images of hundreds of thousands of people and have been used in many challenges/competitions on human activity estimation.

Toshev *et al.* [32] estimated human posture and skeleton, considering the human skeleton as a CNN-based regression. The authors also used a sequence of regression variables to correct the posture and skeleton estimations to get a better estimation. It is important that this method is based on a completed shape of posture and skeleton. When the joints are obstructed, they can be estimated from the completed posture and the skeleton structure. The model in this study is trained on the AlexNet backend network of seven layers, in which the final layer is used to complete the trained model and the target output values of the regression, the number of target values was about 2000 in term of joint coordinates.

Tompson *et al.* [28] created heatmaps by simultaneously running an image through many different resolutions to collect multi-resolution features at the same time. The output is a discrete heatmap instead of continuous regression. A heatmap predicts the probability of joints at each pixel, in which each heatmap area is a candidate of the location of keypoints on the human body (heatmap areas are created as shown in Figure 2).

In addition to the training method and the prediction of heatmaps, Wei *et al.* [33] proposed a network that trains through many phases on the characteristic set of images. They provided a sequential predictive framework focusing on training highly predictive models. Output heatmaps of the previous stage are used as input of the later stages, with which the best accuracy obtained by this model on the MPII dataset [34] was 87.95%.

Andriluka *et al.* [35] published a dataset that is structured and organized similarly to the classic datasets. These benchmark datasets provide training sets, validation sets, and testing sets to train and evaluate deep learning-based methods. In particular, they also established performance metrics for direct and fair comparison across numerous competing approaches. Girdhar *et al.* [16] proposed a 3-D human pose predictor by inflating the 2-D convolutions into 3-D [36] to extend the Mask RCNN [37] with spatio-temporal operations. This work used the Mask RCNN [37] for 2-D human pose estimation, the tracking process was the combination of 2-D human pose estimation results and temporal information. This method achieved high accuracies on the challenging PoseTrack benchmark dataset [35]. However, since this method must predict the box, the segmentation, and the keypoints, the processing time of testing is large (5 frames/s with 8 GPUs). Meanwhile, the method proposed by Cao *et al.* [18] is able to process about 10-15 frames/s with a 12 GB GPU.

III. Human Pose Estimation

The activity of the human body is detected and recognized as well as predicted and estimated based on body parts. 3-D human pose estimation employs either one of the two basic methods: (1) estimating the 3-D human pose from a single image (RGB or depth), and (2) estimating the 3-D human pose from an image sequence. There are many studies on 3-D human pose estimation that use the single-image method [9, 38, 39]. In these studies, the keypoints and joints are estimated on 2-D images and then mapped to the 3-D space. This model is often applied to estimating 3-D human pose, as shown in Figure 3 [40].

In this paper, to estimate the 2-D human pose, we use the method of Cao *et al.* [18]. The architecture of the CNN to train the model is shown in Figure 3. This CNN consists of two branches performing two different jobs. From the input data, a set $\mathbf{F}$ of feature maps is created from image analysis, these confidence maps and affinity fields are detected at the first stage.

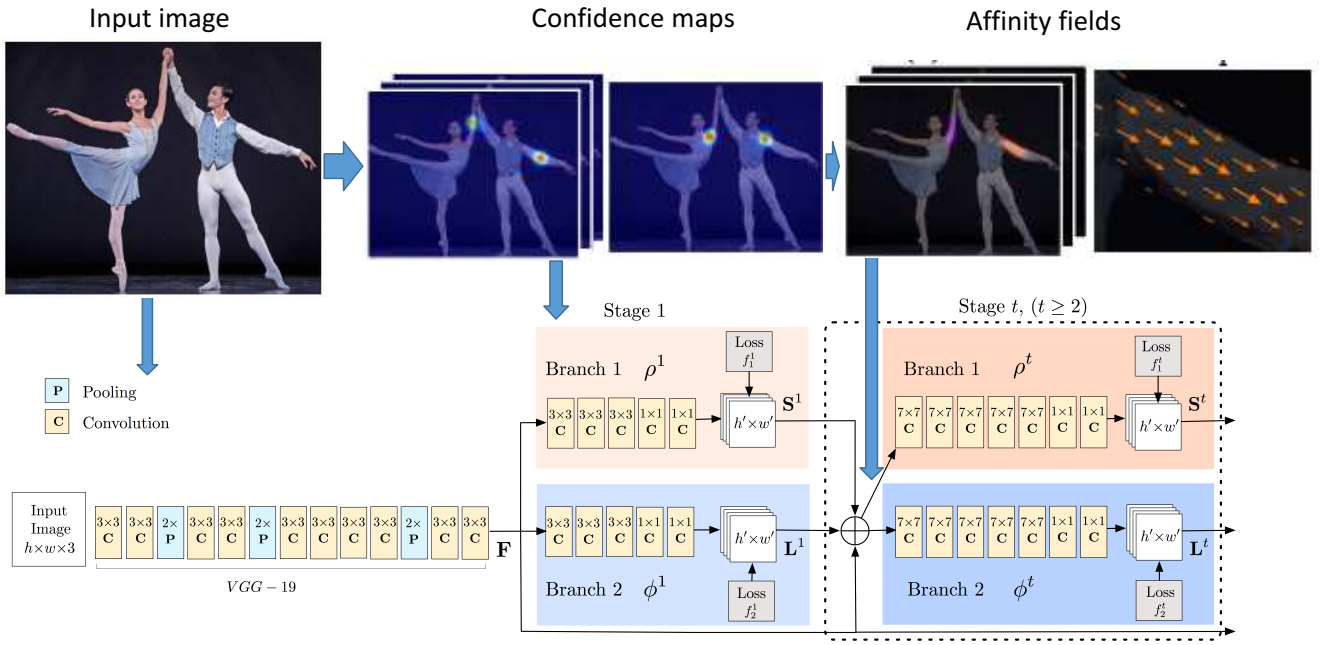


Figure 3. The architecture of the two-branch multi-stage CNN for training the estimation model [18].

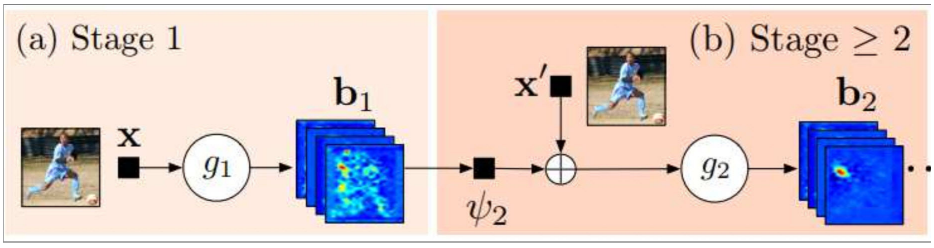


Figure 4. Illustration of the detailed model to predict heatmaps [33].

Details on Cao's model training and prediction (Figure 4) [18] are shown as follows. The input image at stage 1 is an RGB image which has a size of $h \times w$. Features extracted from convolutions with masks of sizes $9\times9, 2\times2, 5\times5, \ldots$ for the training set $\mathbf{X}$ as shown in Figure 5. For each mask, there will be a trained model at each stage. As shown in Figure 4, models g_1 and g_2 at stages 1 and 2 will predict the heatmaps b_1 and b_2, respectively. In Figures 4 and 5, the Convolutional Pose Machines consist of at least 2 stages and the number of phases is a super parameter (usually 3 stages). The second stage takes the resulted heatmaps of the first stage as the input.

Therein, each heatmap indicates the location confidence of the keypoints as a function of (x, y). Keypoints on the training data are displayed on confidence maps as shown in Figure 4. These points are estimated by the trained model as the keypoints on input color images. The first branch (top branch) is used to estimate the keypoints, the second branch (bottom branch) is used to predict the affinity fields matching joints on people.

In addition, we also render a 3-D environment of each video's scene and project the results of 2-D human pose estimation into the 3-D space, based on the intrinsic parameter of the Kinect sensor v1, using the PCL library [41] and the OpenCV library [42] functions. The real coordinates (x_p, y_p, z_p) and color values of each pixel when projected from 2-D to 3-D space are calculated as in Eq. 1.

$$\begin{aligned} X_p &= \frac{(x_a - c_x) * depthvalue(x_a, y_a)}{f_x} \\ Y_p &= \frac{(y_a - c_y) * depthvalue(x_a, y_a)}{f_y} \\ Z_p &= depthvalue(x_a, y_a) \\ C(r, g, b) &= colorvalue(x_a, y_a) \end{aligned} \tag{1}$$

where $depthvalue(x_a, y_a)$ is the depth value of a pixel at (x_a, y_a) on the depth image, and $colorvalue(x_a, y_a)$ returns the RGB color values of that pixel on the color image.

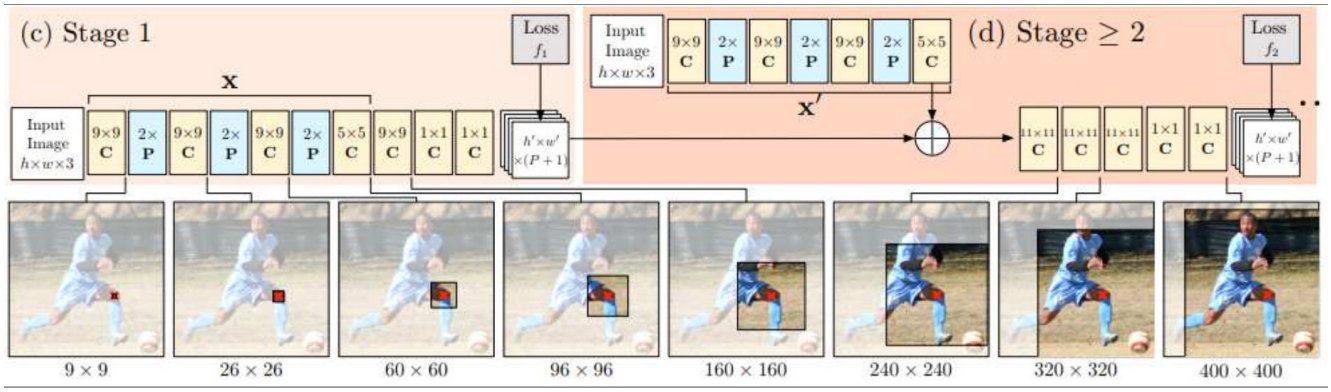


Figure 5. Illustration of the detailed model to extract features for training model and to predict heatmaps at each stage [33].

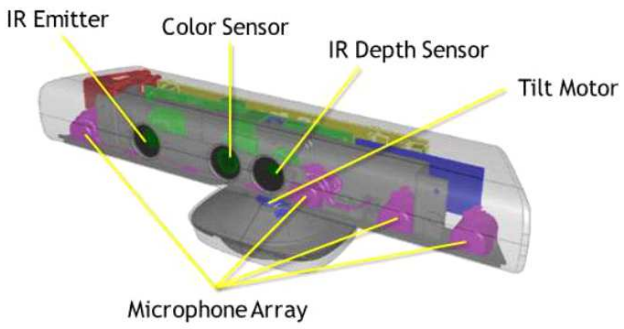


Figure 6. MS Kinect sensor v1.

Figure 7. Illustration of the obtained image from MS Kinect sensor v1 and annotated keypoints of the skeleton model.

In regard to the process of combining color and depth information of a pixel to obtain a point in 3-D space, for cases where the depth values of pixels in the depth image are lost (value is zero), we use the average depth value of pixels in their 50×50 neighborhood.

In our experiments, results of 2-D human pose estimation are the keypoints $\{(x_e, y_e)|e \in \{1, 2, ..., 25\}\}$. The joints are then joined according to a predefined order.

IV. Experimental Result

1. Data Collection

There are many different types of image sensors that can collect information about martial arts teaching and learning. The MS Kinect v1 sensor as seen in Figure 6 is the cheapest sensor today. This type of sensor can collect a lot of information such as color images, depth images, skeletons, acceleration vectors, sound, etc. From the collected data, it is possible to recreate the environment in 3-D space. However, in this work we only use color images captured by the MS Kinect v1 sensor.

To capture data from the sensor, the MS Kinect SDK 1.8 is used [43]. To perform data collection on computers, we use a data collection program developed at MICA Institute [44] with the support of the OpenCV 3.4 libraries [42]. Calibration is required in order to generate 3-D data from color and depth images. Particularly, we apply the calibration methods of Zhou *et al.* [45] and Jean *et al.* [46]. In these two calibration tools, the calibration matrix is used as follows:

$$H_m = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

where (c_x, c_y) is the principle point (usually the image center) and (f_x, f_y) is the focal length vector. The matrix H_m (in Nicolas *et al.* [19]) is given as follows:

$$H_m = \begin{bmatrix} 594.214 & 0 & 339.307 \\ 0 & 591.040 & 242.739 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

In this work, we use two datasets for evaluating the model pose estimation [17]. The first dataset is collected from a

Figure 8. Illustrations on ground-truth for keypoints. Red points are keypoints on the human body. Blue segments show connections between parts of the human body.

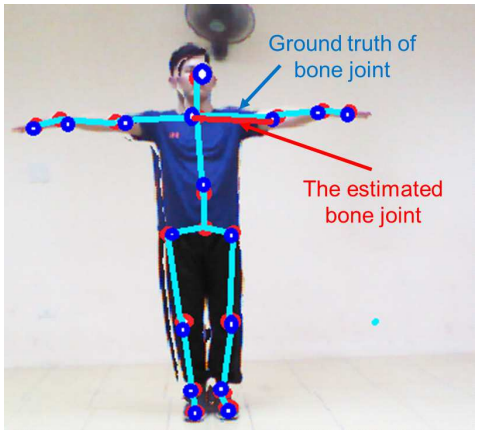


Figure 9. Illustration of the estimated results of the keypoints and joints.

MS Kinect v1 sensor, which can collect data at a rate of about 10 frames/s on a low-performance laptop. The MS Kinect sensor v1 is mounted on a fixed rack; the martial arts instructor presents in a 3×3m space as in Figure 10. It is called "VNMA - VietNam Martial Arts," captured in a martial arts class in Binh Dinh province, Vietnam. Binh Dinh martial art is one of the famous traditional martial arts of Vietnam.

The obtained images (color images and depth images) are 640×480 in pixels. The dataset consists of 14 videos of different postures, with the number of frames listed in Table I and illustrated in Figure 8. This dataset features a martial arts instructor with 14 different postures. The number of frames is the number of poses in each video. The ground-truths for keypoints are manually prepared, as illustrated in Figures 8 and 9. The ground-truth data in each image, which contains a single person, includes 18 keypoints.

TABLE I
NUMBER OF FRAMES IN MARTIAL ARTS POSTURES

Video	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Number of frames	120	74	100	87	80	88	87
Video	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Number of frames	74	71	90	100	97	65	68

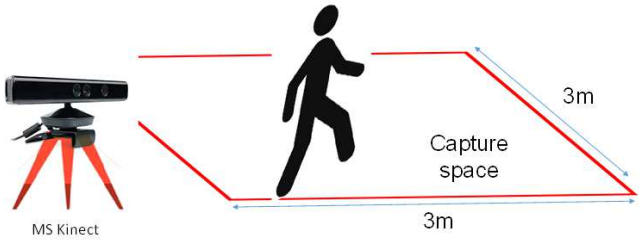


Figure 10. Illustration of MS Kinect v1 sensor settings and data collection space.

The second dataset called Martial Arts, Dancing, and Sports[21] consists of both multi-view RGB videos and depth videos. This dataset contains five action types: Tai-chi, Karate, Hip-hop dance, Jazz dance, and sports. The frame rate is 10 fps for Tai-chi and Karate, or 20 fps for jazz, hip-hop, and sports videos. The resolution of the images are 1024×768 in pixels. However, all the frames in this dataset are resized to 512×384 pixels. Ground-truth data was prepared for 3-D poses, using a MOCAP (MOtion CAPture) system [47] by Motion Analysis. Seven MOCAP cameras were placed on the walls around the capture space to record the positions of markers on the human body. The MOCAP system works at 60 fps. The dataset contains ground-truth data for 19 joints: neck, pelvis, left hip, left knee, left ankle left foot, right hip, right knee, right ankle right foot, left shoulder, left elbow, left wrist, left hand, right shoulder, right elbow, right wrist, right hand, and head.

We use only Tai-chi and Karate videos of color and depth images that contain 11200 frames. The camera calibration matrix of the capture system is given by

$$H_m = \begin{bmatrix} 331 & 0 & 254.097 \\ 0 & 331 & 180.032 \\ 0 & 0 & 1 \end{bmatrix}. \quad (4)$$

Figure 11 illustrates the point cloud data of a scene when a person presents Karate.

We use the 2016 MSCOCO Keypoints Challenge dataset [48] to train the model, because our collected dataset (VNMA - VietNam Martial Arts) is small and the MADS (Martial Arts, Dancing, and Sports) dataset provides

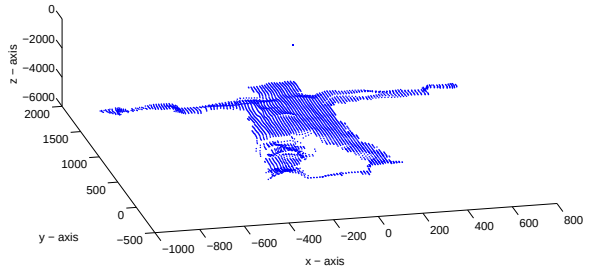


Figure 11. Illustration a point cloud data of a scene. The blue points represents the human body in the 3-D environment.

only 3-D ground-truth data. The training dataset consists of over 100K people and a million labeled keypoints. Each person is marked by 17 keypoints, which are sorted by the following order: "nose," "left_eye," "right_eye," "left_ear," "right_ear," "left_shoulder," "right_shoulder," "left_elbow," "right_elbow," "left_wrist," "right_wrist," "left_hip," "right_hip," "left_knee," "right_knee," "left_ankle," and "right_ankle." The model is trained for estimating and labeling these 17 keypoints on the human body.

A toolkit [49] is used for model training and testing. This toolkit is installed on Ubuntu 16.2 and supported by several libraries: OpenPose C++ library (for testing only) [17], Tensorflow version 1.8 [50], Pytorch [18, 51], Caffe2 [52], Chainer [53], MXnet [54], MatConvnet [55], and CNTK [56]. Readers are referred to [18, 48] for details. The training tool and the testing tool are written in Python/Matlab/C++ language and run on a workstation computer which has the following configurations: Intel (R) Xeon (R) CPU E5-2420 v2 @ 2.20 GHz 16 GB RAM and GTX 1080 Ti GPU with 12 GB RAM. The trained model is used to estimate 25 keypoints in the following order: "nose," "neck," "right shoulder," "right elbow," "right wrist," "left shoulder," "left elbow," "left wrist," "mid hip," "right hip," "right knee," "right ankle," "left hip," "left knee," "left ankle," " right eye," "left eye," "right ear," "left ear," "left big toe," "left small toe," "left heel," "right big toe," "right small toe," and "right heel."

The CNN training parameters are as follows[1]: size of the input images is $368 \times 368 \times 3$ (width x height x number of channels), $batchSize = 16$, $stacks = 4$, and the number of stages is 6 for pooling. Our collected VNMA dataset is still small, therefore we have not divided this dataset into training set, validation set, and testing set. This dataset is used only for testing.

[1]Detailed parameters are shown in https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation/blob/master/training/example_proto/pose_train_test.prototxt.

2. Evaluation Method

As in [18], we use Object Keypoint Similarity (OKS) measure for evaluating the similarities between estimated and ground-truth joints. The formula for calculating OKS is given by [48]:

$$\text{OSK} = \frac{|G_{\text{ground}} - R_{\text{result}}|}{G_{\text{ground}}}, \tag{5}$$

where G_{ground} is the length of a ground-truth joint vector, R_{result} is the length of the corresponding estimated joint vector. The Average Precision (AP) is calculated based on the threshold value of 0.5 for OKS. Therein, if OKS > 0.5, meaning the difference is greater than 50% of length, then it is considered a false estimate, otherwise it is a true estimate.

We also propose another metric called the angle of deflection between an estimated joint $\mathbf{V}_E$ its corresponding ground-truth joint $\mathbf{V}_G$, which is the angle between the two vectors: $A = \arccos(\mathbf{V}_G, \mathbf{V}_E)$. If $A \leq 10^o$ it is considered a true estimate, otherwise it is a false estimate. Denote AD the ratio of true estimates over the total number of joints, and DP the average distance between estimated keypoints and ground-truth keypoints.

The average distance in 2-D space is given by

$$\text{DP}_{\text{2D}} = E\left[\sqrt{(x_g - x_e)^2 + (y_g - y_e)^2}\right], \tag{6}$$

where x_e and y_e are the coordinates of an estimated keypoint, while x_g and y_g are the coordinates of its corresponding ground-truth keypoint, E is the expectation operator.

The average distance in 3-D space is given by

$$\text{DP}_{\text{3D}} = E\left[\sqrt{(x_g - x_e)^2 + (y_g - y_e)^2 + (z_g - z_e)^2}\right], \tag{7}$$

where x_e, y_e, and z_e are the 3-D coordinates of an estimated keypoint, while x_g, y_g, and z_g are the 3-D coordinates of its corresponding ground-truth keypoint.

3. Results and Discussion

Performance of 2-D pose estimation on the VNMA dataset is presented in Table II. The average value of AP in Tabble II is 95.6%. For noisy data performance could be significantly lower, an example is video #4 (AP = 89.6%).

Pose estimation results are shown in Table II and Figure 12, in which 25 keypoints are estimated on the human body. However, the corresponding ground-truth data provides only 20 keypoints on each person, therefore the performance evaluation is performed on only 20 of the 25 estimated keypoints. It can be seen that the estimation results are highly accurate, although the model trained on the MSCOCO Keypoints Challenge dataset [48] and our test data contain a lot of noise. The average accuracy ratio

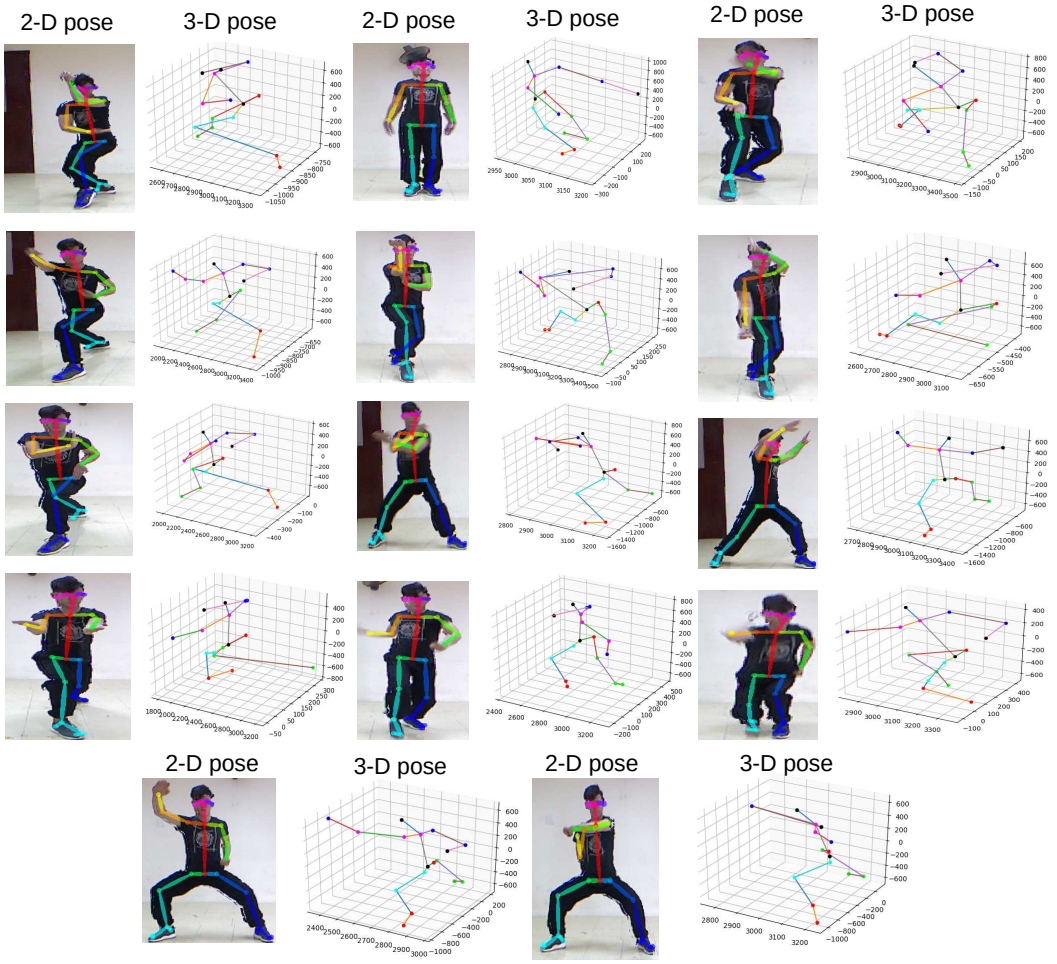


Figure 12. Results of joint estimation shown in 2-D and 3-D spaces.

AD, which is based on the angles of deflection between the estimated joints and the ground-truth, is 95.3% for this test dataset. The average value of DP_{2D} which is the deviation between estimated keypoints and the ground-truth in 2-D image space is 14.73 pixels[2]. Figure 12 shows estimated skeletons in 2-D and 3-D spaces, represented by 17 keypoints[3].

We also measure prediction probabilities in term of Intersection of Union (IOU) for all keypoints of three videos. These probabilities are represented as heatmaps produced by the CNN model, in which each heatmap area shows the confidence scores associated with candidate positions of a keypoint, as presented in Figures 4 and 5. Resulted probability distributions are shown in Figure 13. We can see

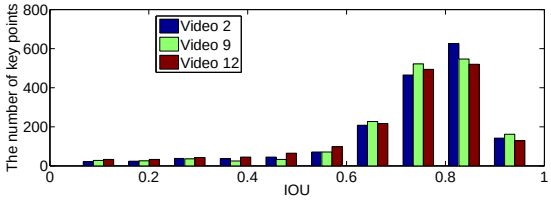


Figure 13. Probability distributions of the keypoint candidates predicted by the CNN model for three martial arts videos.

that all three distributions are peeked within 0.8 to 0.9. That means the trained model in [17] has good predictability.

Performance of 3-D pose estimation on the MADS dataset is presented in Tables III and IV. From the figures in Table III, the average value of AP is 75.93% and average accuracy ratio AD is 79.62% and the estimated 2-D skeleton results are illustrated in Figure 14. Table IV compares performances of the TGP-KNN method [57] and the Pose toolkit [17] on the MADS dataset, using the

[2]Details of the estimated results are available at this link: https://drive.google.com/file/d/1BaFxUbOJvRll5tkgYQW98VOp6LC414Im/view?usp=sharing.

[3]The source code using OpenPose library is shared at the following link: https://drive.google.com/file/d/1OZO8m7TvtZUD55zf1TAU1HTWCk-u4pb8/view?usp=sharing.

TABLE II
AVERAGE PRECISIONS FOR ESTIMATED JOINTS (AP), AVERAGE RATIOS OF TRUE ESTIMATES FOR JOINTS (AD), AND AVERAGE DEVIATIONS OF ESTIMATED KEYPOINTS IN 2-D IMAGE SPACE (DP_{2D}) OBTAINED ON THE VIDEOS OF THE VNMA DATASET

Video	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
AP (%)	95.4	93.7	96.2	89.6	96.1	92.8	97.4	98.8	96.9	94.5	96.9	96.2	95.7	98.2
AD (%)	93.7	94.6	92.8	90.9	95.3	94.6	95.8	97.6	97.8	95.1	97.0	95.8	96.3	96.9
DP_{2D} **(pixels)**	21.2	18.6	9.7	25.9	13.8	15.7	9.4	15.4	12.4	10.1	14.0	12.8	11.3	16.9

TABLE III
AVERAGE PRECISIONS FOR ESTIMATED JOINTS (AP) AND AVERAGE RATIOS OF TRUE ESTIMATES FOR JOINTS (AD) OBTAINED ON THE VIDEOS OF THE MADS DATASET

Video	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
AP (%)	71.20	70.53	72.64	60.53	80.50	73.22	81.95	91.31	55.19	86.65	79.98	87.46
AD (%)	80.45	75.83	80.76	62.46	78.77	78.93	89.13	87.65	59.84	86.46	86.94	88.2

TABLE IV
AVERAGE 3-D DEVIATIONS DP_{3D} (MM) BETWEEN ESTIMATED KEYPOINTS AND THE GROUND-TRUTH FOR VIDEOS OF THE MADS DATASET, USING TWO DIFFERENT METHODS

Video	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
TGP-KNN [57]	214.4	167.4	**246.7**	149.3	**119.5**	197.2	119.4	**114.2**	**214.5**	113.2	**98.4**	133.7
Pose toolkit [17]	**198.23**	**154.38**	252.3	**148.44**	120.34	**188.75**	**112.68**	116.3	**206.65**	**104.29**	102.34	**120.55**

deviations between estimated keypoints and the ground-truth in 3-D space. The TGP-KNN results are those reported in [21]. The processing time to estimate keypoints and joints is about 75 ms per frame.

Figure 15 illustrates the color point cloud of a scene in the VNMA dataset. Therein, the coordinate system (the x-axis is colored red, the y-axis is colored green, and the z-axis is colored blue) of the MS Kinect v1 sensor is shown in the original frame. Figure 16 illustrates 3-D joint estimation results from a scene. The human body is presented by the point cloud.

Figure 17 shows the estimated results in the 2-D space of the proposed dataset. This dataset were collected from a MS Kinect v1 sensor. However, in martial arts presentation, cases vary and the posture of the person must rotate in many different directions. Therefore, there are many cases where many actions are obscured. In order to build a conservation application for training martial arts and evaluating martial arts videos, the bone joints of these actions should be restored.

V. CONCLUSION AND FUTURE WORK

In this paper, we first reviewed some methods for 2-D human pose estimation on RGB images. It was proposed to use a CNN model for estimating keypoints to predict the actions of the martial arts instructor on a traditional martial arts video dataset that we contribute. Finally, we presented methods for evaluating the estimated keypoints and joints by 2-D and 3-D metrics. From the estimated joints human actions can be drawn about. Therefore, training martial arts by videos becomes easier and more explicit.

In martial arts videos, the actions (movements of body, arms, and legs) of a martial arts instructor are not always clear because there are many hidden joints. There are some cases where the joints are obscured in the videos that the model was unable to estimate. In the future, we will conduct studies to estimate obstructed joints. When there are sufficient joints, it is possible to build a visual martial arts teaching model and to evaluate the performance of traditional martial arts representations.

REFERENCES

[1] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E. H. Zahzah, "Human Pose Estimation from Monocular Images: A Comprehensive Survey," *Sensors (Basel, Switzerland)*, vol. 16, no. 12, pp. 1–39, 2016.

[2] M. Rantz, T. Banerjee, E. Cattoor, S. Scott, M. Skubic, and M. Popescu, "Automated fall detection with quality improvement "rewind" to reduce falls in hospital rooms," *J Gerontol Nurs*, vol. 40, no. 1, pp. 13–17, 2014.

[3] R. IgualCarlos, M. Carlos, and I. Plaza, "Challenges, Issues and Trends in Fall Detection Systems," *BioMedical Engineering OnLine*, vol. 12, no. 1, p. 147–158, 2013.

[4] J. Kramer, M. Parker, D. Castro, N. Burrus, and F. Echtler, "Hacking the Kinect," *Apress*, 2012.

Figure 14. Estimated keypoints and joints on video frames showing chains of traditional martial arts actions.

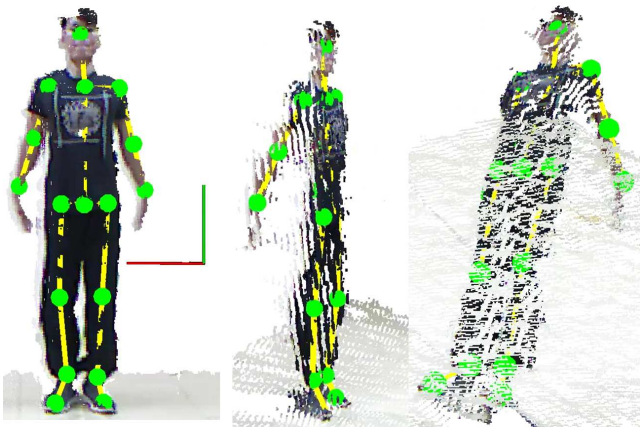


Figure 15. The color point cloud of a scene in a VNMA video, the estimated keypoints are colored green and the estimated joints are colored yellow.

[5] P. Felzenszwalb, D. Mcallester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[6] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *IEEE International Conf. Computer Vision*, 2011, pp. 723–730.

[7] E. M. Berti, A. J. S. Salmerón, and C. R. Viala, "4-Dimensional deformation part model for pose estimation using Kalman filter constraints," *International Journal of Advanced Robotic Systems*, vol. 14, no. 3, pp. 1–13, 2017.

[8] U. Rafi, J. Gall, and B. Leibe, "A semantic occlusion model for human pose estimation from a single depth image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 67–74.

[9] I. Sarandi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3D human pose estimation to occlusion," in *IEEE/RSJ International Conf. Intelligent Robots and Systems*, 2018.

[10] R. Girshick, "Fast R-CNN," in *International Conference on Computer Vision*, 2015, pp. 1440–1448.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 91–99.

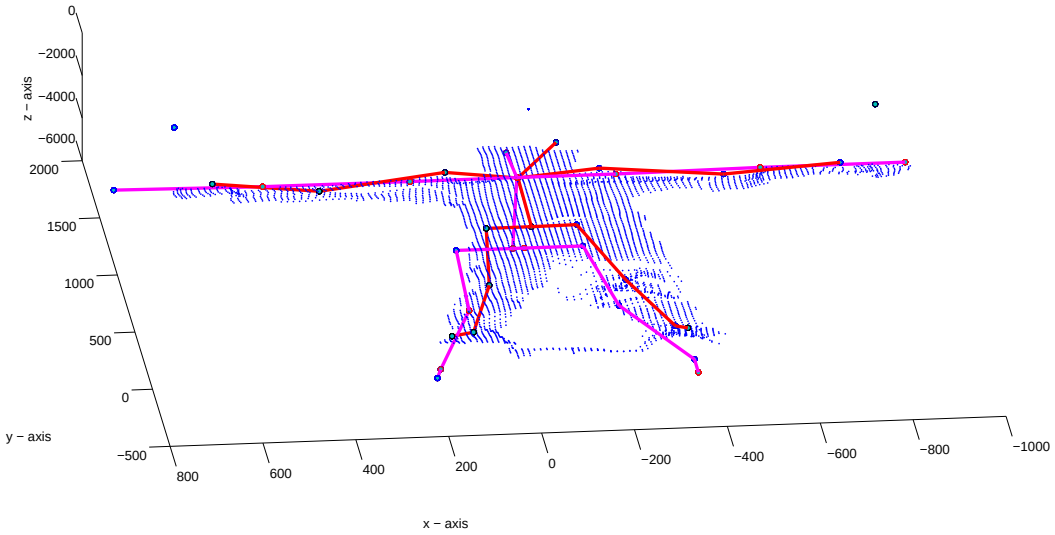


Figure 16. Estimated joints of the human body in 3-D space. The point cloud data of the human body is colored blue. Red segments are the ground-truth joints and magenta segments are the estimated joints.

Figure 17. Estimated keypoints and joints on video frames showing a chain of martial art poses: body joints are colored red, left hand joints are colored green, right hand joints are colored yellow, left foot joints are colored blue, and right foot joints are colored cyan.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Comp. Vision and Pat. Recognition*, 2016, pp. 779–788.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.

[14] D. Osokin, "Real-time 2D multi-person pose estimation on CPU: Lightweight openpose," *ArXiv*, 2018.

[15] K. Brown, "Stereo human keypoint estimation," *Stanford University*, 2017.

[16] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 350–359.

[17] "OpenPose," [Accessed 23 April 2019]. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[18] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity field," 2017, pp. 7291–7299.

[19] B. Nicolas, "Calibrating the depth and color camera," 2018, [Accessed 10 January 2018]. [Online]. Available: http://nicolas.burrus.name/index.php/Research/KinectCalibration

[20] T.-Y. Lin, Y. Cui, G. Patterson, M. R. Ruggero, L. Bourdev, R. Girshick, and P. Dollar, "COCO 2016 Keypoint Detection Task," [Accessed 17 April 2019]. [Online]. Available: http://cocodataset.org/#keypoints-2016

[21] W. Zhang, Z. Liu, L. Zhou, H. Leung, and A. B. Chan, "Martial arts, dancing and sports dataset: a challenging stereo and multi-view dataset for 3D human pose estimation," *Image and Vision Computing*, vol. 61, pp. 22–39, 2017.

[22] T. B. Dinh, "Bao ton va phat huy vo co truyen Binh dinh: Tiep tuc ho tro cac vo duong tieu bieu," 2017, [Accessed 4 April 2019]. [Online]. Available: http://www.baobinhdinh.com.vn/viewer.aspx?macm=12&macmp=12&mabb=88043

[23] ——, "Ai ve Binh Dinh ma coi, con gai Binh Dinh bo roi di quyen," 2019, [Accessed 4 April 2019]. [Online]. Available: http://www.seagullhotel.com.vn/du-lich-binh-dinh/vo-co-truyen-binh-dinh-5

[24] Chinese, "Chinese Kung Fu (Martial Arts)," 2019, [Accessed 4 April 2019]. [Online]. Available: https://www.travelchinaguide.com/intro/martial_arts/

[25] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *IEEE International Conference on Computer Vision*, 2013, pp. 3487–3494.

[26] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited people detection and articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1014–1021.

[27] X. Tao and Z. Yun, "Fall prediction based on biomechanics equilibrium using Kinect," *International Journal of Distributed Sensor Networks*, vol. 13, no. 4, 2017.

[28] J. Tompson, R. Goroshin, A. Jain, Y. Lecun, and C. Bregler, "Efficient object localization using convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[29] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[30] "ECCV 2018 Joint COCO and Mapillary Recognition," [Accessed 18 April 2019]. [Online]. Available: http://cocodataset.org/#home

[31] "MSCOCO Keypoints Challenge 2017)," [Accessed 18 April 2019]. [Online]. Available: https://places-coco2017.github.io

[32] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[33] S.-e. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[34] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 3686–3693.

[35] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, and M. P. I. Informatics, "PoseTrack: A benchmark for human pose estimation and tracking," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 5167–5176.

[36] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[37] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[38] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3D pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5689–5698.

[39] H.-s. Fang, Y. Xu, W. Wang, X. Liu, and S.-c. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," in *AAAI Conference on Artificial Intelligence*, 2018.

[40] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3D human pose estimation : A review of the literature and analysis of covariates," *Computer Vision and Image Understanding*, vol. 152, no. 7, pp. 1–20, 2016.

[41] "How to use random sample consensus model," 2014. [Online]. Available: http://pointclouds.org/documentation/tutorials/random_sample_consensus.php

[42] "Opencv library," 2018, [Accessed 19 April 2019]. [Online]. Available: https://opencv.org/

[43] "Kinect for Windows SDK v1.8," 2012, [Accessed 18 April 2019]. [Online]. Available: https://www.microsoft.com/en-us/download/details.aspx?id=40278

[44] "International Research Institute MICA," 2019, [Accessed 19 April 2019]. [Online]. Available: http://mica.edu.vn/

[45] Z. X, "A study of microsoft Kinect calibration," George Mason University, Tech. Rep., 2012.

[46] J.-Y. B., "Camera calibration toolbox for Matlab," 2019, [Accessed 19 April 2019]. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[47] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, p. 4, 2010.

[48] COCO, "Observations on the calculations of COCO metrics," 2019, [Accessed 24 April 2019]. [Online]. Available: https://github.com/cocodataset/cocoapi/issues/56

[49] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person pose estimation," [Accessed 23 April 2019]. [Online]. Available: https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation

[50] Tensorflow, "Tf-pose-estimation," [Accessed 23 April 2019]. [Online]. Available: https://github.com/ildoonet/tf-pose-estimation

[51] H. Wang, W. P. An, X. Wang, L. Fang, and J. Yuan, "Magnify-Net for multi-person 2D pose estimation," in *IEEE International Conference on Multimedia and Expo*, July 2018, pp. 1–6.

[52] caffe2, "Caffe2-pose-estimation," [Accessed 23 April 2019]. [Online]. Available: https://github.com/eddieyi/caffe2-pose-estimation
[53] "Chainer realtime multi-person pose estimation," [Accessed 23 April 2019]. [Online]. Available: https://github.com/DeNA/Chainer_Realtime_Multi-Person_Pose_Estimation
[54] mxnet, "Reimplementation of human keypoint detection in mxnet," [Accessed 23 April 2019]. [Online]. Available: https://github.com/dragonfly90/mxnet_Realtime_Multi-Person_Pose_Estimation
[55] "MatConvNet realtime multi-person pose estimation," [Accessed 23 April 2019]. [Online]. Available: https://github.com/coocoky/matconvnet_Realtime_Multi-Person_Pose_Estimation
[56] "CNTK realtime multi-person pose estimation," [Accessed 23 April 2019]. [Online]. Available: https://github.com/Hzzone/CNTK_Realtime_Multi-Person_Pose_Estimation
[57] L. Bo and C. Sminchisescu, "Twin Gaussian processes for structured prediction," *International Journal of Computer Vision*, vol. 87, no. 1-2, p. 28, 2010.

Nguyen Tuong Thanh received B.E. degree from Hanoi University Science and Technology in 2002 in Electronics and Telecommunications; He received M.E. degree in Electronic Engineering, University of Transport and Communications. He is now PhD student in Electronic Engineering, Hanoi University of Science and Technology. Currently, he is working at the Faculty of Engineering and Technology, Quy Nhon University. His research interests include computer vision, image processing 2-D, 3-D machine leaning, deep learning.

Le Van Hung received M.Sc. degree at Faculty Information Technology- Hanoi National University of Education (2013). He received PhD degree at International Research Institute MICA HUSTC-NRS/UMI - 2954 - INP Grenoble (2018). Currently, he is a lecture of Tan Trao University. His research interests include Computer vision, RANSAC and RANSAC variation and 3-D object detection, recognition, machine leaning, deep learning.

Pham Thanh Cong received M.Sc. degree at Electronics and Telecommunications, Hanoi University of Technology in 1998. He received PhD degree at Electronics and Telecommunications, Turin Polytechnic University, Italy in 2010. Currently, he is a lecturer of institute of Electronics and Telecommunications, Hanoi University of Science and Technology. His research interests include Super high frequency technology, antennas, telecommunication systems.

INFORMATION FOR AUTHORS

Research and Development on Information and Communication Technology – *The MIC Journal of ICT Research*, or *ICT Research* in short – is a scientific publication of the Journal of Information and Communication, published by Vietnam Ministry of Information and Communications (MIC) since 1999. ICT Research is peer-reviewed, open-access, and published four issues a year: two in its *English Edition* (ISSN: 1859-3534), and two in its *Vietnamese Edition* (ISSN: 1859-3526) (Chuyên san Các công trình nghiên cứu phát triển Công nghệ Thông tin và Truyền thông).

The purpose of ICT Research is to provide a forum for researchers and professionals to disseminate original and innovative ideas in the fields of information technology, communications, and electronics in Vietnam and worldwide. Its Editorial Board is composed of renowned scientists from research institutions throughout Vietnam and other countries.

AUTHOR GUIDELINES

Authors are encouraged to follow good practice of technical paper writing, such as the guidelines on "How to write for technical periodicals & conferences" of the Institute of Electrical and Electronics Engineers.

Submissions must be made online on the website of ICT Research [https://ictmag.vn/ict]. Authors should format the manuscript as closely as possible to the ICT Research manuscript template.

By submitting a manuscript, the authors are required to ensure that the submission has not been previously published, nor is currently submitted for publication elsewhere. Submission also implies that the corresponding author has consent of all authors. The submission file must be in PDF format. To faciliate double-blind review, authors are requested to submit two versions of the manuscript, one with full information of all authors, the other without.

Submissions of revised manuscripts should follow the same guidelines as for the initial manuscript, plus with a document entitled "Response to reviewers' comments" which explains the modifications according to the comments of the anonymous reviewers.

Upon formal acceptance of a manuscript for publication, the authors are required to prepare the final manuscript in LaTeX, using the IEEE Transactions style.

PUBLICATION ETHICS

ICT Research is committed to following the rules, standards, and guidelines set forth by the Committee on Publication Ethics (COPE) in ensuring the publication integrity and in handling cases of misconduct. All involved parties (Editors, Authors, Reviewers, and Publisher) are expected to be aware of and to conform to the respective rules, standards, and guidelines, available on COPE website.

The Editorial Board of ICT Research will immediately screen all articles submitted for publication. All submissions we receive are checked for plagiarism by using online available tools. Any type of plagiarism is unacceptable and is considered unethical publishing behavior. Such manuscripts will be rejected.

PEER REVIEW PROCESS AND PUBLICATION

An area editor of ICT Research (to be considered as the Editor from the point of view of the Authors) is assigned, by the Technical Editor-in-Chief, to each submission for coordinating the review process. The Editor pre-screens the submission and determines whether it is appropriate to the interests of readers of ICT Research. If appropriate, the Editor then initiates the review process. Each submission is then reviewed by at least two Reviewers. The review process is doubly blind, that is, identities of authors and reviewers are not revealed to each other.

Review decisions are: *(A) Accept submission* (i.e., no changes required), *(B1) Revision required* (i.e., minor revision), *B2) Resubmit for review* (i.e., major revision, the submission needs to be re-worked with major changes and is then subject to a second round of review), and *(C) Decline submission* (i.e., reject). Each major round of review may take approximately 2 months. Should authors be requested by the Editor to revise the manuscript, the revised version should be submitted within 6 weeks for a major revision or 2 weeks for a minor revision. No more than 1 major revision and 2 minor revisions are allowed.

ICT Research is published with two issues a year in its English Edition (ISSN: 1859-3534). An accepted article will be immediately published online, with a DOI number, having been copyedited and laid out in compliance with ICT Research editing rules.

COPYRIGHT NOTICE

Upon acceptance for publication, transfer of copyright will be made to the Publisher of ICT Research, who guarantees that full content of the published article is freely distributed on the website of ICT Research. The copyright transfer gives the Publisher of ICT Research full authority to resolve any complaints of misuse or abuse (such as infringement or plagiarism) of the published article. The author has the freedom to redistribute and reuse the published article in any medium or format for any purpose, provided the original published article is properly cited. An article submission implies author agreement with this policy.