Lecture Notes in Electrical Engineering 874

Qilian Liang · Wei Wang · Xin Liu ·
Zhenyu Na · Baoju Zhang *Editors*

# Communications, Signal Processing, and Systems

## Proceedings of the 11th International Conference on Communications, Signal Processing, and Systems, Vol. 3

Springer

# Lecture Notes in Electrical Engineering 874

## Series Editors

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

**China**

Jasmine Dou, Editor (jasmine.dou@springer.com)

**India, Japan, Rest of Asia**

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

**Southeast Asia, Australia, New Zealand**

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

**USA, Canada**

Michael Luby, Senior Editor (michael.luby@springer.com)

**All other Countries**

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

Qilian Liang · Wei Wang · Xin Liu · Zhenyu Na · Baoju Zhang
Editors

# Communications, Signal Processing, and Systems

## Proceedings of the 11th International Conference on Communications, Signal Processing, and Systems, Vol. 3

Springer

*Editors*
Qilian Liang
Department of Electrical Engineering
University of Texas at Arlington
Arlington, TX, USA

Xin Liu
Dalian University of Technology
Dalian, China

Baoju Zhang
College of Electronic and Communication Engineering
Tianjin Normal University
Tianjin, China

Wei Wang
Tianjin Normal University
Tianjin, China

Zhenyu Na
School of Information Science and Technology
Dalian Maritime University
Dalian, China

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-99-2361-8 ISBN 978-981-99-2362-5 (eBook)
https://doi.org/10.1007/978-981-99-2362-5



This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Contents

# Input Design Analysis for the Capacity of Finite Impulse Response Actuator

Deseng Zhang[1], Dazhuan Xu[1(✉)], Boyu Hua[1], and Junwei Bao[2]

[1] College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China
{dszhang,xudazhuan,byhua}@nuaa.edu.cn
[2] College of Physics, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China
broadenway@nuaa.edu.cn

**Abstract.** Mutual information (MI) is usually a computational heavy work when used as optimal metric in control system. In this paper, the modeling of Finite Impulse Response (FIR) system with entropy-rate framework is shown. A practical signal sampling algorithm to calculate the MI in the case of probability density function (PDF) with regenerative property is given. The MI calculation method is based on random sampling, which can reduce the computational complexity. The simulation results show our entropy-rate framework can be employed in analysis for the capacity of FIR actuator.

**Keywords:** FIR Actuator · Information Theory · Entropy

## 1 Introduction

Information theory has been adapted to control problems in various aspects including optimal metric, bound constraint and system explanation. It is well known that information theory and cybernetics share in a common theoretical and engineering background, both of which are concerned with signals and dynamic systems. Most traditional control system analysis assumes that the observed signal is available to controller in its entire decision period [1]. In this paper, the information theory modeling of FIR actuator is a commonly used convolutional structure.

There exists numerous work using information theory to explain and resolve problems in optimal control [2,3]. Touchette and Lloyd [4] proved that the one-step reduction in entropy of the final state is upper bounded by MI between control variables and current state of the system. Minimum of the one-step reduction in entropy can be found by dynamic programming, but the computational complexity of Dynamic Programming (DP) grows exponentially with the number of time steps and control variables. In fact, a quantization problem arises when information is exchanged between the controller and the controlled dynamic object [7–9]. An actuator channel can only transmit a finite number of


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 1–8, 2023.
https://doi.org/10.1007/978-981-99-2362-5_1

states at any given time [10]. There is a trade-off between the fineness of the control process and the time delay required to send the data over the channel. J.Baillieul demonstrated in [12] that adequate data quantization necessitates a bigger channel capacity or more data transmission time. In [13], directed information quantity is adopted as a measure of information transmission in feedback control system. However, the specific expressions for distortion measures other than quadratic distortion is rarely studied and researchers paid little attention to signal distribution in non-Gaussian PDF cases. Recently, authors in [14] have conducted some fundamental limit research in control field using information theory. Information theory is adopted as the main mathematical tool to obtain generic bounds on the variance of estimation errors in time series analysis in [15,16].

In this paper, we will demonstrate how to derive the entropy-rate function about a given actuator with noise, which is based on traditional rate-distortion theory. Then we define the conditional entropy as a distortion metric in an actuator system. Finally, we derive and simulate the MI between system state and output.

This paper is organized as follows. Two preliminaries of our analysis process, entropy-rate function and Sampling A Posterior (SAP) method are introduced in Sect. 2. The modeling of FIR actuator and derivation of the posterior estimation are given in Sect. 3. We derive and calculate the MI between system state and output in Sect. 4. Simulation of a non causal track system is given in Sect. 5.

## 2 Preliminaries

Throughout the paper, we consider real-valued continuous random variables and random vectors, as well as discrete time stochastic processes. All random variables, random vectors, and stochastic processes are zero mean. Given a stochastic process $\{x_k\}$, we assume the excitation of the control system be a set of finite-length K-dimensional independent and identically distributed Gaussian vectors, the system is linear and has memory. The output signal is an N-dimensional Gaussian vector $\mathbf{Y}$.

While classical linear system theory uses frequency response theory to characterize noise and system transfer functions, modern control theory uses linear algebra and state space to establish dynamic equations. Specifically, we consider the basic system with state-space model given by

$$\begin{cases} \hat{\mathbf{x}}_k = \mathbf{x}_k + \mathbf{w}_k \\ \hat{\mathbf{y}}_k = \mathbf{A}\hat{\mathbf{x}}_k \end{cases} \tag{1}$$

where $\mathbf{A}$ is the FIR actuator matrix, $\hat{y}_k$ is real output at step k, $x_k$ is system state at step k. The impluse response sequence h(d) has the length of d. The sequence is then filled with zero to the length of $N_S = N_{cp} + d$, where $N_{cp}$ is the length of circulant prefix. The finite implues response $\mathbf{A}$ during the $i^{th}$ symbol, which is N $\times$ N toeplitz [17],

$$\mathbf{A} = \mathbf{R}_{cp}\mathbf{H}\mathbf{T}_{cp}$$
$$= \begin{bmatrix} h(0) & 0 & \dots & 0 & h(d-1) & \dots & h(1) \\ h(1) & h(0) & 0 & \dots & \dots & \dots & \dots \\ \dots & \dots & \ddots & \dots & \dots & \dots & h(d-1) \\ h(d-1) & \dots & \dots & \ddots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \ddots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \ddots & 0 \\ 0 & \dots & 0 & h(d-1) & \dots & \dots & h(0) \end{bmatrix}, \tag{2}$$

where $\mathbf{R}_{cp} = \begin{bmatrix} 0_{N\times(N_S-N)} & I_N \end{bmatrix}$ is $N \times N_S$, $\mathbf{T}_{cp} = \begin{bmatrix} 0_{N-Ncp} & I_{Ncp} \\ & I_N \end{bmatrix}$ is $N_S \times N$. $\mathbf{y}_k$ is defined as a sequence of designed inputs over some number of timesteps k, $\hat{\mathbf{y}}_k$ is the real actuator output. It is reasonable to assume that direct control over designed input $\mathbf{y}_k$ exists, but not over $\hat{\mathbf{y}}_k$. This is a common noise assumption for many robotic systems in which command input must pass through a lower-level controller. A prototypical example is the steering and throttle inputs for a car which are then used as set-point targets for low level servomotor controllers.

The MI calculation method we used in this paper is called SAP method. It is based on a random sampling of the posterior probability generated by each received signal. The statistical distribution of this estimate will naturally approximate the theoretical posterior probability distribution of the parameter to be estimated. The main constraint is that the PDFs must be regenerative, eg. Gaussian, Cauchy, Gamma, Poisson, Exponentional distribution. The sampling posterior probability estimation method is described in Algorithm 1. Firstly, generate the received signal vector, then match conditional PDF of received signal vector. The key step is generating an estimate by random sampling of the posterior PDF. Through this way the empirical entropy and the average estimation performance will be calculated.

## 3 FIR Actuator Model and Posteriori Estimation

Assuming that the channel is constant (slow fading) during one symbol period, its finite impulse response sequence in the $i_{th}$ symbol period sequence is

$$\begin{aligned} y(n) &= \hat{x}(n) * h(n) = \sum_{l}^{n} \hat{x}(l)h(n-l) \\ &= \hat{x}(n)h(0) + \sum_{l}^{n-1} \hat{x}(l)h(n-l), \end{aligned} \tag{3}$$

where $l = n - d + 1$.

**Algorithm 1:** Sampling from A Posterior Method

```
Data: received signal vector, posterior PDF,
prior entropy H_prior
Result: posterior entropy H, mutual information
initialization;
for j = 1; j ≤ sampling_iteration_num do
    Using random sampling method to get sample points from posterior PDF,
     eg. randsample() in matlab ;
    while sampling_num not end do
        H = H - log(sampling_point);
    end
    H = H / sampling_num;
    I = H_prior - H;
end
return I;
```

The FIR system state Eq. (1) can be extended as

$$\begin{cases} \mathbf{x}_k = A\mathbf{x}_{k-1} + B\mathbf{e}_{k-1} + \mathbf{v}_k \\ \mathbf{y}_k = C\mathbf{x}_k + D\mathbf{e}_k + \mathbf{w}_k \end{cases}, \tag{4}$$

where Gaussian noise $v_k$, $w_k$ are independent. Combine (4) with (1)(2), we give the whole system model, the notation $e_k$ denotes the input of FIR system, $y_k$ denotes the referred signal, $\hat{y}_k$ denotes the output signal. Based on (3), we write the state equation as

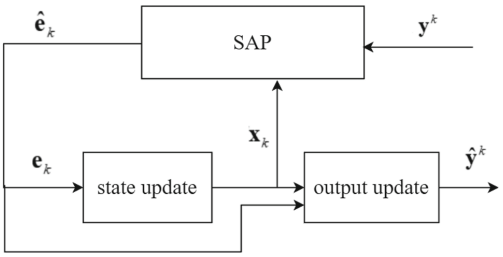



**Fig. 1.** System Input Design with SAP method

$$\begin{cases} p\left(\mathbf{y}_k \mid \mathbf{e}_k, \mathbf{x}_k\right) = f_w\left(\mathbf{y}_k - C\mathbf{x}_k - D\mathbf{e}_k\right) \\ p\left(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{e}_{k-1}\right) = f_v\left(\mathbf{x}_k - A\mathbf{x}_{k-1} - B\mathbf{e}_{k-1}\right) \end{cases}, \tag{5}$$

$$\begin{aligned} & p\left(e_k \mid y_k, y^{k-1}, e^{k-1}, x^{k-1}, x_k\right) \\ & = \frac{p\left(e_k, y_k, y^{k-1}, e^{k-1}, x^{k-1}, x_k\right)}{p\left(y_k, y^{k-1}, e^{k-1}, x^{k-1}, x_k\right)}. \end{aligned} \tag{6}$$

Based on Bayes' Rule, we continue to derive (6) and get the step estimation equation

$$\begin{aligned} &p\left(e_k \mid y_k, y^{k-1}, e^{k-1}, x^{k-1}, x_k\right) \\ &\quad = \frac{f_w f_v p\left(y^{k-1}, e^{k-1}, x^{k-1}\right) p\left(e_k\right)}{E_{p(e_k)}\left[f_w f_v p\left(y^{k-1}, e^{k-1}, x^{k-1}\right)\right]}. \end{aligned} \tag{7}$$

The next stages are derived in the same way. Other PDF with property of regenerativity can also do this SAP algorithm.

## 4 Mutual Information of FIR Actuator

This section will examine the MI-distortion relation of FIR Actuator. The main problem is the processing of MI in this model. SAP algorithm based on sampling method is given to quantitatively figure out the MI between the system output and state.

Based on the actuator model shown in (4), it is reasonable to assume that the stochastic process in both sides of (4) share the same statistical characteristic, thus the auto correlation of a stochastic process $\{\hat{x}_k\}$ is

$$\begin{aligned} R(d) &= E\left[\hat{\mathbf{x}}h(0)[\hat{\mathbf{x}}h(0)]^T\right] \\ &= \begin{bmatrix} R_{xx}(0) & R_{xx}(-1) & \cdots & R_{xx}(-d+1) \\ R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(-d+2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(d-1) & R_{xx}(d-2) & \cdots & R_{xx}(0) \end{bmatrix}, \end{aligned} \tag{8}$$

designed output $y \sim N(0, \Sigma_Y)$, distortion $(y - \hat{y}) \sim N(0, \mathbf{D})$. It is known that circulation matrix $\mathbf{A}$ in (2) can be diagonalized by Fourier matrices. Hence we have $\hat{\mathbf{Y}} = \Sigma\hat{\mathbf{X}}$. To the same effect, the conditional probability distribution representing the relationship between finite precision output and ideal output is

$$p(\hat{\boldsymbol{y}} \mid \boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^N |R_\delta|}} \exp\left\{-\frac{1}{2}(\hat{y} - \boldsymbol{y})R_\delta^{-1}(\hat{\boldsymbol{y}} - \boldsymbol{y})^T\right\}, \tag{9}$$

based on the rate-distortion, we write the conditional entropy as the distortion metric constraint for minimization

$$\frac{1}{2}\log(2\pi e)^N D = H(\mathbf{Y} \mid \hat{\mathbf{Y}}), \tag{10}$$

we have mutual information which indicates the information capacity of the FIR actuator,

$$\begin{aligned} I(\mathbf{y}; \mathbf{x}) &= \iint p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{y} \mid \mathbf{x})}{p(\mathbf{y})} d\mathbf{x}d\mathbf{y} = h(\mathbf{y}) - h(\mathbf{y} \mid \mathbf{x}) \\ &= \frac{1}{2}\ln(2\pi)^N |\boldsymbol{\Sigma}_\mathbf{Y}| + \frac{\sqrt{(2\pi)^N \left(|\boldsymbol{\Sigma}_\mathbf{Y}| + |\mathbf{D}|\right)}}{|\mathbf{D}|} \left[\ln \frac{\sqrt{(2\pi)^N \left(|\boldsymbol{\Sigma}_\mathbf{Y}| + |\mathbf{D}|\right)}}{(2\pi)^N |\boldsymbol{\Sigma}_\mathbf{Y}| |\mathbf{D}|} - 1\right]. \end{aligned} \tag{11}$$

By using SAP algorithm described in Sect. 3 to calculate MI, the results are shown in Fig. 3. Other PDFs with regenerativity such as Cauchy, Gamma, Poisson, Exponentional distribution also apply to this deduction.

## 5 Simulation of a Non Causal Track System

We carry out the system simulation based on (5). The FIR system has 1 Hz bandwidth and it functions as a track system with reference input $y$.

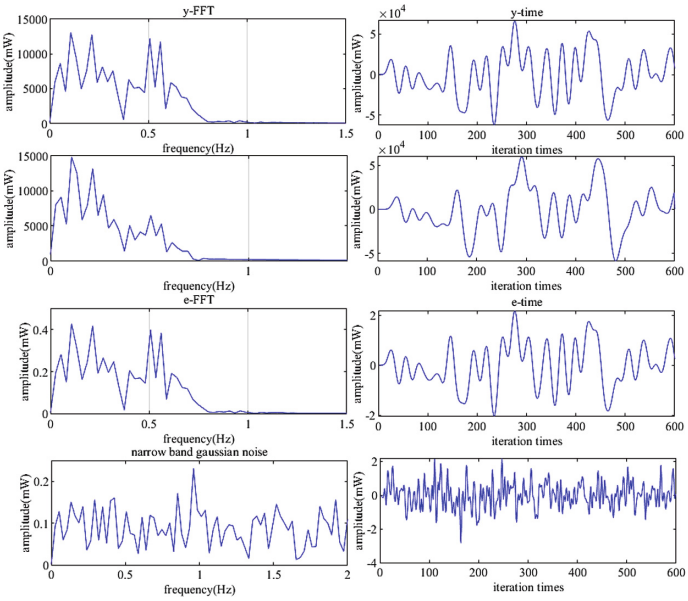



**Fig. 2.** Simulation of the track system

As can be shown in Fig. 2, the left column is the FFT of time series and the right column is the time domain series. The first row is the designed output $y$, and the second row is the actual output $\hat{y}$. The third row is designed input signal $e$ based on estimation Eq. (6) and the forth row is the noise $w$ in system equations. This track system is a lowpass filter 1 Hz bandwidth. The MI between the first row $y$ and the second row $\hat{y}$ is shown in Fig. 3. It can be observed in Fig. 3 that as the tracked frequency $y$ getting closer to the system's cutoff 1 Hz, tracking process performs worse. The worse performance is reflected by the decrease of MI $I(y;\hat{y})$ from 0.3 to 0.2.

MI result in Fig. 3 reflects the capacity of this track system. $I(y;\hat{y})$ tells us the information from $y$ to $\hat{y}$ in Fig. 1. As the input frequency exceeds the allowed entropy-rate relation, the signal tracking process will not keep up. When the

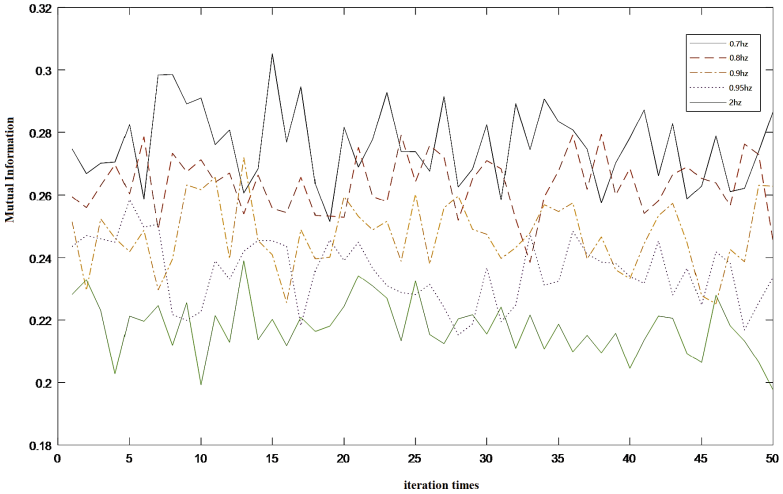



**Fig. 3.** Mutual Information $I(y;\hat{y})$ with different frequenciesies from 0.7 Hz 2 Hz

tracked frequency gradually 1 Hz, the MI decreases at the same time. The same phenomenon also exists in Mean Square Error (MSE), as input frequency getting closer to cutoff frequency, MSE between $y$ and $\hat{y}$ also decreases.

## 6 Conclusion

In this paper, we proposed an entropy-rate framework for the analysis of a FIR actuator model and gave the input estimation equation and posterior PDF recursive deduction. The simulation result about the MI of system is based on SAP method, and the calculation complexity of SAP method is reduced The MI metric can reflect the capacity of FIR actuator based on simulation results. Our entropy-rate relation can be a feasible framework to explain control system's capacity. Other PDFs with regenerativity such as Cauchy, Gamma, Poisson, Exponentional distribution also apply to this analysis framework.

**Acknowledgement.** This work was supported in part by the Key Project of Aeronautical Science Foundation of China under Grant NO. 2020Z073009001, in part by the open research fund of National Mobile Communications Research Laboratory, Southeast University, No. 2022D04, in part by the Future Network Scientific Research Fund Project, No. FNSRFP-2021-YB-04.

## References

1. Nair, G.N., Fagnani, F., Zampieri, S., Evans, R.J.: Feedback control under data rate constraints: an overview. Proc. IEEE **95**(1), 108–137 (2007)
2. Tatikonda, S., Mitter, S.: Control under communication constraints. IEEE Trans. Autom. Control **49**(7), 1056–1068 (2004)

3. Borkar, V.S., Mitter, S.: LQG control with communication constraints. In: Communications, Computation, Control, and Signal Processing: A Tribute to Thomas Kailath. Kluwer, Norwell (1997)
4. Touchette, H., Lloyd, S.: Information-theoretic approach to the study of control systems. Physica A **331**(1), 140–172 (2004)
5. Bania, P.: Bayesian input design for linear dynamical model discrimination. Entropy **21**(4), 351 (2019)
6. Bania, P.: An information based approach to stochastic control problems. Int. J. Appl. Math. Comput. Sci. **30**(1), 23–34 (2020)
7. Zhu, Q., et al.: A novel 3D non-stationary wireless MIMO channel simulator and hardware emulator. IEEE Trans. Commun. **66**(9), 3865–3878 (2018)
8. Zhu, Q., Zhao, Z., Mao, K., Chen, X., Liu, W., Wu, Q.: A real-time hardware emulator for 3D non-stationary U2V channels. IEEE Trans. Circ. Syst. I: Reg. Pap. **68**(9), 3951–3964 (2021)
9. Zhu, Q., Bai, F., Pang, M., et al.: Geometry-based stochastic line-of-sight probability model for A2G channels under urban scenarios. IEEE Trans. Ant. Propag. **70**, 5784–5794 (2022)
10. Li, K., Baillieul, J.: Robust quantization for digital finite communication bandwidth (DFCB) control. IEEE Trans. Autom. Control **49**(9), 1573–1584 (2004)
11. Wong, W.S., Brockett, R.W.: Systems with finite communication bandwidth constraints. II. Stabilization with limited information feedback. IEEE Trans. Autom. Control **44**(5), 1049–1053 (1999)
12. Baillieul, J.: Matching conditions and geometric invariants for second-order control systems. In: Proceedings of the 38th IEEE Conference on Decision and Control, vol. 2, pp. 1664–1670 (1999)
13. Charalambous, C.D., Kourtellaris, C.K., Tzortzis, I.: Information transfer of control strategies: dualities of stochastic optimal control theory and feedback capacity of information theory. IEEE Trans. Autom. Control **62**(10), 5010–5025 (2017)
14. Nekouei, E., Tanaka, T., Skoglund, M., et al.: Information-theoretic approaches to privacy in estimation and control. Ann. Rev. Control **47**, 412–422 (2019)
15. Fang, S., Skoglund, M., Johansson, K.H., Ishii, H., Zhu, Q.: Generic variance bounds on estimation and prediction errors in time series analysis: an entropy perspective. In: 2019 IEEE Information Theory Workshop (ITW), pp. 1–5 (2019)
16. Fang, S., Zhu, Q.: Fundamental limits of obfuscation for linear gaussian dynamical systems: an information-theoretic approach. In: 2021 American Control Conference (ACC), pp. 4574–4579 (2021)
17. Grenander, U., Szego, G.: Toeplitz Forms and Their Applications. University of California Press, Berkeley (1958)

# Research on the Improved Design and Implementation of Signal Processing and Detection Algorithm for High Speed Moving Target

Xinyu Zhang(✉), Shangyue Wang(✉), Ran Zhang(✉), and Pengcheng Zhao(✉)

Beijing Institute of Space Long March Vehicle, Beijing, China
13261255685@163.com, wangshy310@foxmail.com, 1fxranzhang@163.com, zhaopengcheng0958@163.com

**Abstract.** Signal processing and detection technology is one of the important directions of radar signal processing. In order to track high speed moving target stably, the selection of reasonable signal detection technology is particularly important in radar signal processing. In this paper, the common filter of constant false alarm detection in signal detection algorithm is analyzed. The relationship among detection probability false alarm probability and signal to noise ratio in signal detection is introduced. Then, a signal detection method for high speed moving target is proposed. Based on the real data of high speed moving target, several detection methods in radar signal processing algorithm are compared and analyzed, and the advantages and disadvantages of different methods are described.

**Keywords:** Radar Signal Processing · Target Detection · Constant False-Alarm Rate

## 1 Introduction

As an important tool for detecting targets, radar plays an important role in the military field [1]. The targets detected by radar are mainly moving targets [2]. When the target is detected, the threshold is used to judge whether the target exists. When the sampling point amplitude exceeds the threshold, the target is considered to exist; otherwise, the target is considered not to exist. Naiman-Pearson criterion refers to the decision criterion that maximizes the detection probability under the given false alarm probability [5]. In the real environment, both clutter and noise change in real time. If the threshold is fixed, it cannot adapt to the real-time changes of clutter and noise. Therefore, the strength and variation of clutter must be analyzed in real time. In this way, the clutter environment information can be used in threshold detection. The method of spatial constant false alarm rate detection processing mainly estimates the noise parameters by the amplitude value of the unit near the sampling unit to be detected. In radar signal detection, the method of spatial constant false alarm is generally used to detect the target.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 9–15, 2023.
https://doi.org/10.1007/978-981-99-2362-5_2

## 2 Spatial Constant False Alarm Detection

The spatial constant false alarm technique takes different sampling points of fast time of radar echo as different range units to detect the target. The threshold value is set to determine whether the signal of the current range unit is the target signal. If the amplitude value of the current range unit exceeds the threshold, it is considered that there is a target on the current range unit. Otherwise, the signal on the current range unit is considered as clutter signal and noise signal. The spatial constant false alarm adaptively adjusts the threshold by the amplitude of the unit near the current range unit.

Unit average constant false alarm detection determines the detection threshold according to the amplitude average value of units near the unit to be detected. This approach is based on two assumptions. First, it is assumed that the amplitudes of the sampling points near the sampling unit to be detected are independent and the statistical characteristics are consistent with those of the sampling points to be detected. Second, the assumption condition is that the sampling point near the sampling unit to be detected does not contain the target [3].

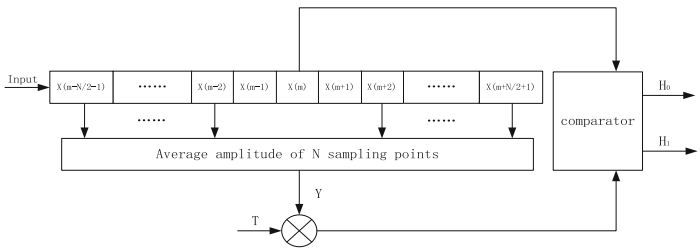



**Fig. 1.** Processing block diagram of unit average constant false alarm detection

The input in Fig. 1 is the signal after passing through the square law detector, where $X(m-1)$ and $X(m+1)$ are the amplitudes of the protection units, and the number of protection units can be modified according to the actual situation. It is mainly related to the number of units occupied by the target sidelobe. The protection unit is not involved in the estimation of the detection threshold. The remaining $N$ nearby sampled values are referred to as reference cells. The mean value $Y$ of the $N$ reference cells was multiplied by the threshold factor $T$. The product is used as the threshold value to compare with the current sampling point $X(m)$. This determines whether the current sampled value is the target. Then, get [3]:

$$P_{fa} = (1 + T')^{-N} \tag{1}$$

At this time, the false alarm probability $P_{fa}$ and detection probability $P_d$ are not related to the clutter power $\mu$, but only related to the threshold factor $T'$ and the reference unit $N$. When the reference unit $N$ is fixed, only the threshold factor $T'$ can be adjusted to achieve the desired false alarm probability.

The CFAR detector with small unit divided the sampling points near the current sampling point into two parts. Take the mean of the two parts and compare them. The

CFAR detector with small unit takes the product of the two parts with smaller mean value and the threshold factor as the threshold for judgment. They are shown in Fig. 2. In the figure, A and B are the mean values of sampling points in the first half and the second half, respectively.

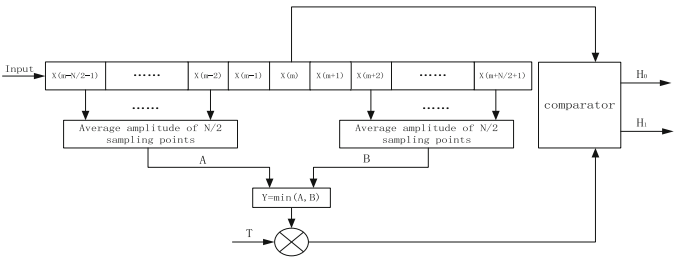



**Fig. 2.** Processing block diagram of unit selected small constant false alarm detection

The false alarm probability $P_{fa}$ can be expressed as follows [4]:

$$P_{fa} = 2\sum_{i=0}^{N-1} \frac{(N+i-1)!}{i!(N-1)!}(2+T)^{-N-i} \tag{2}$$

It can be seen that the false alarm probability is independent of the average clutter power. The detection probability under the same false alarm probability increases with the improvement of SNR and the detection probability is improved by changing the threshold factor under the same SNR condition. At the same time, the probability of false alarm also increases.

## 3 Design and Implementation of High Speed Moving Target Detection Algorithm

### 3.1 Fast Threshold and Slow Threshold Detection

When the radar detects the high speed moving target, because the detection speed of the target is very fast, the spatial background environment around the target changes dramatically. In the actual detection, the method of spatial constant false alarm detection is used to detect the target. Signal processing is controlled by the master control and can select fast threshold and slow threshold detection.

The detection method of fast threshold is protection unit 4. The small constant false alarm detector is selected as the unit of detection unit 32. According to the formula, the expression of the false alarm probability is:

$$P_{f_a} = 2\sum_{i=0}^{32-1} \frac{(32+i-1)!}{i!(N-1)!}(2+T)^{-32-i} \tag{3}$$

It can be seen from the above equation that the false alarm probability is negatively correlated with the radar threshold factor. The higher the threshold factor, the lower the false alarm probability. Otherwise, the lower the threshold factor, the higher the false alarm probability.

The change curve of detection probability with SNR under different false alarm probabilities is shown in Fig. 3.

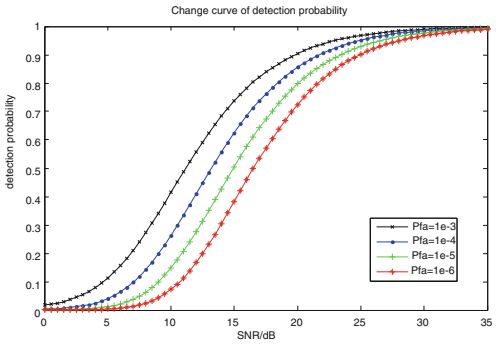



**Fig. 3.** Change curve of detection probability of fast threshold detection with SNR under different false alarm probabilities

The slow threshold detection method selects a small constant false alarm detector for the detection unit 32 and then selects a small constant false alarm detector for the 32-point detection unit at the left and right ends. For any point of the sampling sequence, the detection can be equivalent to four segments of 32 points of the selected small detector. The processing block diagram is as follows (Fig. 4):

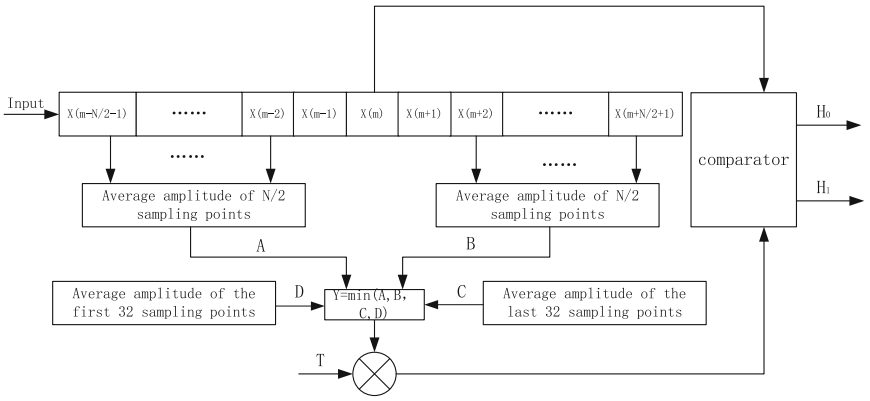



**Fig. 4.** Processing block diagram of fast threshold detector

The fast threshold and slow threshold of high speed moving target are detected respectively. The effect comparison is as follows (Fig. 5):

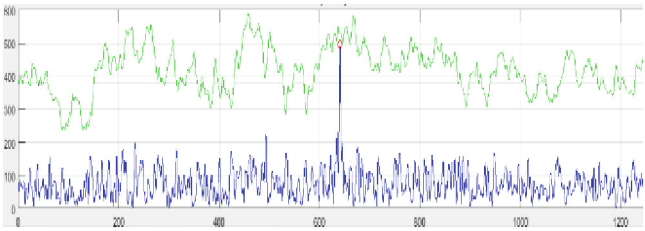



**Fig. 5.** Comparison of fast threshold and slow threshold detection effects

In the figure, under the condition of the same threshold factor, the fast threshold and slow threshold detectors are respectively used to detect the target. Because the slow threshold increases the small operation at both ends than the fast threshold, the false alarm probability is higher under the condition of the same threshold factor, which brings higher detection probability.

### 3.2 Improved Fast Threshold CFAR Detection Method

In the actual detection scene, when there is no interference, there are more scatterers in the tail of the moving target. At this point, under the slow threshold, a large number of targets will be detected. Under the fast threshold, the small target is occluded by the large target. In view of the above two phenomena, the fast threshold detection method can be improved.

The modified fast threshold algorithm flow is as follows:

1) Set the protection interval to 4, and set the left and right points respectively. And then choose small to generate noise sequence A.
2) Set the protection interval to M2. Take a point left and a point left. And then choose small to generate noise sequence B. The value of M2 depends on the interference harmonic period L and the waveform.
3) Noise sequence A and B are chosen small to generate the final noise sequence
4) Make 32-point smoothing for the selected noise sequence. The final threshold is then multiplied by the CfarK value issued by the master control, plus the DC-offset CfarB.
5) Target detection is performed according to the threshold and signal modulus value
6) Delete the target whose SNR is less than CfarK (reduce false alarm rate)
7) The target that has passed the threshold of 3 or more points is retained. The target with less than three points but SNR greater than 10 is also retained (to reduce the probability that the small target will not cross the threshold and be deleted because the large target occludes the small target). The remaining targets are deleted.

By setting two different protection intervals, the probability of using harmonics or targets as noise is reduced as much as possible. After 32 point smoothing, the noise sequence is closer to the real noise floor.

Its detection block diagram is shown as follows (Fig. 6):

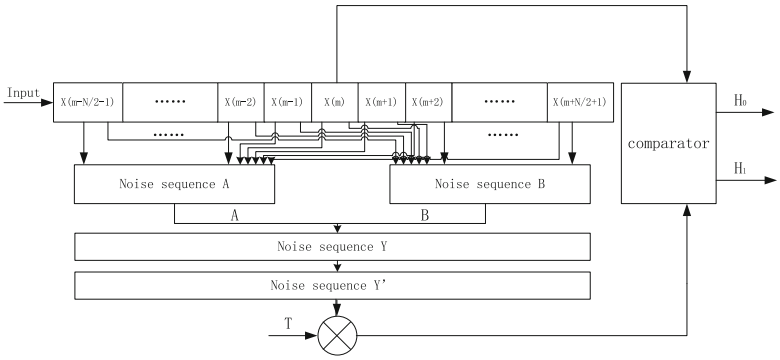



**Fig. 6.** Processing block diagram of the improved fast threshold detector

Where, the expression of noise sequence A is:

$$A(m) = \min(X(m-4), X(m+4)) \tag{4}$$

The expression of noise sequence B is:

$$B(m) = \min(X(m-M2), X(m+M1)) \tag{5}$$

The expression of noise sequence Y is:

$$Y(m) = \min(A(m), B(m)) \tag{6}$$

The expression of noise sequence Y' is:

$$Y'(m) = \frac{1}{32}\sum_{i=0}^{31} Y(m+i) \tag{7}$$

According to the probability density function of X, the probability density function of Y can be derived theoretically. Accordingly, the relationship between the false alarm probability of Y and the threshold factor is derived.

Three kinds of thresholds are used for target detection and comparison, as shown in the figure below (Fig. 7):

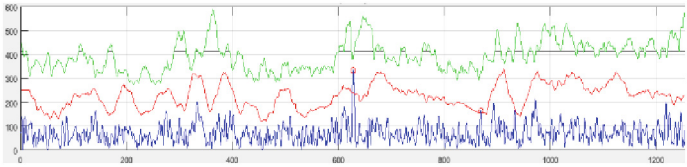



**Fig. 7.** Comparison of detection effects of fast threshold, slow threshold and improved fast threshold

In the figure, due to the different relationship between the false alarm probability and the threshold factor of different detectors, the false alarm probability generated

by detecting the target with the same threshold factor is also different. The detection probability of the improved fast threshold is the highest, and correspondingly its false alarm probability is relatively high. Therefore, the improved fast threshold not only detects the main target located in the center of the channel. It also targets a noise-like signal at the far end. The main advantage of the improved fast threshold detector is that it can be better adapted to specific detection scenarios in combat. In order to achieve the same false alarm probability, the threshold factor of the improved fast threshold detector should be adjusted to some extent.

## 4 Conclusion

Firstly, this paper discusses the unit average of the signal detection algorithm, the selection of small left and right units, the selection of three kinds of constant false alarm detectors, and the relation curve between the SNR and the detection probability of various detectors under different false alarm probabilities. Finally, the fast threshold and slow threshold detection methods and the implementation process of the improved fast threshold detection are described. According to the actual detection data, different detection methods are compared and analyzed.

## References

1. Xiang, J., Zhang, M.: Radar System, pp. 179–184. Publishing House of Electronics Industry, Beijing (2001)
2. Zuo, Q., Xu, G., Ma, L., et al.: Introduction to Radar System, pp. 9-10 Publishing House of Electronics Industry, Beijing (2006)
3. Jian, G., Yong, H., You, H.: MIMO array radar CFAR detector based on adaptive pulse compression-capon filter. Sci. China Inform. Sci. **10**, 10–11 (2011)
4. Yibin, R.: Key Technology Research and Signal Processing System Design of Portable Battlefield Reconnaissance Radar. Nanjing University of Science and Technology, Nanjing (2016)
5. Shen, F.: Signal Statistical Analysis and Processing, pp. 45–49. University of Science and Technology of China Press, Hefei (2001)

# A Transformer-Based Network for Hyperspectral Image Classification

Jizhen Yu and Zhengtao Li(✉)

School of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
1179970008@qq.com

**Abstract.** For hyperspectral classification, many pixel-based approaches have been developed to deal with multidimensional hyperspectral data, in addition to the orderless vector-based feature extraction, the sequence-based feature representation approaches also prove its effect on hyper-spectral classification. Based on the sequence-based data structure, we borrow the Transformer architecture from the knowledge of NLP (Natural Language Processing) to analyze hyperspectral pixels as sequential data and then decide information categories via network reasoning. In the proposed model, after first obtaining the spatial features of single band, we can obtain the spectral embeddings and position embeddings in hyperspectral sequence; Secondly, with adopting multilayer encoder module derived from Transformer which has powerful sequence expression ability, we can dig depth correlation between sequence data; Finally, a feedforward layer and a softmax layer are used to predict class categories. The experimental results on two publicly accessible images demonstrate that our proposed approach can obtain competitive performance compared with other state-of-the-art methods.

**Keywords:** Hyperspectral classification · sequential correlation · self-attention · transformer

## 1 Introduction

Hyperspectral signatures record hundreds of reflectance values over a large domain of electromagnetic spectrum, and based on hyperspectral images, a widespread classification applications have been implemented, such as: the monitoring of land changes, scene interpretation, resource management and disaster assessment.

Recently, numerous types of supervised classification approaches have been researched in the literature, including support vector machines [6, 11], decision trees [5], random forests [7, 8] and sparse representation [4, 16]. And all these vector-based classification methods do not explicitly model the spectral correlation of hyperspectral pixels. As a powerful feature extraction technology, deep learning has achieved great success in hyperspectral classification. Convolutional neural networks (CNNs), which are introduced for hyperspectral data classification [3, 10], have shown its powerful ability to extract spatial and spectral features for category discrimination. In the


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 16–25, 2023.
https://doi.org/10.1007/978-981-99-2362-5_3

above methods, hyperspectral pixels can be considered as points with orderless multi-dimensional features, in addition, the pixels can also be viewed as a set of orderly and continuing spectra sequences in the spectral space [13, 17] which can be modeled with RNN-like networks. Therefore, how to extract the correlations of sequence information has important reference significance for hyperspectral classification.

In 2017, Google proposed the transformer architecture to model the sequential correlations, and this architecture effectively exploits multi-layer self-attention modules to mine correlations between sequences. The multi-head self-attention modules abandon the RNN-like network architecture which transfer sequence information recursively, but adopt a direct sequence correlation extraction method. With self-attention module, we can directly calculate long-range dependencies for any position points in sequence, so it can shorten the length of dependency paths between different positions in sequence. Relying on powerful mining capabilities of sequence correlations, the architecture has achieved the state of the art at the time in many NLP (Natural Language Processing) tasks [1, 14], such as machine translation, document classification, sentence summarization, machine comprehension and so on. The self-attention module imitates the attention mechanism of human brain to perform information selection, the attention mechanism is a approach of rapidly screening out high-value information from a large amount of information. In addition to the self-attention module, the attention mechanism has achieved many successful applications in other fields, including image captioning [9, 15] and sequence learning [2, 12]. In this paper, we introduce the transformer architecture into the classification networks to model the spectral correlations. In this context, we propose a novel transformer based networks to solve the multiclass classification for hyperspectral imagery, and the networks can be described as the following three parts: first, the hyperspectral pixel cube is divided into spectral sequence in the spectral dimension, and the single-band spectral image is transformed through CNN layers into spectral embedding; second, the spectral embeddings are feed into the multi-layer encoder modules derived from the transformer architecture, and we can obtain the correlation expression of spectral sequence. Benefiting from deep mining of sequence correlation, finally, the hyperspectral class can be predicted more accurately through the feedforward layer and the softmax layer. The remainder of this paper is organized as follows. In the next section, the classification method based on transformer architecture is described. Section 3 illustrates the study data sets and the results of experiment. Finally, conclusion of this work is given in Sect. 4.

## 2 Proposed Methodology

### 2.1 Transformer-Based Architecture for Classification

The proposed model adopts the transformer architecture derived from NLP to model the spectral sequence data, and the overall block diagram for hyperspectral classification is shown in Fig. 1. The model architecture includes the following parts: the spectral feature embedding module, position embedding module, the multi-layer Transformer-encoder module and a feedforward layer and a soft-max output layer. In this section we will introduce these parts as follows.

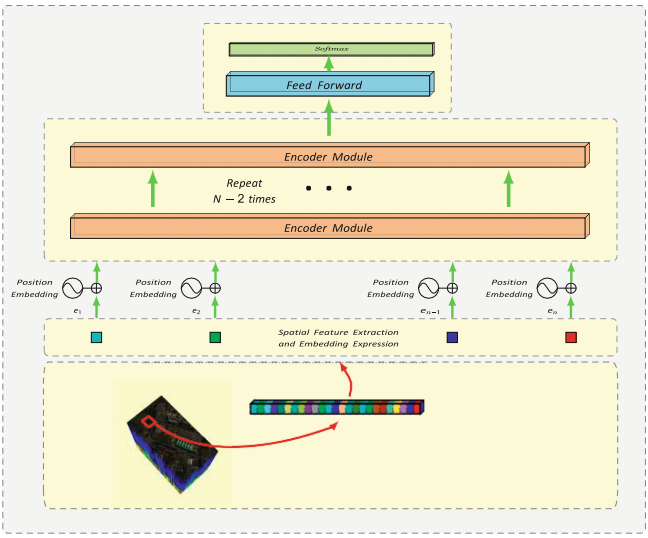



**Fig. 1.** Classification scheme for hyperspectral image based on CNN-Transformer sequence correlation extraction. The lower part of the figure indicates spectral feature embeddings and position embeddings; and the upper part expresses the sequence feature extraction with multi-layer encoder modules.

**Spectral Feature Embedding and Position Embedding** For The Pixel-Based. hyperspectral classification, each pixel can be expressed through a patch of size ($w$, $h$, $n$), and the patch cube is a neighbor window surrounding the pixel. The transformer-encoder module takes a 1-dimensional vector sequence as input embeddings, so we divide the patch cube along the spectral dimension to obtain the image sequence $n$ ($w$, $h$), and then we use CNN layers to extract the spatial feature of each band image of shape ($w$, $h$). Next, we transform the spatial features of spectral band image into 1-dimension vector. Finally, we can get the spectral embedding sequence $W_E = [e_1, ..., e_n]$, and the length of 1-dimension embedding is $l$.

Because the transformer architecture has no recurrence and convolution, the information of the order in the sequence must be injected into the model. The position embeddings of sequence indicate the relative or absolute position information in the sequence, and position embedding sequence $W_P$ can be described by "positional encodings". The positional encodings have the same vector length $d_{model}$ as the spectral embeddings $l$. Here, we introduce the functions in [14] to define positional embedding:

$$\begin{aligned} PE_{(p,2i)} &= sin\left(p/10000^{2i/d_{model}}\right) \\ PE_{(p,2i+1)} &= cos\left(p/10000^{2i/d_{model}}\right) \end{aligned} \tag{1}$$

where $p$ is the position and $i$ indicates the dimension of embedding.

**Multi-layer Transformer-Encoder Module** Similar to the previous rnn-like model for modeling sequence information, the transformer architecture [14] which is proposed by

Google in 2017 has shown powerful ability in the field of NLP. Different from the recursive RNN and LSTM structure, the transformer model adopts an all-attention architecture which is called self-attention module to deal with sequential encoding mechanism. The self-attention module not only has the excellent expression ability of the relationship between sequence information, but also has advantages in terms of parallel ability. And benefiting from the parallel computation of self-attention module which handles entire input sequence at a time for training, we can obtain faster training speed than the recursive rnn-like model.

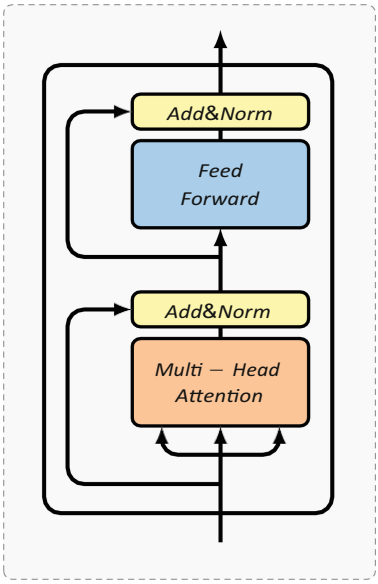



**Fig. 2.** The encoder module used in the classification model.

For the sequence modeling and sequence classification, they are different sequence task paradigms. For sequence modeling, the most typical is the language modeling. For an sentence sequence $U = \{u_1, ..., u_n\}$, the standard language modeling objective can be represented to maximize the likelihood: $L(U) = \sum_i logP(u_i u_{i-k}, ..., u_{i-1}; \Theta)$. When using the transformer architecture to construct the language model, the parallel training of sentence sequences can be achieved through forward masking mechanism which acts on the encoder and decoder. The mask mechanism shield the future word embedding, so that only leftward information of the sequence is visible, while rightward information is blind. For sequence classification task, the model output is the classification label rather than the sequence. And the label is visible to the entire sequence information, so the classification model can utilize the whole sequence information. Therefore, in the sequence classification task, we can only take advantage of the encoder module from transformer and discard the decoder module and mask mechanism.

The encoder structure can be shown in Fig. 2. And in the proposed model, we use the multi-layer transformer-encoder module to express sequence correlation. As shown

in Fig. 2, there are two sub-layers in encoder module: multi-head self-attention is the first part, and the second part is position-wise fully connected feed-forward network. In addition, there are residual connections and layer normalizations in each encoder module.

This module takes spectral embeddings and position embeddings as input, so $h_0$ can be formed as:

$$h_0 = W_E + W_P \tag{2}$$

where spectral embedding $W_E$ is the sequence features $[e_1, ..., e_n]$ which includes the entire bands, and position embedding sequence $W_P$ represents the position embeddings for each band position $[PE_1, ..., PE_n]$. The mulyi-layer encoder module can be described as follows:

$$h_l = transformer\_encoder(h_{l-1}) \forall i \in [1, n] \tag{3}$$

where $n$ is the number of layers.

**Output Layer for Classification.** To predict the type label, we add feedforward layer and softmax layers to the network. For the classification task, we have a labeled dataset, where each pixel-based patch cube has a label $y$. We pass the input sample through embedding layer and multi-layer transformer-encoder module to get the last encoder's activation $h_l$, then add feedforward layer with parameters $W_y$ and softmax layer to predict $y$:

$$P(y|G_1, ..., G_n) = softmax\big(h_l W_y\big) \tag{4}$$

By the formula, we get the objective to maximize:

$$L(C) = \sum_{(G,y)} logP(y|G_1, ..., G_n) \tag{5}$$

In the above formula, the conditional probability $P$ is modeled with neural networks, and the model can be trained using stochastic gradient descent.

## 3 Experimental Results

### 3.1 Description of Datasets

In this section, we experimented and evaluated the proposed framework with two well-known hyperspectral datasets, including University of Pavia and Indian Pines. In particular:

**Indian Pines Dataset** It describes a region in North-western Indiana, this dataset was obtained by AVIRIS sensor whose spectral acquisition range is from 400 nm to 2500 nm. After excluding water absorption bands, the dataset contains 200 available spectral reflectance bands. And across the region of 145 $\times$ 145 pixels, we have 10249 labeled pixels for 16 different crop categories.

**University of Pavia Dataset** This dataset was collected by ROSIS-3 (reflective optics imaging spectrometer) over the University of Pavia (PU), Italy, and it contains 103 valuable bands in a spectral range from 430 nm to 860 nm. In the study region of 610 × 340 pixels, there are 42776 labeled pixels which cover 9 land cover types.

### 3.2 Experimental Design

In order to evaluate the effects of different classification models, a lot of evaluation criteria have been developed. In this paper, the criteria of Overall Accuracy (OA), Average Accuracy (AA) and Kappa coefficient have been introduced to evaluate the performance of various classification models. For a fair comparison, we randomly separate the experimental datasets into training and testing parts for all comparison methods, and we implement the experiments more than ten times with different initial training part. Finally, we record the average results of the measures into results table below.

In this paper, we conduct comparative experiments on several traditional and deep learning architecture classification methods, such as RF, SVM-RBF, CNN, CNN-GRU and CNN-Transformer architecture model. (1) RF-200: The random forest model is constructed with 200 decision trees; (2) SVM-RBF: We utilize the libsvm package to perform SVM classifier with RBF kernel, and fivefold cross-validation is taken to optimize the hyperplane parameters; (3) CNN: For the CNN-based classifier, we adopt a large window size of 15 × 15 for spatial feature extraction, and the subsequent two convolution layers and two pooling layers are adopted in the model, in which the convolution kernel are 5 × 5, 3 × 3 and the pooling kernel is 2 × 2; (4) CNN-GRU: The CNN-GRU model utilizes the convolution layers to extract spatial feature and produce the spectral feature embedding, and then use the GRU module to express the correlation of spectral sequence. This model has the same network architecture as CNN-Transformer, except for the difference between GRU and Transformer in sequence modeling;

(5) CNN-Transformer: The proposed CNN-Transformer model adopts CNN to extract the spatial features, and the Transformer architecture with multi-layer encoder module is used to express the correlation of spectral embeddings.

### 3.3 Classification Results Analysis

For hyperspectral classification, various deep models have achieved excellent classification result, CNN model handles the spatial correlation and spectral correlation directly, while the RNN-like models treat the hyperspectral pixel profile as a spectral sequence. The RNN-like models adopt recursive structure to express the correlation in the sequence, which limits the ability to express sequence information. In order to improve the sequence expression ability of RNN-like models, a new activation function PRetanh (Parametric Rectified tanh) [13] has been proposed to handle this problem. Different from the PRetanh activation function which is constructed for RNN-like model to handle long-term hyperspectral sequence correlation, we utilize the Transformer architecture to calculate the spectral correlations directly on hyperspectral profile, and we can obtain the better classification results compared with CNN and RNN-like models.

In experiments, the traditional RF and SVM-RBF methods handle the pixel-based profiles to extract the low-level features, while deep network based models consider spatial information more. For CNN-GRU and CNN-Transformer, different datasets have different spatial window sizes. The window size of Pavia University and Indian Pines is $7 \times 7$. After dividing the hyperspectral window cube into the orderly sequence along the spectral dimension, we use 3 CNN layers to extract spatial feature and obtain the spectral embedding for each spectral band. Then 4-layer encoder module of Transformer is adopted to ex- press spectral correlation. Lastly, a feed forward module and a soft-max layer are utilized to predict categories, and the feed forward module consists of two fully connected layers with dimensions 80 and 30, respectively. For the proposed CNN-Transformer network, it is trained with Adadelta algorithm, and a low learning rate of 0.001 is taken to train the model. And in the experiments, the training epoch number of CNN-Transformer and compared CNN-GRU is set to 1000, which ensures the convergence of models.

For the two experimental datasets, Table 1 and 2 show the classification accuracies of comparison methods on Indian Pines and University of Pavia. It can be seen that the classification results of traditional pixel-based approaches are very poor. Compared with RF and SVM-RBF methods, the deep-learning based methods of CNN, CNN-GRU and CNN-Transformer can obtain more accurate results than traditional models. Benefiting from adopting spatial information, these CNN-based models can achieve the more cleaner results. For sequence modeling, CNN-GRU handles the spectral sequence with recursive calculation, while CNN-Transformer can calculate the spectral correlation directly. That is, without long-term recursive information loss, we can obtain the better corre-lation expression and more accurate classification results. It is clear that the proposed CNN-Transformer model yields more accurate results than the other models. Specifically, as shown in Table 1, compared with CNN-GRU, CNN-Transformer model increases the accuracy significantly by 3.77% of OA, 2.55% of AA, and 0.0431 of the Kappa coefficient, respectively. And in Table 2, compared with CNN-GRU, the increments of OA, AA, Kappa coefficient obtained by CNN-Transformer model are 1.48%, 1.74%, 0.0197 and 0.87%, 0.39%, 0.0098 respectively.

**Table 1.** Classification Accuracies of Different Techniques in Percentages for the Indian Pines Dataset. The Best Accuracy in Each Row is Shown in Bold.

| Class | RF | SVM-RBF | CNN | CNN-GRU | Transformer |
|---|---|---|---|---|---|
| 1 | 40.63 | 60.87 | 100.00 | 100.00 | **100.00** |
| 2 | 73.70 | 77.73 | 96.26 | 87.63 | **98.36** |
| 3 | 65.06 | 67.23 | 97.60 | 97.79 | **99.73** |
| 4 | 50.60 | 65.25 | **97.22** | 97.18 | 97.18 |
| 5 | 92.60 | 93.80 | **98.61** | 96.90 | 97.93 |
| 6 | 96.48 | 97.53 | 100.00 | 99.77 | **100.00** |

(*continued*)

**Table 1.** (*continued*)

| Class | RF | SVM-RBF | CNN | CNN-GRU | Transformer |
|---|---|---|---|---|---|
| 7 | 35.00 | 35.71 | 100.00 | 100.00 | **100.00** |
| 8 | 99.10 | 98.74 | 99.77 | 99.53 | **99.77** |
| 9 | 21.43 | 30.00 | 100.00 | 91.67 | **100.00** |
| 10 | 82.94 | 84.98 | 91.78 | 93.65 | **99.66** |
| 11 | 88.19 | 90.47 | 98.64 | 96.33 | **99.73** |
| 12 | 66.27 | 76.09 | 97.75 | 92.42 | **99.44** |
| 13 | 93.01 | 93.14 | **100.00** | 99.46 | 100.00 |
| 14 | 96.61 | 98.10 | 99.47 | 98.55 | **99.91** |
| 15 | 53.33 | 58.55 | 94.83 | 98.71 | **99.71** |
| 16 | 86.15 | 86.96 | **100.00** | 100.00 | 100.00 |
| OA | 82.29 | 85.29 | 97.66 | 95.71 | **99.48** |
| AA | 71.32 | 75.95 | 98.26 | 96.91 | **99.46** |
| Kappa | 0.7964 | 0.8312 | 0.9733 | 0.9510 | **0.9941** |

**Table 2.** Classification Accuracies of Different Techniques in Percentages for the University of Pavia Dataset. The Best Accuracy in Each Row is Shown in Bold.

| *Class* | *RF* | *SVM-RBF* | *CNN* | *CNN-GRU* | *Transformer* |
|---|---|---|---|---|---|
| 1 | 93.13 | 93.31 | 99.40 | 95.09 | **99.65** |
| 2 | 97.75 | 99.09 | 99.93 | 99.49 | **100.00** |
| 3 | 67.92 | 62.86 | 93.49 | 95.98 | **97.46** |
| 4 | 90.28 | 92.38 | **99.89** | 99.67 | 99.89 |
| 5 | 98.27 | 99.75 | 99.75 | 99.92 | **100.00** |
| 6 | 64.54 | 71.90 | 99.73 | 99.36 | **99.98** |
| 7 | 72.68 | 71.43 | 99.75 | 95.24 | **100.00** |
| 8 | 90.19 | 89.59 | 97.10 | 95.89 | **98.73** |
| 9 | 99.18 | 99.88 | 99.65 | 99.88 | **100.00** |
| OA | 89.75 | 91.10 | 99.26 | 98.22 | **99.70** |
| AA | 85.99 | 86.69 | 98.74 | 97.78 | **99.52** |
| Kappa | 0.7964 | 0.8312 | 0.9733 | 0.9510 | **0.9941** |

## 4 Conclusions

In this letter, we propose a Transformer-based networks to perform hyperspectral classification. The pixelbased hyperspectral sample can be seen as a spectral sequence, therefore, we borrowed the Transformer architecture from the NLP knowledge, which has strong sequence expression modeling ability to model the correlation between spectral bands in hyperspectral classification. In the network, after first obtaining the spatial features through CNN, we obtain the spectral embeddings and position embeddings of spectral sequence. Then, we introduce the multi-layer encoder module derived from Transformer to represent the frequency band correlation. Finally, the class labels are predicted through feedforward and softmax layer. The experiment results show that the proposed sequence-based model can achieve the state of art performance for hyperspectral classification.

**Acknowledgements.** This paper is supported by the 2021 school level scientific research project "Mutual learning between symbolic logic and knowledge graph" of Tianjin Normal University.

## References

1. Alec Radford, T.S., Narasimhan, K., Sutskever, I.: Improving language understanding by generative pre-training (2018)
2. Bluche, T.: Joint line segmentation and transcription for end-to-end handwritten paragraph recognition (2016)
3. Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P.: Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans. Geosci. Remote Sens. **54**(10), 6232–6251 (2016)
4. Chen, Y., Nasrabadi, N.M., Tran, T.D.: Hyperspectral image classification via kernel sparse representation. IEEE Trans. Geosci. Remote Sens. **51**(1), 217–231 (2013)
5. Delalieux, S., Somers, B., Haest, B., Spanhove, T., Borre, J.V., Mu¨cher, C.A.: Heathland conservation status mapping through integration of hyperspectral mixture analysis and decision tree classifiers. Remote Sens. Environ. **126**(4), 222–231 (2012)
6. Gualtieri, J.A., Chettri, S.: Support vector machines for classification of hyper-spectral data. In: IEEE International Geoscience & Remote Sensing Symposium (2000)
7. Ham, J., Chen, Y., Crawford, M.M., Ghosh, J.: Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sens. **43**(3), 492–501 (2005)
8. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)
9. Long, C., Zhang, H., Xiao, J., Nie, L., Chua, T.S.: Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning (2016)
10. Makantasis, K., Karantzalos, K., Doulamis, A., Doulamis, N.: Deep supervised learning for hyperspectral data classification through convolutional neural networks. In Geoscience & Remote Sensing Symposium (2015)
11. Melgani, F., Bruzzone, L.: Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens. **42**(8), 1778–1790 (2004)
12. Miech, A., Laptev, I., Sivic, J.: Learnable pooling with context gating for video classification (2017)

13. Mou, L., Ghamisi, P., Xiao, X.Z.: Deep recurrent neural networks for hyper-spectral image classification. IEEE Trans. Geosci. Remote Sens. **55**(7), 1–17 (2017)
14. Vaswani, A., et al.: Attention is all you need (2017)
15. Xu, K., Ba, J., Kiros, R., Cho, K., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. Comput. Sci. 2048– 2057 (2015)
16. Yi, C., Nasrabadi, N.M., Tran, T.D.: Hyperspectral image classification using dictionary-based sparse representation. IEEE Trans. Geosci. Remote Sens. **49**(10), 3973–3985 (2011)
17. Zhang, X., Sun, Y., Jiang, K., Li, C., Jiao, L., Zhou, H.: Spatial sequential recurrent neural network for hyperspectral image classification. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. **99**, 1–15 (2018)

# A Time Jitter DDMA MIMO Automotive Radar Waveform

Jianhu Liu[1], Hongfei Lian[2(✉)], Qiao Chen[1], and Sijia Chen[1]

[1] Beijing Rxbit Electronic Technology Co., Ltd., Beijing, China
{liujianhu,chenqiao,chensijia}@racodf.com
[2] National Key Laboratory of Science and Technology on Automatic Target Recognition, Changsha, China
lianhongfei97@163.com

**Abstract.** Aiming at the problems of low utilization of time and space and Doppler ambiguity of high-velocity targets in the traditional multiple input multiple output (MIMO) waveform of automotive radar, a time jitter Doppler frequency division multiplexing automotive radar waveform is proposed. This waveform jitters the pulse repetition interval stagger in the time domain, uses the phase difference between odd and even sequences to realize velocity ambiguity resolution, and realizes waveform orthogonality in the frequency domain through unequal spacing Doppler frequency modulation between multiple antennas, which improves the time-space transmission efficiency. In addition, an echo separation method based on non coherent accumulation is proposed, which can effectively solve the problem of target matching under Doppler frequency division multiplexing waveform. The effectiveness of the proposed waveform and signal processing method is verified by simulation experiments, and the waveform has strong application value in automotive radar engineering practice.

**Keywords:** Automotive radar · Multiple input multiple output · Time jitter · Doppler frequency division multiplexing · Incoherent accumulation

## 1 Introduction

Automotive Millimeter wave radar has wide coverage, high reliability, all-weather and all-weather target detection capabilities, and with the development of technology, the cost will be lower and lower [1–5]. This makes the Automotive Millimeter Wave Radar become one of the irreplaceable core sensors in the advanced driving assistance system. Automotive radar can realize high-resolution range and velocity estimation by transmitting linear frequency modulation continuous wave (FMCW) in millimeter wave band, and the cost is greatly reduced compared with laser radar [6–10]. Due to the limitation of physical space, the on-board millimeter wave radar is small in size, and due to the complex driving environment of vehicles, it is necessary to reconstruct the environmental information with high accuracy [11, 12]. Therefore, most of the on-board millimeter wave radars adopt multiple input multiple output (MIMO) system. MIMO automotive


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 26–32, 2023.
https://doi.org/10.1007/978-981-99-2362-5_4

radar uses linear frequency modulation continuous wave combined with other mechanisms to achieve waveform orthogonality. The commonly used orthogonal waveforms include time division multiplexing, frequency division multiplexing, code division multiplexing and Doppler frequency division multiplexing. The time division multiplexing MIMO system is the most commonly used in automotive radar. Only one antenna transmits in a pulse time period. Its transmission mechanism is simple and easy to implement in engineering, so it is widely used in automotive radar. However, in the time-division MIMO waveform, due to the time-division transmission of each antenna, not only the utilization rate of time and space is low, but also the pulse repetition rate is reduced, which aggravates the problem of Doppler ambiguity. In the frequency division multiplexing orthogonal waveform, the carrier frequencies of multiple transmitting antennas are different, but due to hardware limitations, there is only one VCO in the existing linear frequency modulation continuous wave automotive radar chip, which can not realize frequency division multiplexing, so it is not feasible in engineering. In the code division multiplexing waveform, multiple transmitting antennas transmit mutually orthogonal code sequences, and the receiving end can realize the separation of transmitted signal echoes after matched filtering. However, in this waveform, the code sequences cannot achieve ideal complete orthogonality, so there will be interference redundancy between the code sequences, resulting in the rise of bottom noise, which seriously affects the detection and estimation of targets, especially weak and small targets. The Doppler frequency division multiplexing waveform modulates the linear phase between the pulses of multiple transmitting antennas to achieve separation in the Doppler dimension. However, this waveform reduces the Doppler tolerance and is prone to Doppler ambiguity, which will seriously affect the detection performance of automotive radar.

In view of the above problems, this paper proposes a new waveform of time jitter Doppler frequency modulation based on the multi transmit and multi receive automotive radar system. The waveform performs pulse repetition interval stagger jitter in time domain to achieve sequence diversity, uses jitter phase difference between odd and even sequences to complete velocity ambiguity resolution, and achieves Doppler frequency orthogonality through inter pulse phase modulation in frequency domain to adapt to multi-channel angle measurement processing. This waveform can effectively solve the problem of vehicle radar velocity ambiguity, and can also take into account the multi-channel angle measurement.

## 2 Time Jitter DDMA MIMO Waveform

### 2.1 Waveform Description

This paper proposes a time jitter DDMA MIMO waveform based on the traditional linear frequency modulation sequence. The time-frequency relationship of the transmitted signal is shown in Fig. 1. The waveform realizes sequence interleaving diversity by pulse repetition interval stagger jitter in the time domain, and realizes Doppler frequency orthogonality by inter pulse phase modulation of multiple transmit antennas in the frequency domain.

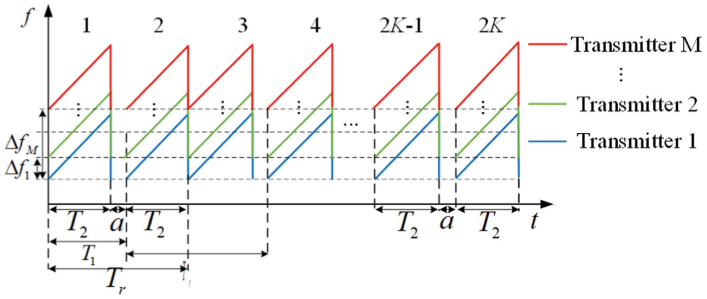



**Fig. 1.** Time frequency diagram of time jitter DDMA MIMO waveform

In the time domain, the waveform can be decomposed into odd and even pulse sequences of length $K$, wherein the pulse time interval in the odd pulse sequence is $T_1$, and the pulse time interval in the even pulse sequence is $T_2$, and $T_1 = T_2 + a$. The frequency modulation duration of all pulses in this waveform is $T_2$ and the frequency modulation bandwidth is consistent. In the frequency domain, the three transmitting antennas realize unequal interval Doppler frequency division multiplexing through inter pulse phase modulation. Among them, the transmitting antenna 1 is not modulated, the Doppler frequency difference between the transmitting antenna 2 and the transmitting antenna 1 is $\Delta f_1$, the Doppler frequency difference between the transmitting antenna 3 and the transmitting antenna 1 is $\Delta f_2$, …, The Doppler frequency difference between the transmitting antenna m and the transmitting antenna 1 is $\Delta f_{M-1}$. In order to reliably overcome the maximum velocity limit without the assumption of a single target, the waveform performs Doppler frequency modulation between multiple transmit antennas based on a vacant subband. Take the antenna array with three transmitters and four receivers as an example. If one empty subband is set, the modulation phase relationship of multiple transmitting antennas is shown in Fig. 2 and Fig. 3.

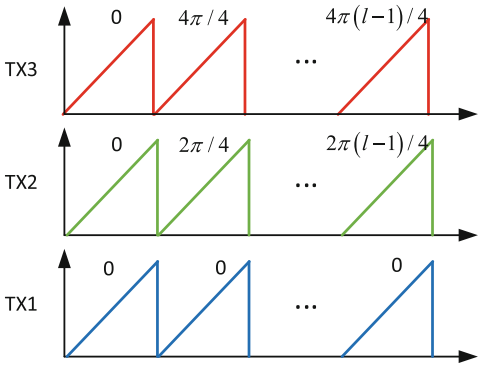



**Fig. 2.** Schematic diagram of modulation phase of time jitter DDMA waveform

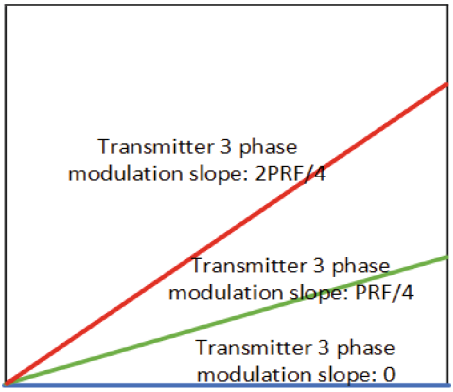



**Fig. 3.** Schematic diagram of modulation phase slope of time jitter DDMA waveform

In the above modulation method, the Doppler frequency division multiplexing between multiple antennas is realized by inter pulse phase modulation. The modulation phase of the $m$ ($1 \leqslant m \leqslant M$) th pulse of the $k$ ($1 \leqslant k \leqslant K$) th transmitting antenna in the odd and even sequences can be expressed as

$$\varphi_{mk} = \exp[j2\pi \frac{f_r}{M+1}(m-1)(k-1)T_r] \tag{1}$$

## 2.2 Time Jitter DDMA MIMO Waveform Signal Processing

Since the pulse repetition intervals of the two sub pulses in this waveform are different, the odd and even frequency modulation sequences need to be processed separately. The equivalent repetition period of the two pulse sequences is $T_r$. From the signal modulation scheme, after performing two-dimensional FFT on odd sequence and even sequence, there are multiple focal points in the two-dimensional spectrum of the parity sequence. Therefore, first of all, it is necessary to match the target to obtain the ambiguous Doppler frequency of the target. Here, a target matching method based on non coherent accumulation is proposed.

Suppose that the number of transmitting antennas is $M$, the number of receiving antennas is $N$, and the number of Doppler frequency modulated subbands is $P(M < P)$. The steps of the method are as follows.

Step 1: Non coherent accumulation of data of all receiving channels.

Step 2: Divide the matrix obtained in step 1 into $P$ equally spaced subbands according to the Doppler dimension.

Step 3: According to the modulation method of the transmitting antenna, m consecutive sub bands in the $P$ frequency bands are selected in sequence, and then non coherent accumulation is performed again to form $M$ accumulation bands.

Step 4: Select the target with the largest peak value of the corresponding position in the $M$ accumulation frequency bands as the target position to complete the target matching process.

After the target matching is completed by the above method, the range and ambiguous Doppler index of the target can be obtained, and the range and ambiguous Doppler frequency of the target can be obtained.

Next, the phase difference ambiguity resolution is performed. Referring to the idea in [13], first extract the phase sum of the peak points in the two sequences $\xi_1$ and $\xi_2$, and consider the phase winding, then define

$$\varphi_{\mathrm{d}} = \xi_2 - \xi_1 - 2\pi f_{D,amb}(T_2 + a) \tag{2}$$

Next, define

$$\delta_{\mathrm{d}} = \frac{\varphi_{\mathrm{d}}}{2\pi} - \left[\frac{\varphi_{\mathrm{d}}}{2\pi}\right] \tag{3}$$

It can be deduced that the relationship $\delta_{\mathrm{d}}$ with Doppler ambiguous number $q$ is

$$q = \begin{cases} \delta_{\mathrm{d}} \dfrac{2T_{\mathrm{r}}}{a}, & -\dfrac{1}{4} < \delta_{\mathrm{d}} < \dfrac{1}{4} \\ \left(\delta_{\mathrm{d}} + \dfrac{1}{2}\right)\dfrac{2T_{\mathrm{r}}}{a}, & \dfrac{1}{4} < \delta_{\mathrm{d}} < \dfrac{1}{2} \\ \left(\delta_{\mathrm{d}} - \dfrac{1}{2}\right)\dfrac{2T_{\mathrm{r}}}{a}, & -\dfrac{1}{2} < \delta_{\mathrm{d}} < \dfrac{1}{4} \end{cases} \tag{4}$$

The Doppler ambiguity number of the target can be obtained from (4), and the target Doppler frequency is

$$f_{\mathrm{D}} = f_{\mathrm{D,amb}} + q \cdot f_{\mathrm{D,max}}, q \in \mathbb{Z} \tag{5}$$

So far, the target velocity ambiguity resolution is completed. Finally, a virtual transceiver array is formed by Doppler frequency division multiplexing to complete multi-channel angle measurement.

## 3 Simulation

This section takes the 3-transmit and 4-receiver radar as an example. The waveform parameters of the simulation test are shown in Table 1.

Set four targets with different distances, velocitys and angles. The target parameters and estimation results are shown in the Fig. 4 and Table 2.

It can be seen from the results in Table 2 that the waveform proposed in this paper can accurately estimate the target parameters, which verifies the effectiveness of the waveform.

**Table 1.** Typical parameter examples of automotive radar system

| Parameter | Value |
|---|---|
| $f_0$/GHz | 77 |
| band /MHz | 150 |
| Number of transmitter | 3 |
| Number of receiver | 4 |
| PRT/μs | 50 |
| Number of pulses | 192 |
| $T_1$/μs | 60 |
| $T_2$/μs | 40 |

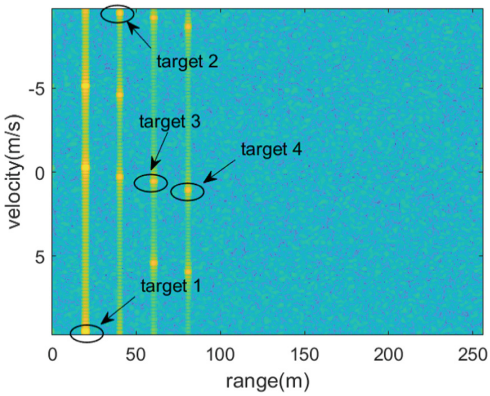



**Fig. 4.** Multi target spectrum results

**Table 2.** Simulation target parameter estimation results

| | Target 1 | Target 2 | Target 3 | Target 4 |
|---|---|---|---|---|
| Target distance/m | 20 | 40 | 60 | 80 |
| Target velocity/(m/s) | −10 | 10 | 20 | 40 |
| Target angle/(°) | 10 | 20 | 30 | 40 |
| Estimated distance/m | 19.96 | 40.3 | 60.2 | 80.62 |
| Estimated ambiguous velocity /(m/s) | 9.47 | −9.47 | 0.3 | 1.10 |
| Estimated $q$ | −1 | 1 | 1 | 2 |
| Estimated velocity /(m/s) | −10.01 | 10.01 | 20.01 | 40.06 |
| Estimated angle/(°) | 10.08° | 20.02° | 30.05° | 20.02° |

## 4 Conclusion

In order to solve the problems of low utilization of time and space and Doppler ambiguity of high-velocity targets, a new waveform of time jitter Doppler frequency division multiplexing is proposed in this paper. The waveform performs sequence diversity in the time domain, velocity ambiguity resolution by using the phase difference caused by pulse repetition interval jitter, and Doppler frequency division multiplexing by inter pulse phase modulation of multiple transmit antennas in the frequency domain. The multi-target simulation test results show that the waveform has excellent non ambiguous velocity measurement and multi-channel angle measurement capabilities while taking into account the transmission efficiency. It has strong application value for the engineering practice of automotive radar.

## References

1. Stockle, C., Herrmann, S., Dirndorfer, T., Utschick, W.: Automated vehicular safety systems: robust function and sensor design. IEEE Signal Process. Mag. **37**(4), 24–33 (2020)
2. Sun, S., Zhang, Y.D.: 4D automotive radar sensing for autonomous vehicles: a sparsity-oriented approach. IEEE J. Select. Top. Sign. Process. **15**(4), 879–891 (2021)
3. Sun, S., Petropulu, A.P., Poor, H.V.: MIMO radar for advanced driver-assistance systems and autonomous driving: advantages and challenges. IEEE Sign. Process. Mag. **37**(4), 98–117 (2020)
4. Bilik, I., Longman, O., Villeval, S., Tabrikian, J.: The rise of radar for autonomous vehicles: signal processing solutions and future research directions. IEEE Signal Process. Mag. **36**(5), 20–31 (2019)
5. Sun, S., Zhang, Y.D.: Four-dimensional high-resolution automotive radar imaging exploiting joint sparse-frequency and sparse-array design. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8413–8417 (2021)
6. Metz, C., et al.: Fully integrated automotive radar sensor with versatile resolution. IEEE Trans. Microw. Theory Tech. **49**(12), 2560–2566 (2001)
7. Steinhauer, M., Ruoss, H.-O., Irion, H., Menzel, W.: Millimeter-wave-radar sensor based on a transceiver array for automotive applications. IEEE Trans. Microw. Theory Tech. **56**(2), 261–269 (2008)
8. Lin, Y.-C., Lee, T.-S., Pan, Y.-H., Lin, K.-H.: Low-Complexity high-resolution parameter estimation for automotive MIMO radars. IEEE Access **8**, 16127–16138 (2020)
9. Wollitzer, M., et al.: Multifunctional radar sensor for automotive application. IEEE Trans. Microw. Theory Tech. **46**(5), 701–708 (1998)
10. Uysal, F.: Phase-coded FMCW automotive radar: system design and interference mitigation. IEEE Trans. Veh. Technol. **69**(1), 270–281 (2020)
11. Ding, X., et al.: Theory and practice: a two-channel automotive radar for three-dimensional object detection. Eur. Radar Conf. **2015**, 265–268 (2015)
12. Xu, Z., et al.: Simultaneous monitoring of multiple people's vital sign leveraging a single phased-MIMO radar. IEEE J. Electromagn. RF Microw. Med. Biol. **6**(3), 311–320 (2022)
13. Wang, Y.K.: Research on signal processing method for chirp squeeze automotive radar. In: Nanjing University of Science & Technology, p. 149 (2018)

# Heterogeneous Physical Layer Network Coding in the Presence of Symbol Asynchrony

Shenshen Li and L. F. Xie(✉)

Faculty of EECS, Ningbo University, Ningbo 317500, China
xielingfu@nbu.edu.cn

**Abstract.** In this paper, we address the vulnerability of Heterogeneous Physical-Layer Network Coding (HePNC) to a phase offset in end users' signals. A measure studied previously is to have different network coding schemes for different phase offsets. Our study, however, considers symbol asynchrony between the end users in HePNC and investigates the joint effect of the phase and symbol offsets on HePNC. For this, we adopt an oversampling method on superimposed signals, construct a factor graph to capture the correlation between the signal samples, and design a belief propagation algorithm on the graph to detect the users' transmit symbols. Several network coding schemes including the simple bitwise XOR are considered in our study. By simulation, we show that the symbol asynchrony can be a remedy to counter the adverse effect of the phase offset and in particular, for a symbol offset around half a symbol duration, the XOR coding suffices for all phase offsets.

**Keywords:** Heterogeneous physical-layer network coding · symbol asynchrony · phase offset · belief propagation

## 1 Introduction

Physical-layer network coding (PNC) [1] is a spectral-efficient transmission technique that has been a research focus for over a decade. It was pioneered in 2006 in [1], where the relay of a two-way relay network (TWRC) was empowered by PNC to turn interference into network coded information and boost network throughput. Since then, enormous efforts have been devoted to understanding PNC under various system settings, e.g., in MIMO systems [2], in channel-coded systems [3], and in multi-access non-relay systems [4].

This paper studies PNC in TWRC, but with asymmetric uplink qualities of the end nodes. It is called homogeneous PNC (HoPNC) if the end nodes in TWRC use the same modulation, and it is called heterogeneous PNC (HePNC) otherwise [5]. When it comes to a TWRC with asymmetric uplink qualities, HePNC is usually preferred over HoPNC, as pointed out in [5].

For both HoPNC and HePNC, there is a line of work aiming at minimizing the error probability of the network coding operation under various modulation


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 33–40, 2023.
https://doi.org/10.1007/978-981-99-2362-5_5

schemes (e.g., PAM [6,9], QAM [7], and PSK [8]). A proper constellation design, a proper bit labeling, and a proper PNC mapping (i.e., a network coding operation) are the studied means for this aim. Particularly, it was shown in [5] that for either HoPNC or HePNC, the PNC mapping should vary with the phase offset between the signals of the end nodes. [10] found that a symbol offset between the end nodes' signals at the relay had a positive role in mitigating the adverse effect of the phase offset in HoPNC. For HePNC, however, it is unknown if this also works. This paper attempts to answer this question by exploring the joint effect of the symbol and phase offsets on HePNC, and it makes the following contributions.

- First, we study a HePNC with BPSK and QPSK used respectively at the two end nodes. As in [10], we deal with the symbol asynchrony by an oversampling method and a factor graph to capture the correlation between the samples. Then, a belief propagation (BP) algorithm for message passing is proposed for signal detection and network coding.
- Second, we consider several network coding schemes, including the bitwise XOR, and conduct simulation to study the joint effect of the symbol and phase offsets on these schemes. The simulation results reveal that for a symbol offset around half a symbol duration, the XOR coding suffices for all phase offsets.

## 2 System Model

We consider a two-way relay channel (TWRC) as shown in Fig. 1, where the end nodes A and B wish to exchange packets with each other via the relay R. Flat fading is assumed in TWRC, and the channel coefficient from node $i$, $i \in \{A, B\}$, is denoted by $h_i$. We assume the links from the end nodes to relay R are asymmetric, and without loss of generality, we assume $|h_A| > |h_B|$. Thus, node A may use a higher order modulation than node B. For simplicity, we assume they use QPSK and BPSK respectively. PNC involves two phases: the uplink phase, in which nodes A and B transmit their packets simultaneously to relay R and relay R deduces a network coded packet from the superimposed signal, and the downlink phase, in which relay R broadcasts the network coded packet to nodes A and B, allowing them to obtain the desired packets. The superimposed baseband signal at relay R, involving a relative symbol offset $\Delta\tau$ between nodes A and B, can be expressed as

$$y_R(t) = \sum_{n=1}^{L} \{h_A x_A[n]\, p(t - nT) + h_B x_B[n]\, p(t - \Delta\tau - nT)\} + w_R(t), \quad (1)$$

where $T$ is the symbol duration, $\{x_i[n]\}_{n=1,2,\ldots,L}$ is the packet sent by node $i$ ($x_A[n] \in \{\pm\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}\}$ and $x_B[n] \in \{\pm 1\}$), $p(t)$, a pulse shaping function, is assumed to be the rectangular pulse here, and $w_R(t)$ is white Gaussian noise with power spectral density $\frac{N_0}{2}$ for either the real or the imaginary part.

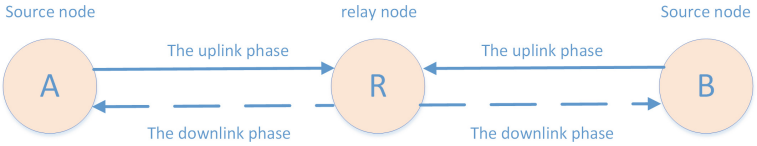



**Fig. 1.** PNC in TWRC.

With matched filtering and the oversampling method in [10] on $y_R(t)$ (see [10] for details) and assuming node A's signal arrives earlier than node B's for no more than $T$ (i.e., $0 \leq \Delta\tau \leq T$), we obtain

$$\begin{aligned} y_R[2n-1] &= h_A x_A[n] + h_B x_B[n-1] + w_R[2n-1], \\ y_R[2n] &= h_A x_A[n] + h_B x_B[n] + w_R[2n], \\ y_R[2L+1] &= h_B x_B[L] + w_R[2L+1], n = 1, 2, ..., L. \end{aligned} \tag{2}$$

The variances of the Gaussian noises $w_R[2n-1]$, $w_R[2n]$, and $w_R[2N+1]$ are $\frac{N_0}{\Delta\tau}$, $\frac{N_0}{T-\Delta\tau}$, and $\frac{N_0}{\Delta\tau}$, respectively.

## 3 Signal Detection and PNC Mapping

As mentioned earlier, relay R has to deduce a network coded packet $\{x_R[n]\} = \{f(x_A[n], x_B[n])\}$ from the overlapped signal $y_R(t)$ or the samples $\{y_R[n]\}$ in (2). Here $f(\cdot)$ is a mapping function, such as the simple bitwise XOR. Note that symbol $x_A[n]$ is associated with 2 bits whereas $x_B[n]$ has 1 bit associated. In bitwise XOR, this 1 bit is XORed with each of the two bits, and $\{x_R[n]\}$ is then the QPSK symbol corresponding to the resultant two bits. For notational convenience, we denote this bitwise XOR by $x_R[n] = x_A[n] \oplus x_B[n]$. The following two steps are employed to obtain $\{x_R[n]\}$.

**Step 1: signal detection of** $(x_A[n], x_B[n])$, $1 \leq n \leq L$. A correlation between any two adjacent samples in (2) can be seen, so $(x_A[n], x_B[n])$ may be detected via the probability computation of $\Pr(x_A[n], x_B[n] | \{y_R[n]\})$. For this computation, we can construct a factor graph to capture the correlation between the samples of (2) and then design a belief propagation (BP) algorithm for message passing over the factor graph.

Plotted in Fig. 2 is the factor graph, which encompasses two types of nodes, namely, the variable nodes (denoted by squares) and the check nodes (denoted by circles). A variable node is a random variable (note that $(x_A[i], x_B[j])$ is abbreviated as $x^{i,j}$ in the figure), and a check node corresponds to a function, describing the correlation between the variable nodes it connects. Four functions appear in Fig. 2; they are defined as follows:

$$\zeta_1(x_1, x_2, x_3, x_4) = \begin{cases} 1, \text{if} x_1 = x_3 \\ 0, \text{otherwise} \end{cases}, \tag{3}$$

$$\zeta_2(x_1, x_2, x_3, x_4) = \begin{cases} 1, \text{if} x_2 = x_4 \\ 0, \text{otherwise} \end{cases}, \tag{4}$$

$$\zeta_3(x_1, x_2, y) = c_1 \cdot \exp(-\Delta\tau |h_A x_1 + h_B x_2 - y|^2 / N_0), \tag{5}$$

$$\zeta_4(x_1, x_2, y) = c_2 \cdot \exp(-(T - \Delta\tau) |h_A x_1 + h_B x_2 - y|^2 / N_0), \tag{6}$$

where $c_1$ and $c_2$ are some constants used for normalization.

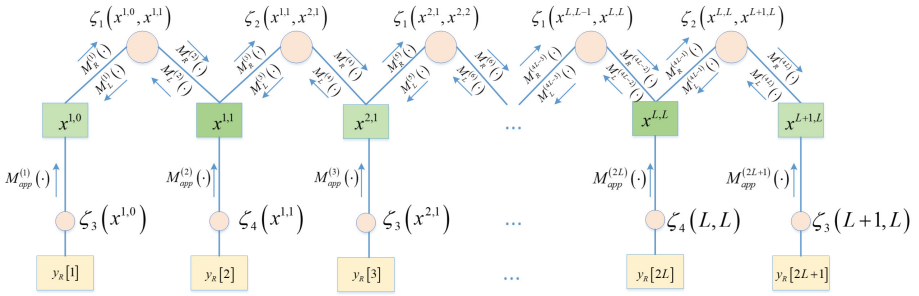



**Fig. 2.** A factor graph illustrating the correlation between the samples $\{y_R[n]\}$.

A BP algorithm is a message-passing algorithm running over a factor graph for an inference problem (e.g., the detection problem here). We thus begin with the message definition in Fig. 2. As shown in the figure, there are three types of messages; the left-bound message $M_L^{(n)}(x_1, x_2)$, $1 \le n \le 4L$, the right-bound message $M_R^{(n)}(x_1, x_2)$, $1 \le n \le 4L$, and the *a posterior* probability message $M_{app}^{(n)}(x_1, x_2)$, $1 \le n \le 2L + 1$. All the messages have $x_1 \in \{\pm\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}\}$ and $x_2 \in \{\pm 1\}$. The messages are initialized and then passed as follows.

*Message Initialization:* For any $n$ and $(x_1, x_2)$, we set $M_L^{(n)}(x_1, x_2) = \frac{1}{8}$ and $M_R^{(n)}(x_1, x_2) = \frac{1}{8}$. For even $n$, $M_{app}^{(n)}(x_1, x_2) = \zeta_4(x_1, x_2, y_R[n])$, and for odd $n$, $1 < n < 2L + 1$, $M_{app}^{(n)}(x_1, x_2) = \zeta_3(x_1, x_2, y_R[n])$. For $n = 1$, $M_{app}^{(1)}(x_1, x_2) = \zeta_3(x_1, x_2, y_R[n])$ with $h_B$ set to 0 in (5), and for $n = 2L + 1$, $M_{app}^{(2L+1)}(x_1, x_2) = \zeta_3(x_1, x_2, y_R[n])$ with $h_A$ set to 0 in (5).

*Message Passing:* There are two directions of message passing; the right-bound passing and the left-bound passing. The former updates $\{M_R^{(n)}(x_1, x_2)\}$ in sequence from $n = 1$ to $n = 4L$, as specified by the update equations below.

$$M_R^{(1)}(x_1, x_2) = M_{app}^{(1)}(x_1, x_2), \tag{7}$$

$$M_R^{(4k-2)}(x_1, x_2) = \sum_{x_j} M_R^{(4k-3)}(x_1, x_j), k = 1, 2, ..., L, \tag{8}$$

$$M_R^{(4k)}(x_1, x_2) = \sum_{x_i} M_R^{(4k-1)}(x_i, x_2), k = 1, 2, ..., L, \tag{9}$$

$$M_R^{(2k+1)}(x_1, x_2) = M_R^{(2k)}(x_1, x_2) \cdot M_{app}^{(k+1)}(x_1, x_2), k = 1, 2, ..., 2L - 1. \tag{10}$$

The left-bound passing updates $\{M_L^{(n)}(x_1, x_2)\}$ in sequence from $n = 4L$ to $n = 1$. The update equations are as follows.

$$M_L^{(4L)}(x_1, x_2) = M_{app}^{(2L+1)}(x_1, x_2), \tag{11}$$

$$M_L^{(4k-1)}(x_1, x_2) = \sum_{x_i} M_L^{(4k)}(x_i, x_2), k = 1, 2, ..., L \tag{12}$$

$$M_L^{(4k-3)}(x_1, x_2) = \sum_{x_j} M_L^{(4k-2)}(x_1, x_j), k = 1, 2, ..., L \tag{13}$$

$$M_L^{(2k)}(x_1, x_2) = M_L^{(2k+1)}(x_1, x_2) \cdot M_{app}^{(k+1)}(x_1, x_2), k = 1, 2, ..., 2L - 1. \tag{14}$$

No iterative message passing is needed for the signal detection in Fig. 2 in that the factor graph has a tree/chain structure without loops. After the above message passing, $\Pr(x_A[n], x_B[n]|\{y_R[n]\})$, $1 \leq n \leq L$, is computed from the variable node $x^{n,n}$ as follows:

$$\begin{aligned}&\Pr(x_A[n], x_B[n]|\{y_R[n]\}) \\ &= M_L^{(4n-2)}(x_A[n], x_B[n]) \cdot M_{app}^{(2n)}(x_A[n], x_B[n]) \cdot M_R^{(4n-1)}(x_A[n], x_B[n])\end{aligned} \tag{15}$$

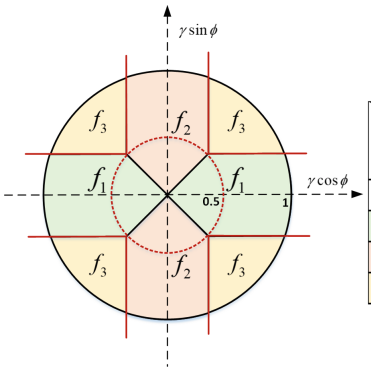
**Step 2: PNC mapping of** $(x_A[n], x_B[n])$, $1 \leq n \leq L$. Consider a specific PNC mapping $f(\cdot)$ and let $\Re(f)$ be the range of $f$. With (15), the PNC mapping $f(x_A[n], x_B[n])$ is done as follows.

$$x_R[n] = \arg\max_{x \in \Re(f)} \sum_{(x_A[n], x_B[n]): f(x_A[n], x_B[n]) = x} \Pr(x_A[n], x_B[n]|\{y_R[n]\}), \tag{16}$$

For $\Delta\tau = 0$ in our HePNC model, [5] pointed out that the design of $f(\cdot)$ ought to be adapted to the phase offset between $h_A$ and $h_B$ to maintain a low BER in PNC mapping. For QPSK used in the downlink at relay R, [5] designed three mapping functions; namely, $f_1(\cdot)$, $f_2(\cdot)$, and $f_3(\cdot)$. In Fig. 3, the right part defines the mapping of each function, and the left part illustrates in what conditions of $\frac{h_B}{h_A}$ these functions are used. In particular, it was also shown in [5] that if relay R kept using a specific mapping, say the bitwise XOR $f_3(\cdot)$, for all values of $\phi$, it then stumbled where $f_1(\cdot)$ and $f_2(\cdot)$ should be used. It is of our interest to investigate if the vulnerability of HePNC to phase offsets, given a specific PNC mapping, can be tackled by exploiting a symbol offset. This is revealed by the simulation results below.

## 4 Simulation Results

This section presents the simulation study on the joint effect of the phase and symbol offsets on HePNC under different PNC mapping functions. We focus primarily on the uplink phase of HePNC and assess the performance of a mapping function in terms of BER.




| BPSK \ QPSK | $\frac{\sqrt{2}}{2}(1+j)$ | $\frac{\sqrt{2}}{2}(-1+j)$ | $\frac{\sqrt{2}}{2}(-1-j)$ | $\frac{\sqrt{2}}{2}(1-j)$ |
|---|---|---|---|---|
| | 00 | 01 | 11 | 10 |
| 0 (under $f_1$ $f_2$ and $f_3$) | 00 | 01 | 11 | 10 |
| 1 (under $f_1$) | 01 | 00 | 10 | 11 |
| 1 (under $f_2$) | 10 | 11 | 01 | 00 |
| 1 (under $f_3$) | 11 | 10 | 00 | 01 |

**Fig. 3.** PNC mapping versus $\frac{h_B}{h_A} = \gamma e^{j\phi}$.

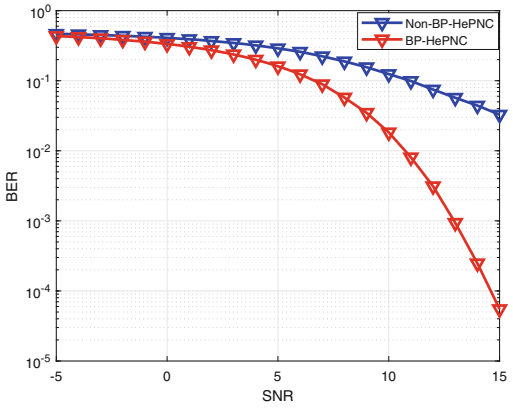



**Fig. 4.** Comparison between BP-HePNC and Non-BP-HePNC.

First, we demonstrate the superiority of our proposed message-passing signal detection (referred to as BP-HePNC hereafter) over a low-complexity detection scheme (referred to as Non-BP-HePNC) that makes only use of the samples without inter-symbol interference (ISI), i.e., $y_R[n]$ of even index. In Non-BP-HePNC, the *a posterior* probability $\Pr(x_A[n], x_B[n]|y_R[2n])$ is firstly computed using (6), and then, (16) is applied for PNC mapping (note we use $\Pr(x_A[n], x_B[n]|y_R[2n])$ in lieu of $\Pr(x_A[n], x_B[n]|\{y_R[n]\})$). We compare the two schemes in the setting as follows: $T = 1$, $\Delta\tau = 0.5$, $\frac{h_B}{h_A} = 0.5e^{j\frac{\pi}{2}}$, and $f = f_3$ (the bitwise XOR). Figure 4 plots the BER curves (against SNR, defined as $10\log\frac{1}{N_0}$) of the two schemes. We see that BP-HePNC outperforms Non-BP-HePNC for SNR above 0 dB. This is because BP-HePNC exploits the correlation between the samples and makes use of all the samples for detecting $(x_A[n], x_B[n])$.

Second, we study the effect of a symbol offset on BP-HePNC by comparing the mapping functions $f = f_3$ and $f = f_2$ under various values of $\Delta\tau$. We fix $|h_A| = |h_B| = 1$ and $\phi = \frac{\pi}{2}$. Referring to Fig. 3, we stress that the optimal

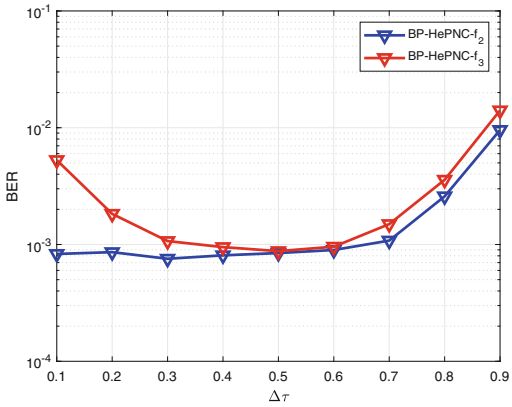
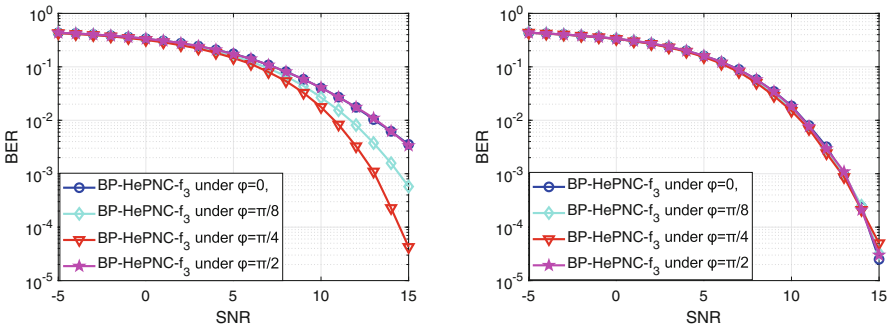



**Fig. 5.** Effect of $\Delta\tau$ on BP-HePNC with different mapping functions.

mapping in this study is $f = f_2$ when $\Delta\tau = 0$. Figure 5 plots the BER curves of BP-HePNC-$f_2$ (BP-HePNC with $f_2$) and BP-HePNC-$f_3$ (BP-HePNC with $f_3$) at SNR = 10 dB. We see that BP-HePNC-$f_3$ falls behind its counterpart for $\Delta\tau > 0.7$ or $\Delta\tau < 0.3$, but it attains a BER nearly the same as that of BP-HePNC-$f_2$ for $0.4 \leq \Delta\tau \leq 0.6$. This demonstrates that a symbol offset $\Delta\tau \neq 0$ (especially when $\Delta\tau \approx 0.5$) could overcome the adverse effect of the phase offset $\phi$ on HePNC, or equivalently, HePNC with bitwise XOR mapping becomes more robust against the phase offset, when a symbol offset comes into play.

Next, we study BP-HePNC-$f_3$ under different phase offsets and SNR values. We set $h_A = 1$ and $h_B = 0.5$. The BER curves of BP-HePNC-$f_3$ are shown in Fig. 6(a) for a low $\Delta\tau = 0.05$ and in Fig. 6(b) for $\Delta\tau = 0.5$ or half a symbol duration. We see from the figures that BP-HePNC-$f_3$ is still at the mercy of the phase offset when $\Delta\tau$ is low, but when $\Delta\tau$ increases to half a symbol duration, the situation is turned around, that is, BP-HePNC-$f_3$ becomes robust against all the phase offsets from 0 to $\frac{\pi}{2}$. Incidentally, BP-HePNC-$f_3$ performs the best at $\phi = \frac{\pi}{4}$ in Fig. 6(a), which is consistent with the condition displayed in Fig. 3.




(a) $\Delta\tau = 0.05$ (b) $\Delta\tau = 0.5$

**Fig. 6.** BP-HePNC-$f_3$ under different phase offsets and SNR values.

## 5 Conclusion

This paper has addressed the phase offset problem in heterogeneous physical-layer network coding (HePNC) in a two-way relay channel (TWRC). We showed that a symbol offset between the signals of the end nodes in the uplink transmission could be exploited to counter the adverse effect of a phase offset between the two signals in HePNC. We proposed a message-passing algorithm for signal detection in HePNC in the presence of a symbol offset. Our simulation study revealed that with a symbol offset around half a symbol duration, it sufficed to use the simple bitwise XOR as the PNC mapping for all possible phase offsets.

## References

1. Zhang, S., Liew, S., Lam, P.: Hot topic: physical-layer network coding. In: Proceedings of MobiCom, pp. 358–365 (2006)
2. Yang, T., Yuan, X., Li, P., Collings, l.B., Yuan, J.: A new physical-layer network coding scheme with Eigen-direction alignment precoding for MIMO two-way relaying. IEEE Trans. Commun. **61**(3), 973–986 (2013)
3. Huang, T., et al.: Design of irregular repeat-accumulate coded physical-layer network coding for Gaussian two-way relay channels. IEEE Trans. Commun. **61**(3), 897–909 (2013)
4. Sun, Z., Yang, L., Yuan, J., Ng, D.W.K.: Physical-layer network coding based decoding scheme for random access. IEEE Trans. Veh. Technol. **68**(4), 3550–3564 (2019)
5. Zhang, H., Cai, L.: Design of channel coded heterogeneous modulation physical layer network coding. IEEE Trans. Veh. Technol. **67**(3), 2219–2230 (2018)
6. Chen, Q., Yang, T., Liu, R.: Asymmetric linear physical-layer network coding. IEEE Commun. Lett. **25**(10), 3214–3218 (2021)
7. Yang, T., Collings, I.B.: On the optimal design and performance of linear physical-layer network coding for fading two-way relay channels. IEEE Trans. Wireless Commun. **13**(2), 956–967 (2014)
8. Arunachala, C., Buch, S.D., Rajan, B.S.: Wireless bidirectional relaying using physical layer network coding with heterogeneous PSK modulation. IEEE Trans. Veh. Technol. **67**(3), 2335–2344 (2018)
9. Yang, H.J., Choi, Y., Chun, J.: Modified high-order PAMs for binary coded physical-layer network coding. IEEE Commun. Lett. **14**(8), 689–691 (2010)
10. Lu, L., Liew, S.C.: Asynchronous physical-layer network coding. IEEE Trans. Wireless Commun. **11**(2), 819–831 (2012)

# The Optimization of the Safety and Energy Efficiency in a UAV-Assisted Communication System

Maolin Yang, Jing Gao(✉), Tingting Han, and Junchi Ma

Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
Jing401@126.com

**Abstract.** It is well known that both safety and energy consumption should be of great concern in unmanned aerial vehicles (UAVs)-assisted communication technology. In this paper, we studied the safety and energy efficiency in a UAV-assisted communication system in which there is a UAV, a ground terminal (GT) and an eavesdropper. The safety and energy efficiency was defined as the total secrecy rate in the case of considering UAV propulsion energy consumption only during its service period. Since the objective function formulated as maximizing the safety and energy efficiency is non-convex, we proposed an iterative sequential convex programming (SCP)-based algorithm to seek a suboptimal solution. Simulation results show that the method proposed in this paper can effectively improve the safety and energy efficiency of the system.

**Keywords:** UAV-assisted communication · safety and energy efficiency · trajectory design

## 1 Introduction

With fast development of wireless communication, the traditional infrastructure can hardly meet the diverse quality of service any more. Unmanned aerial vehicles (UAVs)-assisted communication technology has gained more and more attention because of its high mobility and flexible deployment [1]. However, the issues of security and energy consumption are difficult problems to be solved urgently in UAV-assisted communication technology. The propagation characteristics of wireless channels make ground terminals (GTs) vulnerable to be overheard by eavesdroppers. The security of traditional cryptography is influenced by the type of encryption and key [2]. The proposed in this case, physical layer security technology is proposed to overcome these drawbacks. Wyner [3] defined the rate of eavesdropping channel coding as the secrecy capacity. Zhang [4] et al. improved the safety capacity of the system by designing the trajectory and launch power of the UAV. In addition, UAVs can be used as airborne interference sources to send artificial noise to improve the safety capacity of the system [5].


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 41–49, 2023.
https://doi.org/10.1007/978-981-99-2362-5_6

On the other hands, the energy consumption problem of UAV-assisted communication technology determines the proper operation of the whole communication system [6]. Zeng and Zhang [7, 8] et al. improved the UAV energy efficiency by designing the UAV trajectory.

In this paper, we consider a three-node eavesdropping system, in which the UAV communicates with the GT as an air base station and the ground eavesdropping terminal acts as an eavesdropper. The UAV sends packets to GT while emitting artificial noise to interfere with the eavesdropper. By designing the flight trajectory to balance the secrecy capacity and energy consumption of the UAV, the secure and energy-saving communication between the UAV and the GT is achieved.

### 1.1 System Model and Problem Formulation

### 1.2 System Model

As shown in Fig. 1, we consider a system consists of three nodes, in which the UAV communicates with the GT as a air base station and the eavesdropping terminal acts as an eavesdropper to eavesdrop on the link between the UAV and the GT. The UAV transmits communication signals while emitting artificial noise to interfere with the eavesdropper [9]. It is assumed that the locations of all communication nodes are known, where the eavesdropper's location can be detected by the UAV equipped with an optical camera and SAR [10].

UAV is assumed to fly at a constant altitude $H$ during service time $T$. $H$ is the minimum flight altitude without frequent obstacle avoidance. To simplify the problem, we discretize the service duration $T$ as $N$ time slot, i.e., $T = N\delta_t$. Assuming that $\delta_t$ is small enough, the position of the UAV is approximately constant in each time slot. In this paper, we study the trajectory design problem of UAV in a 3D Cartesian coordinate. The positions of GT and eavesdropper are defined as $q_G \triangleq (0, 0, 0)^T$ and $q_E \triangleq (0, y_E, 0)^T$, respectively. The flight trajectory of UAV in each time slot is defined as $q[n] \triangleq (x_u[n], y_u[n], H)^T, n = 1, \cdots, N$. The starting speed, ending speed, starting position and ending position of the UAV are defined as $v_0, v_F$, $q_0 \triangleq (x_0, y_0)$, $q_F \triangleq (x_F, y_F)$.

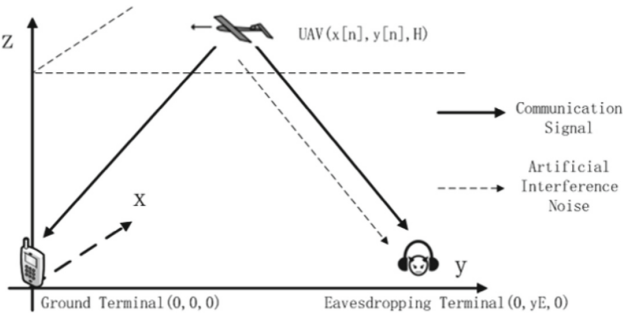



**Fig. 1.** Schematic diagram of the communication system

Assuming that the Doppler effect due to UAV movement is perfectly compensated, and the communication channel between the UAV and the GT is dominated by the line-of-sight channel. The channel between the UAV and the GT and the eavesdropper follows the path loss in free space, respectively

$$h_{UG}[n] = \frac{\beta_0}{\|q[n]\|^2 + H^2}, n = 0, \cdots, N \tag{1a}$$

$$h_{UE}[n] = \frac{\beta_0}{\|q[n] - q_E\|^2 + H^2}, n = 0, \cdots, N \tag{1b}$$

where $\beta_0$ denotes the reference channel power at $d_0 = 1m$, and its magnitude depends on the carrier power, antenna gain, etc.

The achievable secure transmission rate (bits/s/Hz) from the UAV to the GT in the nth time slot in the absence of interference from the eavesdropper is

$$R_{\text{sec}}(\{q[n]\}) = \sum_{n=1}^{N} B\left[\log_2\left(1 + \frac{P[n]h_{UG}[n]}{\delta^2}\right) - \log_2\left(1 + \frac{P[n]h_{UE}[n]}{P_n[n]h_{UE}[n] + \delta^2}\right)\right]^+ \tag{2}$$

where $(x)^+ \triangleq \max(x, 0)$, $P[n]$ and $P_n[n]$ are the power of the UAV transmitting communication signals and the power of transmitting artificial interference noise in each time slot, respectively. $\delta^2$ is the noise power of the channel, $B$ is the channel bandwidth, and $\gamma_0 = \beta_0/\delta^2$ is the reference signal-to-noise ratio.

Since the energy consumption of the UAV used to send and process information is much smaller than the propulsion energy consumption, we only consider the propulsion energy consumption of the UAV [11]. According to [7], the propulsion energy consumption model of a fixed-wing UAV is given as follows

$$\begin{aligned} &E(\{v[n], a[n]\}) \\ &= \sum_{n=1}^{N} \left( c_1 \|v[n]\|^3 + \frac{c_2}{\|v[n]\|} \left( 1 + \frac{\|a[n]\|^2 - \frac{\left(a^T[n]v[n]\right)^2}{\|v[n]\|^2}}{g^2} \right) \right) + \frac{\frac{1}{2}m\left(\|v_F\|^2 - \|v_0\|^2\right)}{\delta_t} \end{aligned} \tag{3}$$

where $m$ is the mass of the UAV, $g = 9.8$ m/s$^2$ is the acceleration of gravity, $c_1, c_2$ are constants determined by factors such as the weight and air density of the UAV, and $1/2m\left(\|v_F\|^2 - \|v_0\|^2\right)$ is the amount of change in the kinetic energy of the UAV.

In the discrete state space, the relationship between coordinates $q[n]$, velocity $v[n]$, and acceleration $a[n]$ of the UAV in the $n$ th time slot is expressed by first- and second-order Taylor expansions as follows [7].

$$v[n+1] = v[n] + a[n]\delta_t, n = 0, \cdots, N \tag{4a}$$

$$q[n+1] = q[n] + v[n]\delta_t + \frac{1}{2}a[n]\delta_t^2, n = 0, \cdots, N \tag{4b}$$

Combining Eq. (2) and (3), the safety and energy efficiency (SEE) of the system is as [12]

$$EE_{\text{sec}}(\{q[n], v[n], a[n]\}) = \frac{R_{\text{sec}}(\{q[n]\})}{E(\{v[n], a[n]\})} \tag{5}$$

### 1.3 Problem Formulation

The research objective in this subsection is SEE of fixed-wing UAV-assisted communication system. In practice, the trajectory $q[n]$, velocity $v[n]$, and acceleration $a[n]$ of UAV are subject to some certain constraints. Assuming that these constraints are satisfied, we aim to maximize SEE by optimizing the trajectory $q[n]$, velocity $v[n]$, and acceleration $a[n]$ of the UAV. The specific problem is formulated as follows.

(P1):

$$\max_{\{q[n],v[n],a[n]\}} \frac{R_{\text{sec}}(\{q[n]\})}{E(\{v[n], a[n]\})} \tag{6a}$$

$$s.t. \quad q[0] = q_0, q[N+1] = q_F \tag{6b}$$

$$v[0] = v_0, v[N+1] = v_F \tag{6c}$$

$$\|v[n]\| \leqslant v_{\max}, n = 1, \cdots, N \tag{6d}$$

$$\|v[n]\| \geqslant v_{\min}, n = 1, \cdots, N \tag{6e}$$

$$\|a[n]\| \leqslant a_{\max}, n = 0, \cdots, N \tag{6f}$$

$$(4a), (4b)$$

where $v_{\max}$ and $v_{\min}$ are the maximum and minimum flight velocities of the UAV, respectively, $a_{\max}$ denotes the maximum acceleration. Since (6a) and (6e) are not convex, standard convex optimization methods can't be employed to solve (P1) directly. In the next section, we first approximate (P1) as convex optimization problems, and then solve it by using sequential convex optimization method.

## 2 Problem Solution

In this section, we approximate (P1) as a convex optimization problem, and obtain a local optimal solution to (P1) under the guarantee that the Karush-Kuhn-Tucker (KKT) condition of the problem is satisfied.

Based on [7], the upper bound for the UAV propulsion energy consumption model $E(\{v[n], a[n]\})$ can be directly given as follows

$$\begin{aligned} &E(\{v[n], a[n]\}) \\ &\leqslant \sum_{n=1}^{N}\left(c_1\|v[n]\|^3 + \frac{c_2}{\|v[n]\|} + \frac{c_2\|a[n]\|^2}{g^2\|v[n]\|}\right) + \frac{\frac{1}{2}m\left(v_F^2 - v_0^2\right)}{\delta_t} \triangleq E_{ub}(\{v[n], a[n]\}) \end{aligned} \tag{7}$$

In Eq. (7) $a^T[n]v[n] = 0, n = 0, \cdots, N$ is tight for constant velocity flight. Introducing slack variables $\mu_n$, $\tau_n$, reformulate the problem as follows. (P1.1):

$$\max_{\{q[n],v[n],a[n]\}} \frac{\sum_{n=1}^{N} B\left(\log_2\left(1 + \frac{P[n]\gamma_0}{\|q[n]\|^2 + H^2}\right) + \log_2\left(\frac{P_n[n]\beta_0}{q[n] - q_E^2 + H^2} + \delta^2\right) - \log_2\left(\frac{P[n]\beta_0}{\mu_n} + \frac{P_n[n]\beta_0}{\mu_n} + \delta^2\right)\right)}{\sum_{n=1}^{N}\left(c_1\|v[n]\|^3 + \frac{c_2}{\tau_n} + \frac{c_2\|a[n]\|^2}{g^2\tau_n}\right) + \frac{\frac{1}{2}m\left(v_F^2 - v_0^2\right)}{\delta_t}} \tag{8a}$$

$$s.t. \quad (4a), (4b), (6b) \sim (6d), (6f) \tag{8b}$$

$$\tau_n \geqslant v_{\min}, n = 1, \cdots, N \tag{8c}$$

$$\|v[n]\|^2 \geqslant \tau_n^2, n = 1, \cdots, N \tag{8d}$$

$$\|q[n] - q_E\|^2 + H^2 \geqslant \mu_n, n = 1, \cdots, N \tag{8e}$$

$$\mu_n \geqslant H^2, n = 1, \cdots, N \tag{8f}$$

It can be shown that if and only if the constraints (8d) and (8e) are equal, the (P1.1) obtains the optimal solution, and the (P1.1) is equivalent to the (P1). It can be seen that the denominator of the objective function is jointly convex for $\{v[n], a[n], \tau_n\}$. The numerator is nonconvex, and new nonconvex constraint (8d), (8e) arises.

For the numerator of the objective function, the convex lower bound of the first-order Taylor approximation is obtained in the jth iteration as

$$\begin{aligned} R_{lb}[n] \geqslant\ & a_j[n]\left(\|q[n]\|^2 - \|q^{(j)}[n]\|^2\right) + b_j[n]\left(\|q[n] - q_E\|^2 - \left\|q^{(j)}[n] - q_E\right\|^2\right) \\ & + c_j - \log_2\left(\frac{P[n]\beta_0 + P_n[n]\beta_0}{\mu_n} + \delta^2\right) \triangleq R_{lb}^{(j)}[n], n = 1, \cdots, N. \end{aligned} \tag{9}$$

where $q^{(j)}[n]$ denotes any feasible coordinate of the UAV in the jth iteration.

$$a_j[n] = -\frac{P[n]\beta_0}{\ln 2\left(\left\|q^{(j)}[n]\right\|^2 + P[n]\gamma_0 + H^2\right)\left(\left\|q^{(j)}[n]\right\|^2 + H^2\right)}$$

$$b_j[n] = -\frac{P_n[n]\beta_0}{\ln 2\left(\sigma^2\left(\left\|q^{(j)}[n] - q_E\right\|^2 + P_n[n]\beta_0\right)\right)\left(\left\|q^{(j)}[n] - q_E\right\|^2 + H^2\right)}$$

$$c_j[n] = \log_2\left(1 + \frac{P[n]\gamma_0}{\|q^{(j)}[n]^2 + H^2}\right) + \log_2\left(\frac{P_n[n]\beta_0}{\left\|q^{(j)}[n] - q_E\right\|^2 + H^2} + \delta^2\right)$$

Similarly, the convex lower bounds of (8d) and (8e) are

$$\begin{aligned}&\psi[n] \triangleq \|v[n]\|^2\\&\geqslant \left\|v^{(j)}[n]\right\|^2 + 2\left(v^{(j)}[n]\right)^T\left(v[n]-v^{(j)}[n]\right) \triangleq \psi_{lb}^{(j)}[n], n=1\cdots,N\end{aligned} \tag{10}$$

$$\begin{aligned}&\Phi[n] \triangleq \|q[n]-q_E\|^2 + H^2\\&\geqslant \left\|q^{(j)}[n]-q_E\right\|^2 + 2\left(q^{(j)}[n]-q_E\right)^T\left(q[n]-q^{(j)}[n]\right) + H^2 \triangleq \Phi_{lb}^{(j)}[n], n=1,\cdots,N\end{aligned} \tag{11}$$

where $v^{(j)}[n]$ denotes the arbitrary feasible speed of the UAV in the jth iteration.

Therefore, for any feasible point $\left\{q^{(j)}[n], v^{(j)}[n]\right\}$, the problem (P1.1) is approximated as follows:

(P1.2):

$$\max_{\{q[n],v[n],a[n]\}} \frac{\sum_{n=1}^{N} B\left(R_{lb}^{(j)}[n]\right)}{\sum_{n=1}^{N}\left(c_1||v[n]||^3 + \frac{c_2}{\tau_n} + \frac{c_2||a[n]||^2}{g^2\tau_n}\right) + \frac{\Delta K}{\delta_t}} \tag{12a}$$

$$s.t. \quad (4a), (4b), (6b) \sim (6d), (6f), (8c), (8f)(10), (11) \tag{12b}$$

where $\Delta K \triangleq 1/2m\left(\|v_F\|^2 - \|v_0\|^2\right)$. From the above conclusion, it is clear that the objective value of (P1.2) is a lower bound of (P1). Moreover, (P1.2) is a fractional programming problem with a convex objective function and all constraints, and the Dinkelbach method can be used to solve it. Therefore, (P1) can be solved by iterative optimization (P1.2). The specific algorithm is as follows.

Algorithm 1: Safety and Energy Efficiency Maximization Algorithm

1.Initialization: finding an initial feasible solution $\{q^{(0)}[n], v^{(0)}[n], a^{(0)}[n]\}$ ,Let $EE_{\text{sec}}^{(0)} = f_{(P1)}^{(0)}\left(\{q^{(0)}[n], v^{(0)}[n], a^{(0)}[n]\}\right), j=0.$

2.**while** $|EE_{\text{sec}}^{(j)} - EE_{\text{sec}}^{(j-1)}|/EE_{\text{sec}}^{(j)} \geqslant \varepsilon$

3.Update： $j = j+1.$

4. Given $\{P[n], P_n[n]\}$,Solve the problem (P1.4) and express the optimal solution as $\{q_*^{(j)}[n], v_*^{(j)}[n], a_*^{(j)}[n]\}.$

5. Update local points $q^{(j)}[n] = q_*^{(j)}[n], v^{(j)}[n] = v_*^{(j)}[n], a^{(j)}[n] = a_*^{(j)}[n].$

6.Let $EE_{\text{sec}}^{(j)} = f_{(P1)}^{(j)}\left(\{q^{(j)}[n], v^{(j)}[n], a^{(j)}[n]\}\right).$

7.*end* **while**

Where $EE_{\text{sec}}^{(j)}$ denotes the safe energy consumption after the jth iteration. $\epsilon$ denotes the convergence accuracy. The solution resulting from the jth iteration is denoted as $\left\{q^{(j)}[n], v^{(j)}[n], a^{(j)}[n]\right\}$.

## 3 Simulation Results

This section will verify the design scheme proposed in this paper through simulation. The parameter are set as follows unless otherwise specified: $H = 100$ m $q_0 = (-1000, 400)$, $q_F = (1000, 400)$, $T = 200$ s, $\delta_t = 0.2$ s. The start and end flight speeds are $v_0 = v_F = 30(q_F - q_0)/\|q_F - q_0\|$, $v_{\max} = 50$ m/s, $v_{\min} = 3$ m/s, $a_{\max} = 5$ m/s$^2$, $c_1 = 9.26 \times 10^{-4}$, $c_2 = 2250$, $B = 1$ MHz, $N_0 = 170$ dBm/B$z$, and $\beta_0 = -60dB$. The horizontal coordinates of the GT is $q_G = (0, 0)$, and eavesdropper is $q_E = (0, 800)$. $P[n] = 80$ mW, $P_n[n] = 20$ mW. The initial flight trajectory is set to straight flight.

In Fig. 2, we depicts the horizontal flight trajectory of the UAV under the maximization of SEE and several benchmark scenarios, where "Straight Flight" corresponds to the initial trajectory. "SR-max" corresponds to the trajectory that considers only the safe transmission rate. It can be seen that the UAV flies around the GT from the starting point to the end point in order to increase the safe transmission rate. "SR-max at SEE" corresponds to the trajectory of (P1) after the objective function is modified to a safe transmission rate, which shows that the UAV bypasses the GT and hovers around the GT for a longer period of time to increase security capacity. "SEE-max" corresponds to the trajectory in Algorithm 1, and it can be seen that the UAV increases the hover radius while hovering close to the ground terminal to improve energy efficiency.

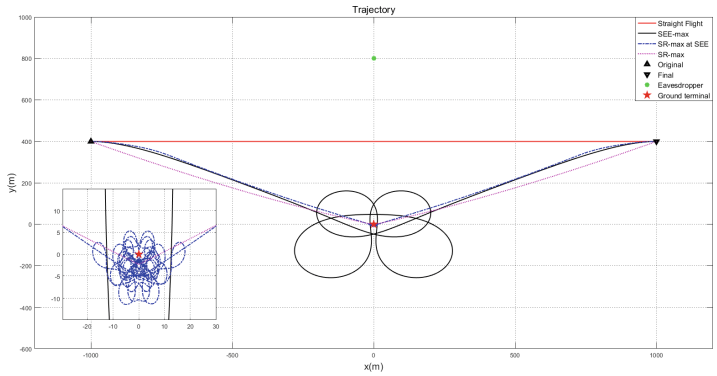



**Fig. 2.** UAV horizontal flight trajectory

In Table 1, we compares the performance of each scheme in Fig. 2. It can be seen that "SR-max" has the largest safe transmission rate. Since its energy consumption is not considered, we only compare the speed and the safe transmission rate. "SR-max at SEE" has the largest energy consumption. The energy efficiency of "SEE-max" is the largest, and the energy efficiency of "Straight Flight" is the smallest. This corresponds to the flight path of the UAV in Fig. 2.

**Table. 1.** Comparison of the Performance of Various Designs

| | $\overline{v}(m/s)$ | $\overline{a}\left(m/s^2\right)$ | $\overline{R}_{\text{sec}}(Mbps)$ | $\overline{E}(W)$ | $EE(kbits/J)$ |
|---|---|---|---|---|---|
| Straight Flight | 10.00 | 0.00 | 1.69 | 225.93 | 7.46 |
| SR-max | 30.50 | | 6.72 | | |
| SR-max at SEE | 14.50 | 4.37 | 6.63 | 540.82 | 12.25 |
| SEE-max | 26.70 | 2.68 | 4.96 | 115.02 | 43.25 |

## 4 Conclusion

In this paper, we study the fixed-wing UAV eavesdropping model with active jamming collaboration. First, we define a secure energy efficiency model for UAVs based on the propulsion energy consumption model for fixed-wing UAVs. Secondly, we formulate the objective function based on the constraints of UAV trajectory and launch power in practice. Since the objective function is a nonconvex problem, we approximate the original problem as a convex optimization problem and propose an iterative algorithm based on SCP. The local optimal solution of the original problem is obtained by solving the convex approximation problem. Finally, the simulation results show that the proposed design can effectively improve the safety energy efficiency of the system.

## References

1. Zhan, P., Yu, K., Swindlehurst, A.L.: Wireless relay communications with unmanned aerial vehicles: performance and optimization. IEEE Trans. Aerospace Electron. Syst. **47**(3), 2068–2085 (2011)
2. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Trans. Inf. Theory **22**(6), 644–654 (1976)
3. Wyner, A.D.: The wire-tap channel. Bell Labs Techn. J. **54**(8) (1975)
4. Zhang, G., Wu, Q., et al.: Securing UAV communications via joint trajectory and power control. IEEE Trans. Wireless Commun. (2019)
5. Li, A., Wu, Q., Zhang, R., et al.: UAV-enabled cooperative jamming for improving secrecy of ground wiretap channel. IEEE Wireless Commun. Lett. (2019)
6. Yan, C., Zhang, S., et al.: Fundamental trade-offs on green wireless networks **49**(6) (2011)
7. Zeng, Y., Zhang, R.: Energy-efficient UAV communication with trajectory optimization. IEEE Trans. Wireless Commun. 3747–3760 (2017)
8. Zeng, Y., Xu, J., Zhang, R.: Energy minimization for wireless communication with rotary-wing UAV. **2018**(4) (2018)
9. Goel, S., Negi, R.: Guaranteeing secrecy using artificial noise. IEEE Trans. Wireless Commun. **7**(6), 2180–2189 (2008)
10. Erdelj, M., Natalizio, E., Chowdhury, K.R., et al.: Help from the Sky: leveraging UAVs for Disaster Management. IEEE Pervasive Comput. **16**(1), 24–32 (2017)

11. Desset, C., Debaillie, B., Giannini, V., et al.: Flexible power modeling of LTE base stations. In: 2012 IEEE Wireless Communications and Networking Conference (WCNC)
12. Jiang, Y., Zou, Y., et al.: Secrecy energy efficiency optimization for artificial noise aided physical-layer security in ofdm-based cognitive radio networks. IEEE Trans. Vehicul. Technol. (2018)

# Multi-scale Channel Attention for Image Registration

Jin Zhang[1,2], Baoju Zhang[1,2(✉)], Bo Zhang[1,2], Cuiping Zhang[1,2], Youchen Sun[1,2], Cong Guo[1,2], and Jiayuan Wang[1,2]

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
wdxyzbj@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** In recent years, image registration methods based on convolutional neural networks have achieved good results and operational advantages in deformable image registration. However, due to the intrinsic property of convolution, namely the limited size of the convolution kernel, the learning of global contextual information is lacking. In order to solve the above problems, this paper proposes a multi-scale channel attention image registration model (MCAReg-Net), which can learn multi-scale contextual information and effectively fuse features of different scales for more accurate image registration. To demonstrate the effectiveness of the experiments, we conduct a series of experiments on the OASIS dataset. The experimental results show that our method can achieve better performance compared with other state-of-the-art methods, which proves the effectiveness of our proposed method.

**Keywords:** multi-scale features · channel attention · deformable image registration · convolutional neural network

## 1 Introduction

Medical image analysis based on computer-aided technology provides a reliable information basis for lesion identification and localization [1], disease diagnosis [2, 3], etc. Among them, the purpose of medical image registration [4] is to perform spatial transformation the moving image, and determine the spatial correspondence between the moving image and the fixed image [5]. It plays a vital role in medical image analysis tasks such as image fusion, medical atlas creation, and time-series image comparison [6, 7].

The traditional registration method [8] uses the image to calculate the similarity measure of the information between the pixels, and seeks the optimal correspondence by

This work was supported in part by 2021 Tianjin Postgraduate Research and Innovation Project 2021YJSS209.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 50–56, 2023.
https://doi.org/10.1007/978-981-99-2362-5_7

minimizing the similarity measure in an iterative manner. However, when such methods operate in high-dimensional space, the amount of calculation is large and the operation time is too long.

In recent years, neural networks have been widely used in medical image registration due to their advantages of being able to learn complex mappings from image features, automatically optimize registration parameters, and establish optimization models [9–11]. Balakrishnan et al. [12] propose an end-to-end unsupervised registration VoxelMorph using CNN, which achieved fast and fine registration without any supervision information. On this basis, in order to solve the problem that the VoxelMorph network does not work well in large displacement deformation, Zhang et al. [13] proposed an end-to-end recursive cascade registration network, which enabled all sub-networks to learn image features assisted by each other and achieve coarse-to-fine image registration. However, these methods cannot solve the problem of unsupervised large deformation image registration well. Mok et al. [14] proposed a large deformation image registration based on Laplacian pyramid network. However, in deep learning-based registration methods, it is very difficult to analyze the influence of hyperparameters and search for optimal regularization parameters because it requires training a large number of independent models with different hyperparameter values. Mok et al. [15] further proposed a conditional image registration method and a new self-supervised learning paradigm for deep deformable image registration. They demonstrate that optimal solutions with arbitrary hyperparameters can be captured by a single deep convolutional neural network.

However, due to the inherent properties of convolution, CNN registration-based methods usually learn local information and lack global information. In response to the above problems, PSPNet [16] used multi-scale features to learn global contextual information at different scales, that is, learning local information and global information at the same time. Hu et al. [17] proposed a channel attention mechanism that allowed the network to perform feature recalibration. However, when the features of different sizes are fused, direct stitching will lose the interdependence between the scale features. In response to these problems, this paper proposes a multi-resolution registration network based on a multi-scale channel attention module(MCAReg-Net):

We use the pyramid pooling module to extract multi-scale features and learn the context information of different regions, so as to obtain global information. We use the channel attention module to interact with features at different scales, which captures the relationship between different multi-scale features, which are better fused together for more accurate registration.

## 2 Registration Method

Define a reference image F, a floating image M, and a condition variable c. The goal of image registration is to estimate a deformation field that aligns M with F. Therefore, the conditional image registration method parameterizes for the function:

$$f_\theta(F, M, c) = \Phi \tag{1.1}$$

where $\theta$ represents the learned parameters in the network.

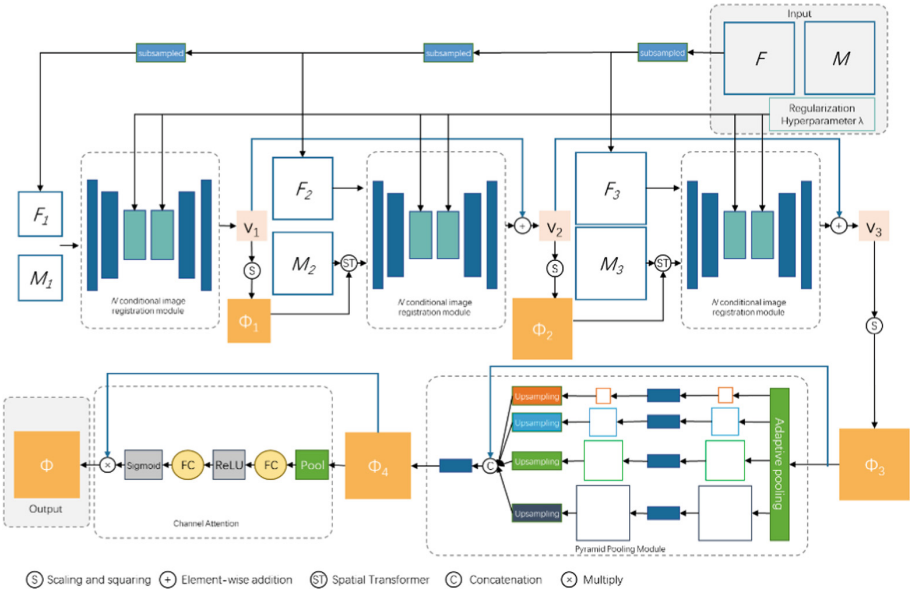



**Fig. 1.** Framework of MCAReg-Net

The overall framework of MCAReg-Net is shown in Fig. 1. In order to consider the complex semantic relationship between fixed and floating images, this paper adopts a multi-resolution strategy to jointly optimize and learn the velocity field of spatial transformation at different spatial resolutions step by step in an end-to-end learning framework. We use L identical CNN-based registration stages to simulate registration in multi-resolution mode. The velocity field estimated at the lower resolution is used to warp the moving image, and the warped moving image is used as input to the next stage to estimate the residual velocity field of the high-resolution spatial transformation. Furthermore, our method cleverly incorporates a pyramid pooling module and a multi-scale channel attention module to improve registration accuracy.

### 2.1 Pyramid Pooling Module

Due to the limited size of convolution kernels, CNN registration-based methods usually learn local information and lack the learning of global information. Multi-scale features have receptive fields of different sizes, which can capture information at different scales and obtain both global and local information. In this paper, multi-scale features are introduced, and the output feature maps of different convolutional layers are fused through the pyramid pooling module to better obtain global information. The structure is shown in Fig. 1.

The pyramid pooling model combines the features of four different pyramid scales. Through adaptive pooling, the pooling kernel covers the whole image, half of the image and a small part of the image. The red in the figure represents the coarsest level global pooling. The following pyramid level divides the feature map into different sub-regions.

The outputs of different levels in the pyramid pooling module contain feature maps of different sizes.

### 2.2 Multi-scale Channel Attention Module

Generally, multi-scale features are directly connected such as PSPNet [16], which results in insufficient use of multi-scale features. Therefore, we use the multi-scale channel attention module to interactively capture the dependencies between multi-scale features, selectively amplify valuable feature channels and suppress useless feature channels.

The structure of the multi-scale channel attention module is shown in Fig. 1. Firstly, the global average pooling is performed on the deformation field $\Phi_4$ output by the pyramid pooling model, and the spatial information of the global receptive field of the feature map is placed into the feature map with a size of $1 \times 1 \times C$ to generate the statistical information of the channel. Formally, the statistic $z$ is generated by narrowing $\Phi_4$ by the space dimension $H \times W$, so the calculation method of the c-th element of $z$ is:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \Phi_{4c}(i, j) \tag{1.2}$$

In order to simplify the model and reduce the parameters, a fully connected neural network is used to perform a nonlinear transformation on the feature map to reduce the dimension. After another ReLU layer, and then a fully connected layer to restore the dimension. Finally, through the sigmoid function, the channel weight $s$ is obtained, and its dimension is $1 \times 1 \times C$:

$$s = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{1.3}$$

where $\sigma$ is the ReLU function, $W_1 \in R^{\frac{C}{r} \times C}$, $W_2 \in R^{C \times \frac{C}{r}}$, and $r$ are the scaling parameters.

Multiply $s$ and the input feature $\Phi_4$ to get the final registration field $\Phi$:

$$\Phi_c = s_c \Phi_{4c} \tag{1.4}$$

### 2.3 Loss Function

The loss function mainly consists of two parts: similarity measure and regularization loss. The similarity measure in this paper adopts the normalized cross-correlation (NCC), and the regularization loss term is smoothed using the diffusion regularizer. Since our network is trained from coarse to fine, the objective function of each stage can be defined as:

$$u_l(F, M(\Phi), \Phi, \lambda) = \sum_{i \in [1..l]} -\frac{1}{2^{(l-i)}} NCC_\omega(Fi, Mi(\Phi)) + \lambda \|\nabla \Phi\|^2 \tag{1.5}$$

where $NCC_\omega$ denotes the local normalized cross-correlation with window size w.

## 3 Experimental Setup and Evaluation Criteria

We used 425 t1-weighted brain MR scans from the OASIS [18] 2D dataset. The image size is 160 × 192, and the dataset is randomly divided into 255, 20, and 150 volumes for training, validation, and test sets, respectively.

The method proposed in this paper is implemented using PyTorch and trained on NVIDIA-SMI GPUs. We employ the Adam optimizer with a learning rate of 0.0001. Due to the limitation of GPU memory, the number of batches is set to 1, that is, a floating image and a reference image are randomly selected each time as a pair of images to participate in training. We train in three stages, first train the network separately from the coarsest stage with 20001 iterations, then train the second stage with 40001 iterations, and finally train the most refined stage with 80001 iterations. In order to simplify the parameters, whenever a new network is added to the training, we freeze the learning parameters of the previous stage training to a fixed 3000 steps and start a new training. In the encoder-decoder network, we set L to 3, N to 5, and the pyramid pooling module is a four-layer module with sizes of 1 × 1, 2 × 2, 3 × 3, and 6 × 6, respectively.

The evaluation indicators in this paper are registration time and accuracy. In order to meet the real-time requirements of clinical registration, the registration performance can be well evaluated by comparing the registration time of different methods. For accuracy, quantitative evaluation is performed by calculating the Dice score.

## 4 Experimental Results and Analysis

The visualization results of image registration in this paper are shown in Fig. 2. Shown in the figure is a sample result obtained randomly from the test set image. It can be seen intuitively from the figure that the registered image has a high similarity with the reference image.

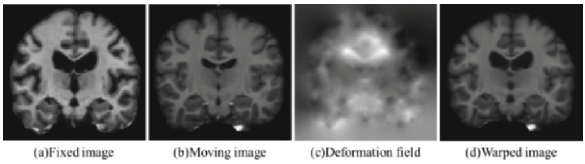



**Fig. 2.** Visualization results

In order to further compare the differences in the registration results of traditional methods such as Demons, SyN, voxelmorph, LapIRN and CIR brain MR images, the registration results of the OASIS test set were calculated using the evaluation index Dice score, and a quantitative analysis was obtained. As shown in Table 1, our proposed method achieves the highest average Dice score among different methods, which is 5.3%, 4.5%, 4.1%, 0.3% and 0.1% higher than Demons, SyN, voxelmorph, LapIRN and CIR, respectively.

In order to evaluate the registration speed of the registration network model proposed in this paper, the average registration time of Demons, SyN, voxelmorph, LapIRN and

CIR is compared, as shown in Table 1. The results show that the registration can be completed within one second when testing the optimal models trained by VoxelMorph, LapIRN, CIR and MCAReg-Net. Among them, the registration time of MCAReg-Net is about half that of VoxelMorph, and about 1/6000 of the traditional method SyN.

**Table 1.** Experimental results

| Method | Dice | Time/s |
|---|---|---|
| Initial | 0.567 | – |
| Demons | 0.715 | 192 |
| SyN | 0.723 | 1439 |
| VoxelMorph | 0.727 | 0.517 |
| LapIRN | 0.765 | 0.331 |
| CIR | 0.767 | 0.216 |
| MCAReg-Net | 0.768 | 0.209 |

## 5 Conclusion

In conclusion, we propose a novel image registration framework based on pyramid pooling model and multi-scale channel attention model for deep learning-based deformable image registration. The experimental results show that the average Dice score of the MCAReg-Net registration framework is improved by 5.3%, 4.5%, 4.1%, 0.3% and 0.1% respectively compared with Demons, SyN, voxelmorph, LapIRN and CIR, which better needs of medical image registration for high accuracy in clinical applications. Brain atlas registration experiments show that the proposed method can improve the accuracy of the registration method without sacrificing the runtime advantage.

## 6 References

1. El-Regaily, S.A., Salem, M.A., Abdel Aziz, M.H., et al.: Survey of computer aided detection systems for lung cancer in computed tomography. Curr. Med. Imag. **14**(1), 3–18 (2018)
2. Shen, W., Zhou, M., Yang, F., et al.: Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recogn. **61**, 663–673 (2017)
3. Kamnitsas, K., Ledig, C., Newcombe, V.F.J., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)
4. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vis. Appl. **31**(1), 1–18 (2020)
5. Alam, F., Rahman, S.U.: Challenges and solutions in multimo- dal medical image subregion detection and registration. J. Med. Imag. Radiat. Sci. **50**(1), 24–30 (2019)
6. Gu, Y., Lu, X., Zhang, B., et al.: Automatic lung nodule detection using multi-scale dot nodule-enhancement filter and weighted support vector machines in chest computed tomography. PLoS ONE **14**(1), e0210551 (2019)

7. Lu, X., Gu, Y., Yang, L., et al.: Multi-level 3D densenets for false-positive reduction in lung nodule detection based on chest computed tomography. Curr. Med. Imag. **16**(8), 1004–1021 (2020)
8. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**(1), 95–113 (2007)
9. Balakrishnan, G., Zhao, A., Sabuncu, M.R., et al.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260 (2018)
10. Shen, Z., Han, X., Xu, Z., et al.: Networks for joint affine and non-parametric image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4224–4233 (2019)
11. Hu, X., Kang, M., Huang, W., Scott, M.R., Wiest, R., Reyes, M.: Dual-stream pyramid registration network. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 382–390. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_43
12. Balakrishnan, G., Zhao, A., Sabuncu, M.R., et al.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
13. Zhang, L., Zhou, L., Li, R., et al.: Cascaded feature warping network for unsupervised medical image registration. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp 913–916. IEEE (2021)
14. Mok, T.C.W., Chung, A.C.S.: Large deformation diffeomorphic image registration with laplacian pyramid networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 211–221. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_21
15. Mok, T.C.W., Chung, A.C.S.: Conditional deformable image registration with convolutional neural network. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 35–45. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_4
16. Zhao, H., Shi, J., Qi, X., et al.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
18. Marcus, D.S., Wang, T.H., Parker, J., et al.: Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**(9), 1498–1507 (2007)

# Design of an IoT-Based Cross-Modality Pedestrian Monitoring System for Contact Tracing in COVID-19 Prevention

Ziyang Bian[1,3], Liang Ma[3], Jianan Li[1(✉)], and Tingfa Xu[1,2(✉)]

[1] Beijing Institute of Technology, Beijing 100081, China
{lijianan,ciom_xtf1}@bit.edu.cn
[2] Beijing Institute of Technology Chongqing Innovation Center, Chongqing 401120, China
[3] North China Research Institute of Electro-Optics, Beijing 100015, China

**Abstract.** COVID-19 epidemic prevention and control has become a regular part of life, and tracking people's trajectory (especially fever patients) in public areas can help curb the spread of the epidemic. IoT technology has dramatically improved the efficiency of epidemic prevention and control. In this paper, we propose an IoT-based cross-modality pedestrian surveillance system architecture with the following characteristics: (1) robust, the system accesses multiple modal information (visible light, infrared, temperature, mobile phone signals), and can achieve trajectory tracking in non-cooperative situations; (2) with flexibility, we develop a nodal visual artificial intelligence software platform for deep neural network training and optimization, using which the deployment can be iteratively optimized intuitively and quickly. Finally, we also discuss the research prospects of artificial intelligence and IoT technologies in the direction of epidemic prevention and control.

**Keywords:** COVID-19 Epidemic · Internet of things · cross-modality · pedestrian monitoring systems

## 1 Introduction

In December 2019, a New Coronavirus named Covid-19 was discovered and spread rapidly worldwide. A year later, in December 2020, more than 70 million people were infected with the virus, resulting in more than 1.6 million deaths. Fever is the initial symptom of covid-19 [1]. Therefore, tracking the trajectory of people in public areas (especially fever patients) is an effective method to curb the spread of the epidemic. A skin temperature extraction method proposed by Aryal in [2] uses a binocular camera with visible light and thermal imaging to detect people's skin temperature. Digital crypto tracking uses the user's smartphone to track other users nearby [3]. Then, if the contact has a positive coronavirus diagnosis, the user will be notified to take preventive measures,


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 57–64, 2023.
https://doi.org/10.1007/978-981-99-2362-5_8

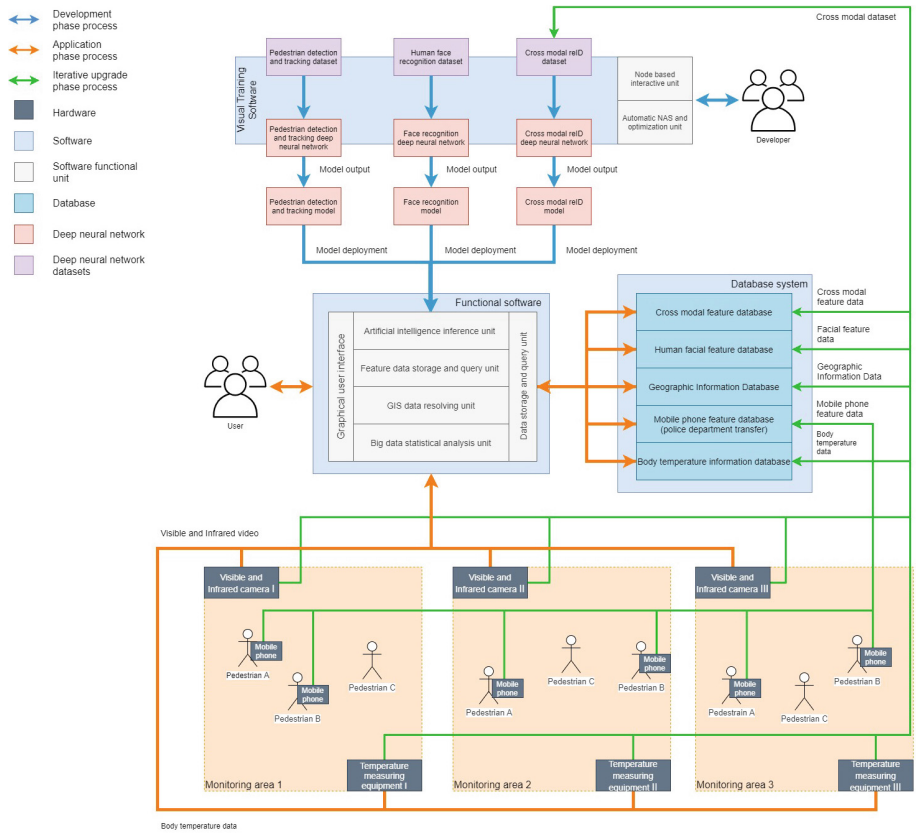



**Fig. 1.** The proposed IoT-based Cross-modality Pedestrian Monitoring System for Contact Tracing accesses various modal information (visible light, infrared, temperature, and mobile phone signal) in the Internet of things architecture.

such as testing or self-isolation. Hamagen [4] of Israel uses the global navigation satellite system (GNSS) for positioning and close contact tracking.

However, the above tracking methods still have the following problems in non-cooperative public areas: (1) the information source is single, and it can not work stably when the light changes sharply and the mobile phone refuses to cooperate. (2) The demand for epidemic prevention and control is changeable and lacks the ability of flexible deployment in different scenarios.

## 2 IoT-Based Cross-Modality Pedestrian Monitoring System Design

As shown in Fig. 1, we use pedestrian detection, face recognition, and pedestrian reidentification technology to build an end-to-end pedestrian monitoring system,

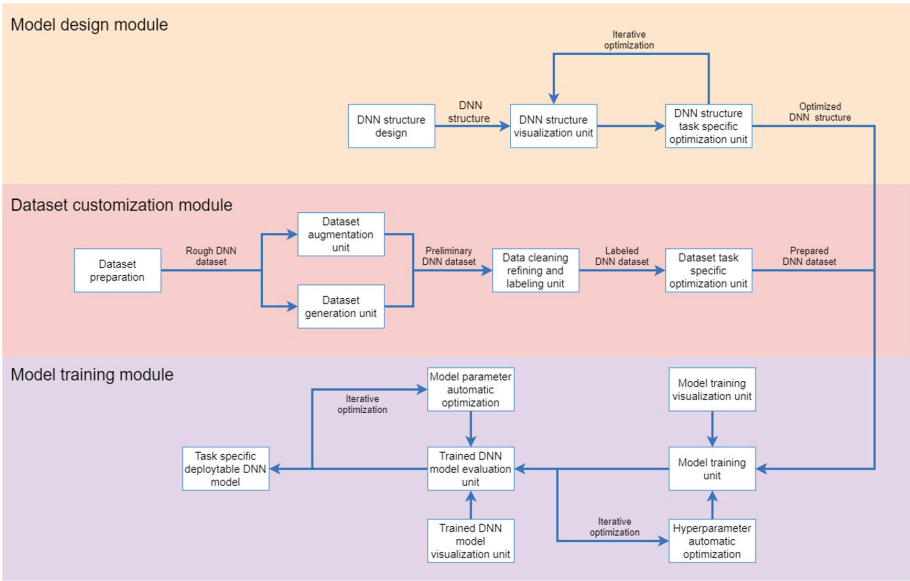



**Fig. 2.** Design of the node artificial intelligence deep neural network training platform.

which users can use without cooperation (such as long distances and poor light conditions). Realize the track tracking of specific pedestrians. To improve the ability of flexible deployment in different scenarios, we have developed a node artificial intelligence deep neural network training and optimization software platform to manage data sets, neural networks, and models visually. Developers can intuitively and quickly optimize the deployment of software models and improve the development efficiency of the software system.

We design a node artificial intelligence deep neural network training platform and a pedestrian monitoring system, as shown in Fig. 2 and Fig. 3. We develop the software using Python and C++, based on the QT graphic interaction framework, OPENCV, FFMPEG video stream processing framework, Torch, TensorFlow artificial intelligence framework, SQLite lightweight database, Redis cache database, and Milvus vector search engine technology. It comprises a GUI user interaction module, device management module, high-speed video stream processing module, data access module, and artificial intelligence inference module. The architecture design of the software follows the principle of modularization and high scalability. Based on general video monitoring functions, it seamlessly supports AI tasks such as pedestrian detection, face recognition, pedestrian reidentification, face detection and recognition, human behavior recognition, and human recognition. It has an interface design widely compatible with other visual AI tasks.

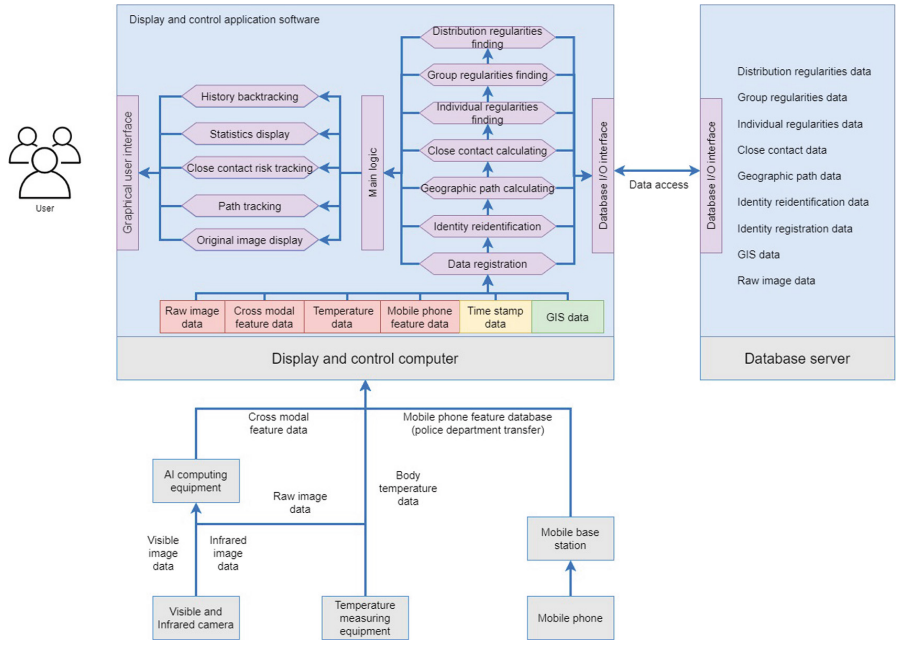



**Fig. 3.** Design of the proposed IoT-based Cross modality Pedestrian Monitoring System for Contact Tracing.

## 3 System Implementation

As shown in Fig. 4, the hardware equipment management module provides the same abstraction layer for multi-modal sensors(such as visible cameras and temperature measuring equipment). It provides a straightforward and convenient equipment management interface, status query interface, video stream control interface, electromechanical control interface, network, and serial port communication interface. To achieve high-speed video stream processing, we use OpenCV and FFMPEG framework. It is responsible for processing the stream fetching, encoding and decoding, and image preprocessing functions of real-time high-speed video streams and is widely compatible with various network camera stream protocols, as well as the encoding, decoding, reading, writing, playing, and control of local video files. Moreover, we use different data serialization methods to provide the system with high flexibility and high-reliability data access support, combining the characteristics of SQLite's convenience and high integration with the high-speed characteristics of the Redis database, as shown in Fig. 5.

The GUI user interaction module is based on the QT graphic framework and Python language binding development, providing a convenient and user-friendly graphical operation interface. As shown in Fig. 6, developers can easily adjust model parameters and network structure through this interactive visual platform to deploy the system quickly Fig. 7.


**Fig. 4.** The hardware implementation of the pedestrian monitoring system.

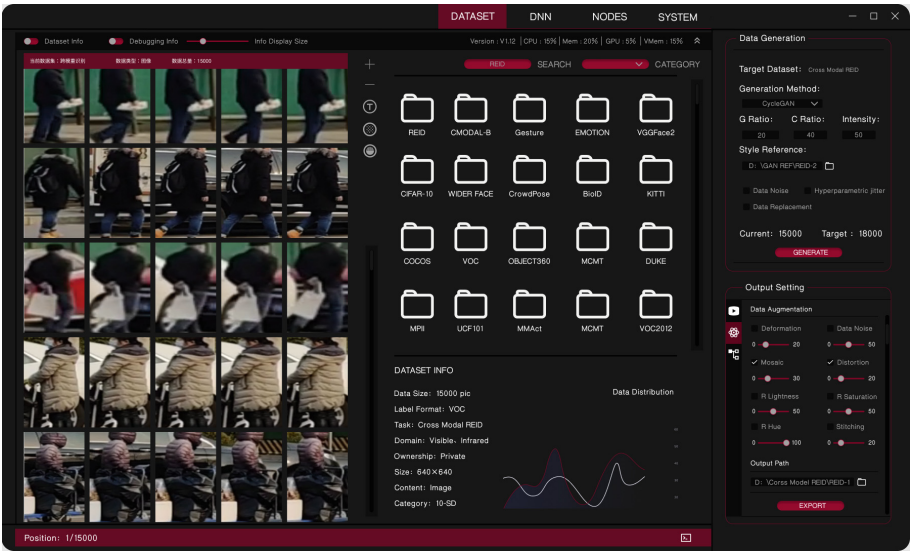



**Fig. 5.** The database of the pedestrian monitoring system.

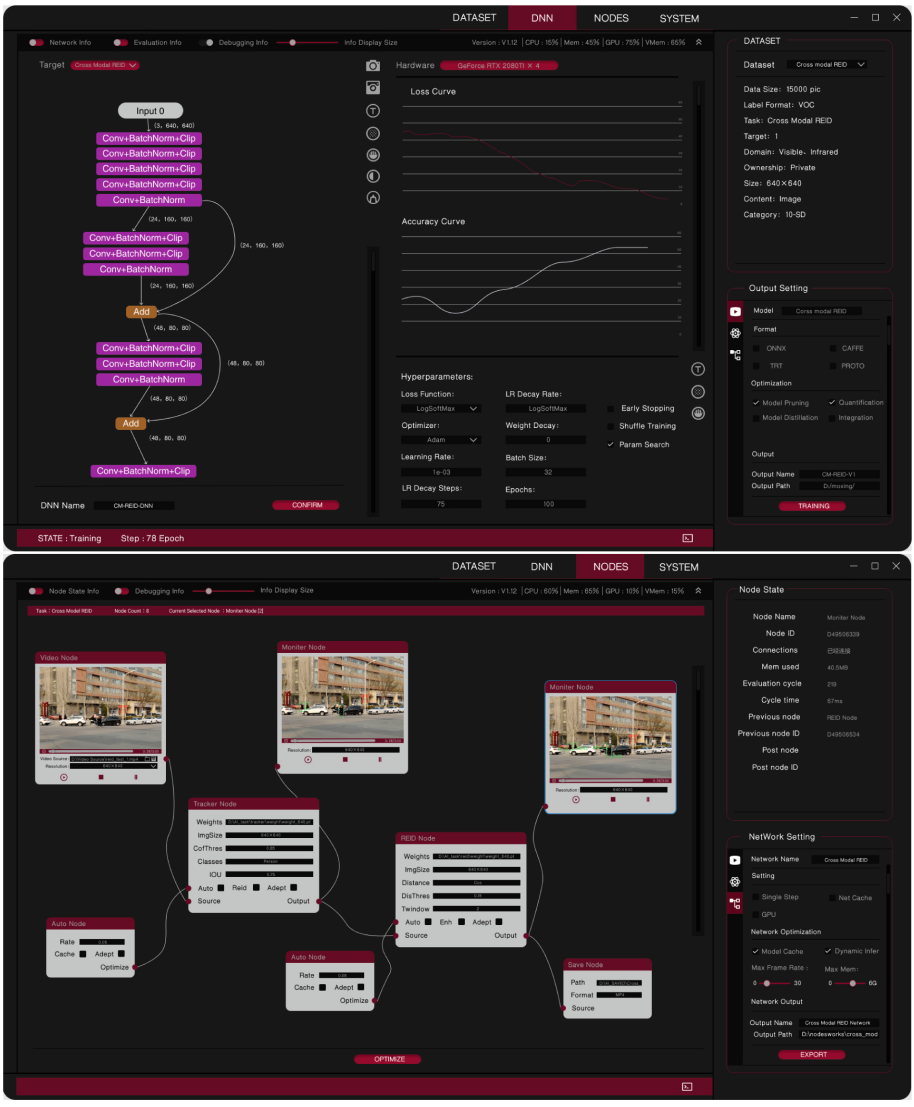



**Fig. 6.** The implementation of the node artificial intelligence deep neural network training platform.

The AI inference module is developed based on Torch and TensorFlow AI frameworks and widely supports various visual AI algorithm models. Among them, the human posture estimation function is developed based on the algorithm of the AlphaPose model [5]; the Face recognition function is developed based on the InsightFace model [6]; The infrared-visible cross-modal pedestrian re identification function is developed based on the HAT model [7]. The function

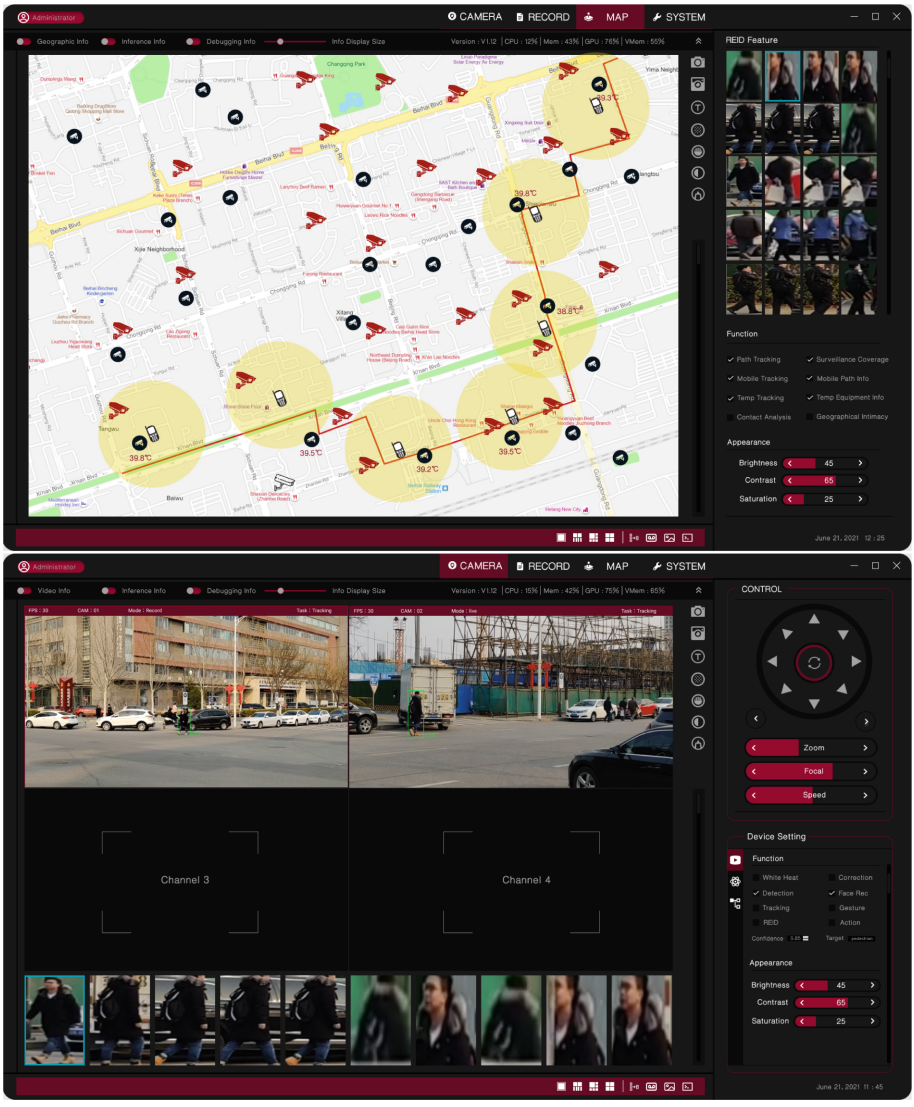



**Fig. 7.** The implementation of the pedestrian monitoring system.

module provides an abstract interface that seamlessly upgrades the supported visual AI task models and is widely compatible with new AI function models.

## 4 Conclusion

Developing the Pedestrian Monitoring System for Contact Tracing is significant for epidemic prevention and control. The proposed system was developed using

state-of-art techniques on the Internet of Things and Artificial Intelligence. The application of visible light, infrared, and other multi-modal data makes it possible to monitor and track pedestrians under all-weather and non-cooperative conditions. The scientific contribution of this work was to use innovative techniques to design and implement an IoT-based cross-modality pedestrian surveillance system. We have also developed a nodal visual artificial intelligence software platform for deep neural network training and optimization, which can be applied in new developments, allowing for fast prototyping. In the future, this research has the task of conducting validation experiments with additional sensors. Besides, this work can be used in developing research in areas such as monitoring in agriculture fields, traffic surveillance, and border defense.

**Acknowledgments.** This research was funded by the National Key Laboratory Foundation of China grant number TCGZ2020C004 and 202020429036.

## References

1. Zhao, D., Yao, F., Wang, L., et al.: A comparative study on the clinical features of COVID-19 pneumonia to other pneumonias. Clin. Infect. Dis. **71**(15), 756–761 (2020)
2. Aryal, A., Becerik-Gerber, B.: Skin temperature extraction using facial landmark detection and thermal imaging for comfort assessment. In: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, pp. 71–80 (2019)
3. Garg, L., et al." Anonymity preserving IoT-based COVID-19 and other infectious disease contact tracing model. IEEE Access **8**, 159 402–159 414 (2020)
4. Israeli Health Ministry: Hamagen (2020). https://govextra.gov.il/ministry-of-health/hamagen-app/download-en/. Accessed 1 Jan 2021
5. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: HybrIK: a hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In: Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3383–3393 (2021)
6. Guo, J., Deng, J., Lattas, A., et al.: Sample and computation redistribution for efficient face detection (2021)
7. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Trans. Inf. Forensics Secur. **16**, 728–739 (2020)

# Low-Light Image Enhancement Algorithm Based on the Fusion of Multi-scale Features and Attention Mechanism

Youchen Sun[1,2], Baoju Zhang[1,2(✉)], Bo Zhang[1,2], Cuiping Zhang[1,2], and Jin Zhang[1,2]

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
wdxyzbj@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** Aiming at the problems of Retinex-Net such as large noise of reflection component, low brightness of illumination component and insufficient feature extraction, a low-light image enhancement algorithm based on fusion of multi-scale features and attention mechanism is proposed. First, the Retinex-Net network is used as the basic model to decompose the input image, and the atrous convolution and ordinary convolution are fused to achieve multi-scale feature extraction to obtain more detailed information; multi-layer attention is introduced into the enhanced network. The force mechanism module enhances the brightness of the details and illumination components; finally, the denoised reflection components and the enhanced illumination components are fused into a normal illumination image output. Experiments show that the algorithm can effectively improve the details of the image and improve the visual effect of the image.

**Keywords:** Low-light image enhancement · Retinex-Net · multiscale feature extraction · attention mechanism · convolutional neural network

## 1 Introduction

In the information age, high-quality images are critical for many computer vision tasks. In reality, good lighting helps to get high-quality images. Conversely, low light results in less useful information being obtained from the image. During the shooting process, due to ambient light intensity and technical limitations, the images obtained by imaging devices such as cameras have problems such as low quality, low contrast, and brightness, so the loss of important image information is not conducive to subsequent feature extraction [1]. To effectively obtain high-quality images from low-light environments, scholars have proposed many low-light image enhancement algorithms. The current research methods for low-light image enhancement can be roughly divided into two categories: traditional methods and deep learning.

This work was supported in part by 2021 Tianjin Postgraduate Research and Innovation Project 2021YJSS209.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 65–72, 2023.
https://doi.org/10.1007/978-981-99-2362-5_9

Compared with traditional methods, deep learning-based methods have better accuracy, robustness, and speed. Therefore, in recent years, researchers have also proposed many methods based on deep learning. Based on the theory of the Retinex algorithm, researchers will combine the Retinex algorithm with convolutional neural networks to improve the visual effect of images, automatically learn the features of images, and solve the problem that Retinex relies on manually setting parameters [2]. Therefore, deep Retinex-based methods have better performance in most cases [3].

Based on the network framework of Retinex-Net, this paper proposes a low-light image enhancement algorithm based on the fusion of multi-scale features and attention mechanisms. Through the fusion of hole convolution and ordinary convolution in the decomposition network, multi-scale feature extraction is realized, and more detailed information of the attention map and reflection map is obtained; the multi-layer attention mechanism module is introduced into the enhancement network to measure the brightness of the illumination component. For enhancement, the improved network improves the details of the image, enhances the visual effect of the image, and improves the overall quality of the image.

## 2 Related Work

### 2.1 Retinex-Net Network Improvement

According to the problems of large reflection component noise, low illumination component brightness, and blurred image in the Retinex-Net network, this paper improves based on the Retinex-Net network and proposes a method based on the fusion of multi-scale features and attention mechanism. Low-light image enhancement algorithm. The algorithm mainly includes two sub-networks trained independently of each other, namely the decomposition network and the enhancement network. In the decomposition network, atrous convolution is used to improve the receptive field of the network without increasing the model parameters, and to obtain illumination and reflection components with more detailed information. Secondly, multiple convolutional block attention models (CBAM) are introduced into the enhancement network to enhance the details of the illumination components and reflection components, and guide the network to correct the illumination components; the reflection map is denoised by the Block Matching 3D(BM3D) algorithm. Finally, image reconstruction is performed, and the enhanced image is obtained by multiplying the processed illumination component and the denoised reflection component. The overall network structure of this paper is shown in Fig. 1. Next, the main parts of the model will be introduced in detail.

### 2.2 Decomposing the Network

In the decomposition network (Dceom-Net), the input low-light $S_{low}$ and normal-light images $S_{normal}$ are decomposed into reflection components $R_{low}$ and $R_{normal}$ and illumination components $I_{low}$ and $I_{normal}$. In order to obtain more feature information, the decomposition network combines the atrous convolution with the ordinary convolution, and the atrous convolution can expand the characteristics of the receptive field without increasing the number of effective units of the convolution kernel. By combining

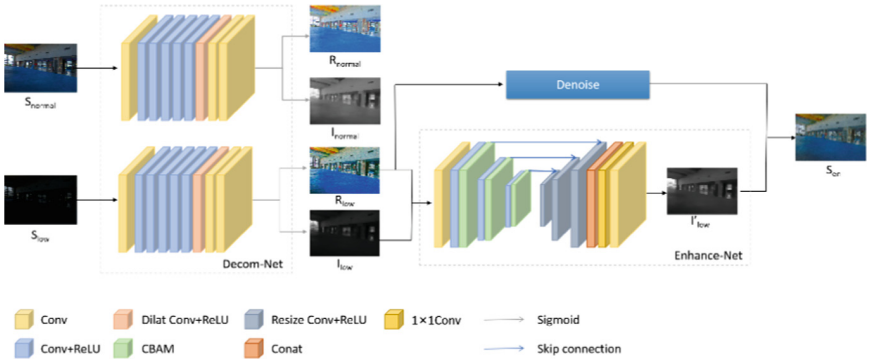



**Fig. 1.** Overall network structure

ordinary convolutions, feature extraction at different scales can be achieved. Therefore, while atrous convolution is used to perform sparse feature extraction on image features, ordinary convolution with the same size of receptive field is used to densely extract the features of the input image, and finally the features output by two different convolution methods are processed in the channel dimension. Stitching.The decomposition network module is shown in Fig. 1.

The loss function LDecom of the decomposition network consists of three parts: reconstruction loss Lrecon, reflection consistency loss Lir and illumination smoothing loss Lis:

$$L_{Decom} = L_{recon} + \lambda_{ir} L_{ir} + \lambda_{is} L_{is} \tag{1}$$

where $\lambda_{ir}$ and $\lambda_{is}$ are the coefficients for reflection consistency and lighting smoothness, respectively. When L1, L2, and SSIM losses are chosen, the L2 norm does not correlate well with the perception of human visual image quality and tends to be locally minimized during training. Although SSIM can better understand the structural features of the image, it is less sensitive to errors in smooth regions, resulting in chromatic aberration [4]. Therefore, this paper chooses to use the L1 norm as a constraint to constrain the loss.

Since each of the decomposition networks $R_{low}$ and $R_{normal}$ can be reconstructed with their respective light maps, the specific form of the reconstruction loss is:

$$L_{recon} = \|R_{low} \cdot I_{low} - S_{low}\|_1 + \|R_{normal} \cdot I_{normal} - S_{normal}\|_1 \tag{2}$$

The reflection consistency loss is used to make the reflection components of the low-illuminance image and the normal-illumination image as consistent as possible. The specific performance of the reflection consistency loss is::

$$L_{ir} = \|R_{low} - R_{noemal}\|_1 \tag{3}$$

To ensure that the illumination component keeps the overall smoothness and the structure and local details of the image, the structure-aware smoothing loss [5] is used as the illumination smoothing loss Lis, which is specifically expressed as:

$$L_{is} = \left\|\nabla I_{low} \cdot \exp\left(-\lambda_g \nabla R_{low}\right)\right\|_1 + \left\|\nabla I_{noemal} \cdot \exp\left(-\lambda_g \nabla R_{normal}\right)\right\|_1 \tag{4}$$

Among them, $\nabla$ represents the gradient in the horizontal and vertical directions of the image, and $\lambda_g$ represents the coefficient of balancing the perceived strength of the structure. The weights $\left(-\lambda_g \nabla R\right)$ allow Lis to relax the smoothness constraints when the gradient of the reflection component is large, even if the smoother areas in the reflection component match the areas corresponding to the lighting component, ensuring that texture details and boundary information in the smoothness constraints are not destroyed.

### 2.3 Enhanced Network

The enhancement network uses the overall framework of the encoder-decoder structure, combined with the U-Net network structure, to enhance the illumination component Ilow obtained by the decomposition network, and perform BM3D noise suppression processing for the reflection component $R_{low}$. The specific structure is shown in Fig. 1.

Aiming at the problems of color distortion, low brightness, and insufficient details in the illumination images obtained by the enhancement network, an attention mechanism Convolutional Block Attention Module(CBAM) is introduced to enhance and correct the details of the illumination images.The structure diagram of CBAM is shown in Fig. 2.

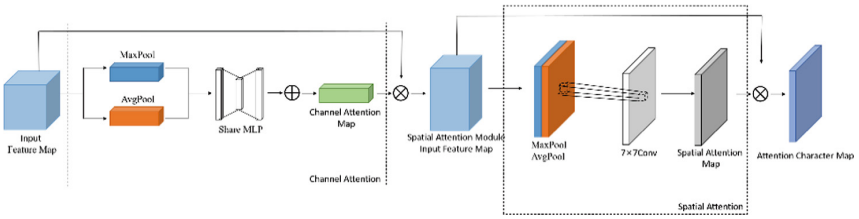



**Fig. 2.** CBAM structure diagram

The enhancement network first performs a connection operation on the illumination components and reflection components obtained in the decomposition stage and then performs a 3 × 3 convolution operation, performs channel selection on the input low-illuminance map illumination component Ilow, and passes the input feature map through two parallel The global maximum connection layer and the intermediate layer are transmitted, and then through the shared network (Share MLP) module, through the activation function ReLU to obtain two activated results. The two output results are added together, and then a sigmoid activation function is used to obtain the output result of the channel module, and then the output result is multiplied by the original image to obtain the output result of the channel module. The output of the channel module is obtained by max pooling and average pooling to obtain two feature maps, splicing the two feature maps, and convolution through a standard convolution layer to obtain a spatial attention map. After another sigmoid, the final attention feature map is obtained. After passing through the attention module, the image is enhanced by the nearest neighbor interpolation method, and up-sampling is performed to ensure that the size of the combined feature map is consistent, and the corresponding sums are added, and then the feature fusion is performed to obtain a feature map with more complete details, and

finally, the network is fine-tuned end-to-end using stochastic gradient descent to obtain augmented images [6].

The loss Len of the augmentation network includes the reconstruction loss $L_{recon}$ and the illumination smoothing loss $L_{is}$. The reconstruction loss is expressed as the multiplication of the denoised reflection component and the corrected illumination component to reconstruct the image to obtain an enhanced image.

The reconstruction loss in the augmented network is $L_{recon} = \left\| \hat{R}_{low} \cdot \hat{I}_{low} - S_{normal} \right\|_1$. $\hat{R}_{low}$ is the reflection component after denoising, $\hat{I}_{low}$ is the illumination component weighted by the $R_{low}$ gradient map, and is the normal illumination image. The calculation of the illumination smoothing loss $L_{is}$ is the same as that of Eq. (4).

While enhancing the illumination image, denoise the decomposed reflection image, and use the BM3D algorithm to suppress the amplified noise in $R_{low}$. Finally, the enhanced illumination image and the denoised reflection image are multiplied element by element to achieve image reconstruction and form the final enhanced image.

## 3 Experimental Design and Results

### 3.1 Dataset

The dataset in this paper is divided into two parts: the real scene dataset LOL and the artificial synthetic dataset Brighting Train [5]. The LOL dataset is a dataset of image pairs collected from real scenes for low-light enhancement. The dataset captures images from multiple real scenes, including 500 pairs of images of the furniture, streets, buildings, etc., of which 485 pairs of images are used as a training set and 15 pairs of images are used as a test set. In RAISE [7], 1000 original images were used to synthesize low-light images, forming 1000 pairs of artificially synthesized datasets as the training set, and named Brighting Train. To make the experimental results more accurate, the training samples are rotated, translated, cropped, and other operations to perform data enhancement preprocessing on the images of the data set, to obtain more data image sets.

### 3.2 Experimental Environment and Parameter Settings

The training and testing experiments of the overall network in this paper are completed based on the PyTorch framework on the NVIDIA GeForce 1080Ti GPU device, using CUDA11.6 to accelerate it, and using the Adam optimizer to optimize the loss function. The loss function coefficients $\lambda_{ir}$ and $\lambda_{is}$ are set to 0.001 and 0.1, respectively. The balanced structure perception strength factor $\lambda_g$ is 10. During network training, the batch size batch_size is set to 16 and the image patch size is set to $48 \times 48$. The training batch epoch is set to 100. The initial value of the learning rate is 0.001, and the number of iterations is selected to perform a learning rate decay every 20 iterations, that is, a decay of 5 times every 20 epochs.

### 3.3 Comparative Experiment

To verify the performance and effectiveness of the algorithm in this paper, a comparison experiment was conducted with other low-light image enhancement methods, including the EnlightenGAN algorithm [8], the MBLLEN algorithm [9], the Retinex-Net algorithm [5], and the RRDNet algorithm [10], Zero-DCE algorithm [11], 4 low-light images were selected in the LOL dataset and compared with five algorithms to reflect the superiority of the improved algorithm in this paper. The comparison results are shown in Fig. 3.


**Fig. 3.** Experimental comparison of each algorithm in the LOL dataset

As can be seen from Fig. 3, in the EnlightenGAN algorithm, through the fourth low-light image, it can be found that there are obvious shadows in the image, such as the shadow at the intersection of the book on the desk and the desk is more obvious. The MBLLEN algorithm is bright as a whole, which makes the picture overexposed. For example, the brightness of the stainless steel kitchen utensils in the second picture is high, which is more obvious. The Retinex-Net algorithm produces a blurring phenomenon in the comparison image, the image is noisy, the color is slightly distorted, the edges of some objects are too prominent, and artifacts appear in some areas. The RRDNet algorithm has insufficient brightness enhancement, and the overall image is dark, resulting in the phenomenon of underexposure and serious problems in image denoising, and image sharpening is serious. In the Zero-DCE algorithm, the brightness of the enhanced image is slightly lower and the color saturation is weaker. The overall brightness of the image enhanced by the algorithm in this paper is more in line with the visual perception of the human eye be improved. Therefore, from a subjective point of view, the experimental results of our algorithm are better than other algorithms.

Since people have different visual and sensory preferences, comparisons using only visual quality are one-sided. Therefore, this paper adopts two evaluation indicators, PSNR and SSIM, to objectively evaluate the image quality. The experimental results are shown in Table 1. As can be seen from Table 1, the algorithm in this paper has achieved good results in the PSNR and SSIM scores, with the highest PSNR value and the second highest SSIM value. The PSNR value of the algorithm in this paper is 0.72dB higher than that of the MBLLEN algorithm with the second highest PSNR value, and the SSIM value is only 0.036 lower than the highest value. In summary, the results of the algorithm in this

paper have been significantly improved, which proves the effectiveness of the algorithm in low-light image enhancement, and the improved network can effectively improve the processing capability of low-light images.

**Table 1.** Average objective evaluation of different algorithms on the LOL dataset

| Algorithm | PSNR↑ | SSIM↑ |
|---|---|---|
| EnlightenGAN [8] | 17.483 | 0.677 |
| MBLLEN [9] | 17.902 | **0.715** |
| Retinex-Net [5] | 16.774 | 0.462 |
| RRDNet [10] | 11.392 | 0.468 |
| Zero-DCE [11] | 16.796 | 0.589 |
| Ours | **18.622** | 0.679 |

## 4 Conclusion

Aiming at the problems of Retinex-Net such as large noise of reflection component, low brightness of illumination component and insufficient feature extraction, a low-light image enhancement algorithm based on fusion of multi-scale features and attention mechanism is proposed. The algorithm in this paper combines atrous convolution with ordinary convolution in the decomposition network to achieve multi-scale feature extraction, so as to obtain illumination components and reflection components with more detailed information. The attention mechanism CBAM module is introduced into the enhancement network to enhance and correct the brightness of the illumination components. Finally, the denoised reflection component and the enhanced illumination component are constructed into a normal illumination image. On the LOL data set, the algorithm in this paper is compared with some other advanced algorithms from both subjective and objective aspects. The experimental results show that the algorithm in this paper improves the brightness of the image, the sharpness and texture details are significantly improved, and finally the quality of the image is improved.

## References

1. Li, C., Guo, C., Han, L.H., et al.: Low-light image and video enhancement using deep learning: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **01**, 1 (2021)
2. Ou, J.M., Hu, X., Yang, J.X.: Low-Light Image enhancement algorithm based on improved retinex-net. Pattern Recogn. Artif. Intell. **34**(01), 77–86 (2021)
3. Wang, M.M., Peng, D.L.: Retinex-ADNet: a low-light image enhancement system. J. Chin. Comput. Syst. **43**(02), 367–371 (2022)
4. Zhao, H., Gallo, O., et al.: Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imaging **3**, 27–47 (2017)

5. Wei, C., Wang, W., Yang, W., et al.: Deep retinex decomposition for low-light enhancement (2018). arXiv preprint arXiv:1808.04560
6. Liu, J.M., He, N., Yin, X.J.: Low lllumination image enhancement based on retinex-UNet algorithm. Comput. Eng. Appl. **56**(22), 211–216 (2020)
7. Dang-Nguyen, D.T., Pasquini, C., Conotter, V., et al.: RAISE: a raw images dataset for digital image forensics. In: Proceedings of the 6th ACM Multimedia Systems Conference, pp. 219–224 (2015)
8. Jiang, Y., et al.: Enlightengan: deep light enhancement without paired supervision (2019). arXiv preprint arXiv:1906.06972
9. Lv, F., Lu, F., Wu, J., Lim, C.: Mbllen: low-light image/video enhancement using cnns. In: BMVC (2018)
10. Zhu, A., Zhang, L., Shen, Y., Ma, Y., Zhao, S., Zhou, Y.: Zero-shot restoration of underexposed images via robust retinex decomposition. In: ICME, pp. 1–6 (2020)
11. Guo, C., et al.: Zeroreference deep curve estimation for low-light image enhancement. In: CVPR, pp. 1780–1789 (2020)

# Research on Improved Image Classification Algorithm Based on Darknet53 Model

Shangchen Zou[1,2], Baoju Zhang[1,2(✉)], and Bo Zhang[1,2]

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
wdxyzbj@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** An improved Image classification algorithm based on Darknet53 model is proposed in this paper to solve the problem that a large amount of the same gradient information is repeatedly used to update the weights of different dense layers in the Darknet53 network in the process of back propagation. Firstly, the residual block in the original network is replaced by THE CSP module, referring to the idea of cutting off the gradient flow in CSPnet to prevent too much repeated gradient information. Second, the Mish activation function is used to replace the Leaky ReLU function, which can transmit information more smoothly and achieve better accuracy and generalization. Finally, a SPP module is added to the end of the original network structure to solve the multi-scale problem of the main part of image classification. The experimental results show that the improved Darknet has achieved better performance, and the accuracy is improved by 1.3% compared with the original network; At the same time, compared with resnet50 and resnet101, the improved Darknet has better effect on image classification.

**Keywords:** Image classification · Deep learning · CNN

## 1 Introduction

Image classification is one of the research hotspots in the field of computer vision in recent years. It is also the basic algorithm for other applications such as target detection [1], face recognition [2], and pose estimation [3]. Therefore, image classification technology has high academic research and practical application value. Image classification, that is, given an input image, the classification algorithm is used to determine the category of the image. This paper only focuses on the relevant theories of deep learning. The following is a brief introduction of some important and common image classification networks.

LeCun et al. [4] proposed Lenet-5 network, which is a classical convolutional neural network containing some important features, which are still the core of CNN network up to now; Krizhevsky et al. [5] proposed AlexNet network with five convolutional layers. It uses ReLU activation function to solve the gradient dispersion problem caused by Sigmoid when the network is deep. Dropout technology is also introduced to reduce network


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 73–80, 2023.
https://doi.org/10.1007/978-981-99-2362-5_10

overfitting and the calculated quantity of training models. Simonyan et al. [6] explored the relationship between depth and performance of convolutional neural network on the basis of AlexNet network and proposed VGG network, which realized 16–19 layers deep convolutional neural network by repeatedly stacking 3 × 3 small convolutional kernel and 2 × 2 maximum pooling layer, and proved that the deeper the neural network, the better the effect. He et al. [7] proposed ResNet network, namely ResNetV1, which solved the degradation of deep network training by setting shortcut structure. Szegedy et al. [8] proposed GoogLeNet network from the direction of optimizing convolutional neural network structure and reducing network complexity, which is cascaded by several Inception modules, uses a global pooling layer to replace the final full connection layer, and adds an auxiliary classifier. Huang et al. [9] proposed a DenseNet network based on ResNet, which connects each layer of the network with all the previous layers in a feedforward manner, and designs each layer to be narrow to reduce redundancy and ensure accuracy while reducing network parameters.

This paper proposes an improved image classification algorithm based on Darknet53 model. By changing the activation function, replacing the residual module of feature extraction layer, and adding SPP module at the end of the network, The accuracy of the improved network with deeper convolution layers has also been improved on the basis of maintaining the original processing speed.

## 2 The Improved Darknet

### 2.1 Darknet53 Model

DarkNet53 is an image feature extraction backbone network proposed by Joseph Redmon [10]. The network consists of five block units, each of which contains residual units with repeated structures (each convolution layer is actually Conv + BN + Leaky_Relu), which are repeated 1, 2, 8, 8 and 4 times respectively. All convolution layers total 53 layers, making the network structure relatively complex. The network model combines features of deep residual networks ResNet and DarkNet19 to achieve classification accuracy comparable to the most advanced classifiers with fewer floating point operations and faster speeds. In addition, the network replaces all pooling layers in feature extraction layer with convolution layer with convolution kernel size of 3 × 3 and step size of 2, which not only achieves the purpose of down-sampling, but also increases the network depth and further improves the accuracy.

### 2.2 Mish

Mish [11] is a self-regularized non-monotone neural activation function whose mathematical expression is:

$$f(x) = x * \tanh(ln(1 + e^x)) \tag{1}$$

As an activation function, it has the following three characteristics: 1. No upper limit but lower limit; 2. Smooth; 3. Non-monotonic. There is a lower limit which can guarantee the regularization effect of the network. Mish is smoother than Relu activation function,

which makes the network better at solving and model generalization. The non-monotonic property allows the small negative input to maintain a certain negative output so as to improve the expressiveness of the network and promote the propagation of gradient flow.

During the experiment, the network structure of the original darknet53 model is maintained, and only the activation function in the convolution layer is changed by leaky_Relu is all replaced by Mish, and a comparative experiment is done on cifar10 dataset. The experimental results show that after the activation function is replaced by Mish, the accuracy of image classification network increases by 0.4%.

### 2.3 CSP Structure

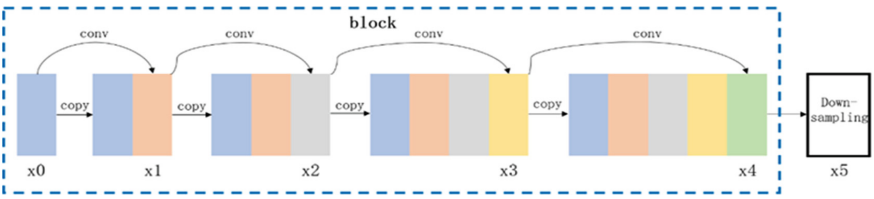



**Fig. 1.** Block in Darknet53

Each stage in the Darknet53 network consists of a segment of repeatedly stacked residual modules and a convolutional layer used for down-sampling, as shown in Fig. 1 (block5 as an example). Then the relation between input and output can be expressed as:

$$\begin{aligned} &x1 = w1 * x0, x2 = w2 * [x0, x1], \\ &x3 = w3 * [x0, x1, x2], x4 = w4 * [x0, x1, x2, x3] \end{aligned} \tag{2}$$

where wi represents the weight of each convolution layer and * represents the convolution operation. If back propagation is used to update weight parameters, the weight update equation can be written as:

$$\begin{aligned} &w1' = f(w1, g0), w2' = f(w2, g0, g1), \\ &w3' = f(w3, g0, g1, g2), w4' = f(w4, g0, g1, g2, g3) \end{aligned} \tag{3}$$

where f is the weight update parameter, and gi represents the gradient propagating to the i− th convolution layer in the block. It can be seen from the above formula that a large amount of gradient information is repeatedly used to update the weights of different convolution layers, which will cause different convolution layers to repeatedly learn the same gradient information.

In order to solve the above problems, referring to the idea of CSPnet, this paper replaces the block in the original darknet53 network with CSP_ block. The principle of CSP structure is shown in Fig. 2. It divides the input characteristic graph x0 into $x0'$ and $x0''$. The former is directly connected to the end of the stage, and the latter is stacked

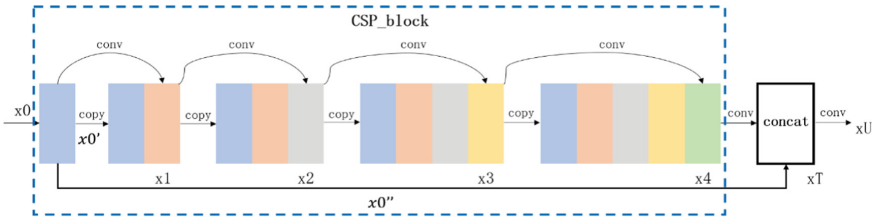



**Fig. 2.** CSP_block

with x0′ in depth at the end through the residual blocks stacked in the block, and then output to the next stage.

The feedforward equation of the network structure is:

$$\begin{aligned} &\mathrm{xk} = \mathrm{wk} * [\mathrm{x0}', \mathrm{x1}, ..., \mathrm{xk} - 1], \mathrm{xT} = \mathrm{wT} * [\mathrm{x0}', \mathrm{x1}, ..., \mathrm{xk}], \\ &\mathrm{xU} = \mathrm{wU} * [\mathrm{x0}'', \mathrm{x}T] \end{aligned} \tag{4}$$

The reverse transfer weight update equation is:

$$\begin{aligned} &\mathrm{wk}' = \mathrm{f}(\mathrm{wk}, \mathrm{g0}', \mathrm{g1}, ..., \mathrm{gk} - 1), \mathrm{w}T' = \mathrm{f}(\mathrm{w}T, \mathrm{g0}', \mathrm{g1}, ..., \mathrm{gk}), \\ &\mathrm{wU}' = \mathrm{f}(\mathrm{wU}, \mathrm{g0}'', \mathrm{g}T) \end{aligned} \tag{5}$$

It can be seen from the above two equations that the gradient from CSP_block is integrated separately, and the gradient of input characteristic diagram x0″ without CSP_block is also integrated separately. That is, wT″ and wU′ do not contain each other's repeated gradient information. Therefore, CSP structure retains the advantage of reusing the original network features, prevents the use of too much repeated gradient information by truncating the gradient flow, shortens the path of gradient backflow, and obtains richer gradient combinations while reducing the amount of calculation. Based on this theory, this paper replaces the block in the original darknet53 with the CSP_block network structure in Fig. 3.

During the experiment, this paper only replaces the block of darknet53 network with the CSP_block structure described above, keeping other parameters and experimental environment unchanged. A comparative experiment is done on cifar10 dataset. The experimental results show that after replacing the block structure, the accuracy of image classification network increases by 0.3%, and the network training time and processing speed remain basically unchanged.

### 2.4 SPP Structure

In the original darknet53 network structure, the size of convolution kernel of each layer is fixed, that is, the receptive field is also fixed, and the size of the main body in the input image varies, so the classification accuracy of the main body with large size difference is very unstable in the classification process. Based on the above reasons, this paper adds SPP (spatial pyramid pooling) structure to increase receptive field and solve such problems before the network end pooling layer and full connection layer (Fig. 4).

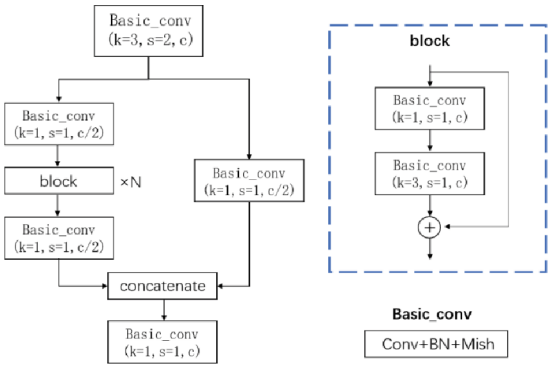



**Fig. 3.** CSP_ Block network structure

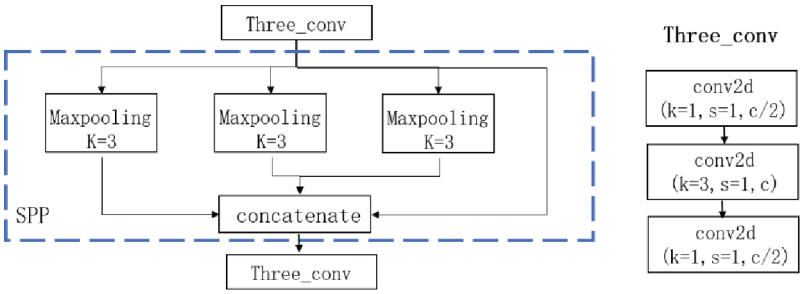



**Fig. 4.** SPP network structure

During the experiment, keeping the original darknet53 model and other network parameters unchanged, this paper only adds a layer of SPP structure described above before the end full connection layer, and makes a comparative experiment on cifar10 dataset. The experimental results show that after adding SPP structure, the accuracy of image classification network increases by 0.5%, and the network training time is slightly prolonged.

### 2.5 Improved Darknet Classification Network Structure

On the basis of the previous article, the improved Darknet model proposed in this paper is shown in Fig. 5. The model includes five CSP_block residual module, 3 transitional convolution layers and 1 SPP module. CSP_block and the previous convolution layer include convolution, BN and Mish. SPP module and its over convolution layer include convolution, BN and Leaky_Relu activation functions. The reason for this design is that in the control experiment, it is found that setting the activation function of the terminal excessive convolution layer as the leaky_Relu function has a higher accuracy of classification.

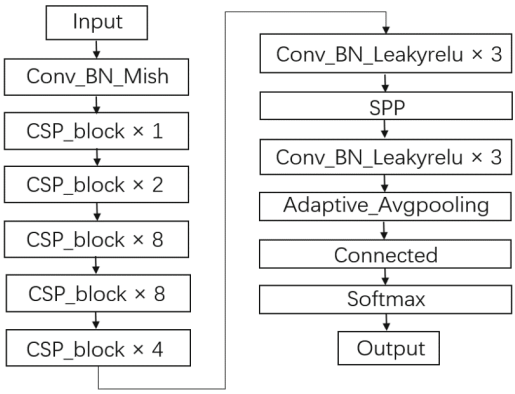



**Fig. 5.** Improved Darknet network model

## 3 Experimental Results and Analysis

### 3.1 Dataset

This paper adopts the publicly available cifar-10 dataset, which has a total of 10 categories of pictures, 6000 pictures in each category, a total of 60000 pictures, and the size of each picture is $32 \times 32 \times 3$. The dataset is divided by a ratio of 5:1, with 50000 pictures as the training set and 10000 as the verification set.

### 3.2 Evaluation Function

In order to evaluate the performance of the algorithm, this paper directly adopts the training-verification method to calculate the accuracy of the classification network in the verification set. The accuracy of the validation set is used to evaluate the performance of the model, and the formula is:

$$\text{Accuracy} = \text{R}/\text{T} \tag{6}$$

where: R is the sample with correct classification, and T is the total samples of validation set.

### 3.3 Experimental Result

Adam optimization algorithm is selected in this experiment, the initial learning rate is set to 0.0001, and the learning rate decays by 75% every 20 training times. The batch size for one training is 32. The loss function uses the cross entropy loss function commonly used in classification problems. During the training process, draw the loss value of the training set and the accuracy curve of the verification set respectively to observe the degree of model training. When the loss function of the training set tends to 0, it indicates that the model has converged. The experimental results show that when the number of iterations of the training set reaches 200, the effect of the model is the best.

The algorithm in this paper is compared with the original darknet53 network and resnet50 and resnet101. The results in Table 1 show that the accuracy of the improved Darknet network is improved by 1.3% compared with the original structure, and it also performs better than RESNET network.

**Table 1.** Comparison with the original network and other network structure

| Method | Batch_size | Accuracy |
|---|---|---|
| resnet50 | 32 | 0.761 |
| resnet101 | 32 | 0.770 |
| Darknet53 | 32 | 0.828 |
| Improved Darknet | 32 | 0.841 |

## 4 Conclusion

This paper proposes an improved network to solve the problem that a large amount of the same gradient information is repeatedly learned in darknet53. Firstly, the residual block in the original network is replaced by CSP_block by referring to the idea of cutting off the gradient flow to prevent too much repeated gradient information in CSPnet; Secondly, the Mish activation function is used to replace the Leaky_Relu function, so that the network can transfer information more smoothly, so as to obtain better accuracy and generalization; Finally, SPP module is added at the end of the original network structure to solve the multi-scale problem of the main part in image classification. The experimental results show that the accuracy of the improved Darknet model is improved by 1.3% compared with the original network. Subsequently, this paper will conduct promotion research on the target detection network, and strive to further improve the efficiency and accuracy of target detection by improving the feature extraction backbone network and optimizing the feature path aggregation and loss function.

## References

1. Ouyang, W., Zeng, X., Wang, X., et al.: DeepID-Net: object detection with deformable part based convolutional neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(7), 1320–1334 (2017)
2. Hu, G., Yang, Y.X., Yi, D., et al.: When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In: International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015, pp. 142–150. IEEE Press, Piscataway (2015)
3. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: HigherHRNet: scale-aware representation learning for bottom-up human pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5385–5394 (2020). https://doi.org/10.1109/CVPR42600.2020.00543
4. Lecun, Y., Bottou, L., Bengio, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)

5. Krizhevsky, A., Sutskever, I.E., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 3–6 December 2012, pp. 1097–1105. Curran Associates Inc., Red Hook (2012)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference of Learning Representation, San Diego, CA, 7–9 May 2015 (2015). arXiv:1409.1556v6 [cs.CV]
7. He, K.M., Zhang, X.Y., Ren, S.Q., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, 27–30 June 2016. IEEE Computer Society, Los Alamitos (2016)
8. Gedy, C., Liu, W., Jia, Y.Q., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 July 2015, pp. 1–9. IEEE Press, Piscataway (2015)
9. Huang, G., Liu, Z., Maaten, L.V.D., et al.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 21–26 July 2017, pp. 2261–2269. IEEE Press, Piscataway (2017)
10. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement.arXiv Preprint arXiv: 1804.02767 (2018)
11. Misra, D.: Mish: a self regularized non-monotonic neural activation function (2019)

# Location-Aware Heterogeneous Graph Neural Network for Region Recommendation

Liantao Bai[1], Yaxing Liu[1], Jun Wang[2(✉)], and Hengpeng Xu[1(✉)]

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
xuhengpeng@mail.nankai.edu.cn

[2] College of Mathematics and Statistics Science, Ludong University, Yantai, Shandong 264025, China
junwang@mail.nankai.edu.cn

**Abstract.** With the diversification of human activity and travel demand in urban space, recommending ROIs (region-of-interest) to users is important for both satisfying commercial demands and better understanding user urban lifestyles. Current researches mainly resort to the traditional POI-level (point-of-interest) or neural network-based recommendation methods for ROI recommendation, in disregard of the rich heterogeneous graph information, such as user-region-user, region-category-region, just to name a few. In this work, we employ the heterogeneous graph to address this issue, considering heterogeneous graph contains more comprehensive information and rich semantics. We propose a novel meta-path based graph attention network for ROI recommendation, called MRec. MRec is a newly devised heterogeneous graph neural network, which is equipped with both node-level and semantic-level attentions. Specially, the node-level attention aims to learn the importance between a node and its meta-path based neighbors, while the semantic-level attention is to learn the importance of different meta-paths. This mechanism contributes to effectively embedding users and ROIs in a hierarchical manner of fully considering both node and semantic-level component information. An extensive experiment on two real-world datasets demonstrates the effectiveness of the proposed framework.

**Keywords:** ROI recommendation · Graph neural network · Heterogeneous information network · Attention network

## 1 Introduction

The rapid urbanization process has been nurturing big and complex urban space and changing our lifestyles on an unprecedented scale and speed. Recommending ROIs to users, especially those irregular ROIs, is widely recognized as a fundamental task to tackle the urban challenges (e.g., traffic control, public safety,

L. Bai and Y. Liu—Equally contributed to this work.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 81–89, 2023.
https://doi.org/10.1007/978-981-99-2362-5_11

and air pollution). Different urban ROIs may satisfy different social demands, such as the demands of entertainment, dwelling, education, and communications. This functionality divergence of ROIs naturally leads to various user ROI-level mobility patterns. For example, users usually check in the supermarkets in their residential quarters, while office appliances are more popular in business areas. The major focus of this study is how to recommend ROIs to users by modeling ROI-level mobility patterns in different urban ROIs, accompanying with handling the feedback from the various interactions in LBSN (location-based social network) based HIN (heterogeneous information network).

In real-world, users' check-in behaviors induce rich interaction relations, which deeply underlie the LBSN but are critical for ROI recommendation. Figure 1 demonstrates a general user-region interaction scenario which is in the form of heterogeneous graph comprising four types of nodes (a.k.a. entities): users (U), locations (L), regions (R), and category (C). Besides the attribute information, these nodes or objects also share rich interaction information, corresponding to meta-path in HINs. For instance, a U-R-U meta-path indicates two users co-visiting the same ROI, and a R-C-R meta-path denotes two ROIs belonging to the same category.

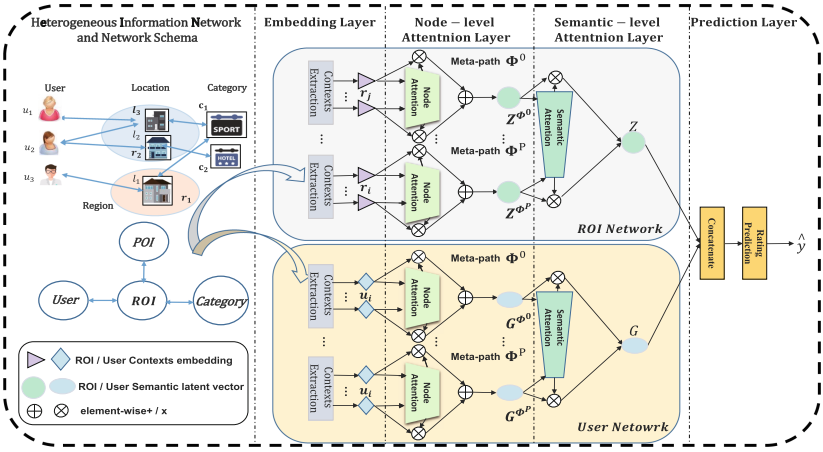



**Fig. 1.** Main architecture of MRec

In this study, we design a novel meta-path based graph neural network with attention mechanism for ROI recommendation (dubbed MRec), mainly addressing the following challenges:

**Rich Context Information.** The decision of checking in a ROI is typically determined by various context information associated with historical check-ins. *To conquer the challenge of jointly capturing rich contextual information and incorporating heterogeneous information*, MRec leverages an embedding layer to incorporate all kinds of contextual factors in a unified manner.

**Node-Level Attention.** A certain number of links in HINs may not convey meaningful semantics, which is extremely common in LBSNs since most ROIs satisfy different social demands. For example, when using the meta-path User-ROI-User as shown in Fig. 1(the users who checked-in the same ROI), user $u_1$ and user $u_2$ both checked-in ROI $r_2$, to better identify the genre of $r_2$ as residential community, the model should pay more attention to $u_2$, rather than $u_1$, when $u_2$ resides at $r_2$.

*To solve the challenge of distinguishing subtle difference between links and selecting the most informative links*, MRec devises the node-level attention to learn the importance of meta-path based neighbors in a heterogeneous graph.

**Semantic-Level Attention.** Different meta-paths embed different semantic information. Given $u_1$ and $u_2$ are online friends and $u_1$ and $u_3$ reside in the same community in Fig. 1, $u_1$ can both connect to $u_2$ via U-R-U and connect to $u_3$ via U-C-U. When identifying the preferences influenced by online friends, U-C-U should be more emphasized than U-R-U when $u_2$ lives in $r_2$. Thus, *to tackle the challenge of finding the most semantically meaningful meta-paths*, MRec introduces the semantic-level attention to learn the importance of meta-paths.

In this paper, we recommend ROIs by jointly utilizing users' static latent states and the latent semantic representations of users and ROIs that are enhanced via meta-path embedding. Our major contributions are as follows:

- This is the first region recommendation study that leverages the heterogeneous graph neural network with attentions to learn ROI's latent semantic representation that is enhanced via meta-path embedding.
- The node-level and semantic-level attention modules are seamlessly fused to comprehensively learn the importance of meta-path based neighborhoods and meta-paths.
- Extensive evaluations on real-world datasets demonstrate that the proposed model outperforms the state-of-the-art recommendation approaches by a considerable margin.

## 2 The Proposed Model

In this section, we introduce our meta-path based graph attentive neural network model MRec. First, we introduce the general architecture of MRec. Then, we present two attention-based layers, which are the main concerns in this work. Next, we propose the prediction layer, which is a neural latent factor model designed for rating prediction. Finally, we go through the optimization details of MRec. Some critical notations are displayed in Table 1.

### 2.1 Model

We propose to address the region recommendation problem under the representation learning (RL) framework. Under the RL paradigm, each user or item is embedded as a vector, which encodes the inherent properties of the user or item (e.g., preferences of a user, spatial-temporal attributes of a region, etc.).

**Table 1.** Notations

| SYM. | DESCRIPTION |
|---|---|
| $h_{r_i}$ | feature embedding of the ROI $r_i$ |
| $Z_{r_i}$ | latent semantic vector of the region $r_i$ |
| $Z^{\Phi}_{r_i}$ | semantic-specific node embedding for the ROI $r_i$ |
| $N^{\Phi}_i$ | meta-path based neighbors of the ROI $r_i$ |
| $\alpha^{\Phi}_{r_{ij}}$ | weight of the meta-path based ROI pair $(r_i, r_j)$ |
| $\beta_{\Phi}$ | weight of the meta-path $\Phi$ |

**Node-level Attention.** The meta-path based neighbors of each node in a heterogenous graph have different importance in learning the semantic embedding of this node. Hence, the goal of the node-level attention in $Net_r$ is to learn the importance of meta-path based neighbors and find the most meaningful ones for the ROI $r$. By the way of aggregating the representations of these meaningful neighbors, we can learn the semantic embedding of $r$ as follows:

$$Z^{\Phi}_{r_i} = \delta\left(W.Aggre_{Node}(\{h_{r_j}, r_j \in N^{\Phi}_i\}) + b\right), \tag{1}$$

where $Aggre_{Node}$ is the aggregation function to combine meta-path based neighbors, $\delta$ denotes the non-linear activation function, and $W$ and $b$ are the weight and bias of a neural network.

A popular formulation of $Aggre_{items}$ is the mean operator wherein we can take the element-wise mean of the vectors on $(\{h_{r_j}, r_j \in N^{\Phi}_i\})$. This mean-based aggregation is a linear approximation as

$$Z^{\Phi}_{r_i} = \delta\left(W.\left\{\sum_{r_j \in N^{\Phi}_i} \alpha^{\Phi}_{r_i} h_{r_j}\right\} + b\right), \tag{2}$$

where $\alpha^{\Phi}_{r_i}$ is fixed to $\frac{1}{|N^{\Phi}_i|}$ for all meta-path based neighbors of the ROI $r_i$, assuming that all neighbors equally contribute to modeling $r_i$. However, this may be not appropriate due to the fact that the meta-path based neighbors show different importance in learning the node embedding. Hence, we allow the neighbors to differently contribute to ROI's latent semantic embedding by assigning them different weights.

Specifically, we tweak $\alpha_{ms}$ to be aware of the target user $u_m$, i,e., assigning an individualized weight for each$(u_m, r_s)$ pair as

$$Z^{\Phi}_{r_i} = \delta\left(W.\left\{\sum_{r_j \in N^{\Phi}_i} \alpha^{\Phi}_{r_{ij}} h_{r_j}\right\} + b\right), \tag{3}$$

where $\alpha_{r_{ij}}$ denotes the attention weight of the neighbor node $r_j$ paired with $r_i$ in contributing to the semantic-specific latent embedding of $r_i$ when given the meta-path $\phi$. In addition, we need to inject the structural information into the

model. Inspired by the work in [2], we parameterize the node-level attention score $\alpha_{r_ij}$ via the graph attention network layer. Formally, the node-level attention score is defined as

$$\alpha_{r_{ij}}^{\Phi} = \frac{\exp\left(\delta(\partial_{\Phi}^{T}.\left(h_{r_i}||h_{r_j}\right])\right)}{\sum\limits_{k\in N_i^{\Phi}} \exp\left(\delta(\partial_{\Phi}^{T}.\left(h_{r_i}||h_{r_k}\right])\right)}. \tag{4}$$

Given the meta-path set $\{\Phi_0, \Phi_1, \ldots, \Phi_P\}$, after feeding the ROI embedding into the node-level attention, we can obtain $P$ groups of semantic-specific node embeddings, denoted as $\{Z_{\Phi_0}, Z_{\Phi_1}, \ldots, Z_{\Phi_p}\}$.

**Semantic-Level Attention.** Each user or ROI in a heterogeneous graph possesses multiple kinds of semantics information since it may be included by a set of meta-paths. However, the semantic-specific node embedding captured by the node-level attention module can merely reflect an individual meta-path information. To learn a more comprehensive node embedding for users and ROIs, we propose the semantic-level attention to automatically learn the importance of different meta-paths and fuse them for the semantic embedding. Given $P$ groups of semantic-specific node embeddings learned from the node-level attention module, the learned latent semantic vector of the node can be aggregated as follows:

$$Z = \delta\left(W.Aggre_{Semantic}(\{Z^{\Phi_p}, p \in [1, ..., P]\})\right), \tag{5}$$

where $Aggre_{Semantic}$ is the aggregation function to fuse the semantic-specific embeddings of ROIs.

A natural formulation of $Aggre_{Semantic}$ is also the mean operator, which estimate the element-wise mean of the vectors in $\{Z^{\Phi_p}, p \in [1, ..., P]\}$ as follows:

$$Z = \delta\left(W.\left\{\sum_{p=1}^{P} \beta Z^{\Phi_p}\right\} + b\right), \tag{6}$$

where $\beta$ is fixed to $\frac{1}{P}$ for all POIs in $r_s$, considering all meta-paths with equal contribution on the ROI semantic representation construction. As mentioned before, different meta-paths may extract distinctive semantic information. Thus, we perform an attention mechanism with a two-layer neural network to calculate the semantic-level attention score $\beta$ for the final embedding $Z$, which is defined as

$$\beta_{\Phi_p} = \frac{\exp(\beta_{\Phi_p}^{*})}{\sum_{p=1}^{P} \beta_{\Phi_p}^{*}}, \quad \beta_{\Phi_p}^{*} = \frac{1}{|\nu|} \sum_{i\in\nu} q^{T}.\delta(W.z_{r_i}^{\Phi_p} + b), \tag{7}$$

where $q^T$ is the traverse of the semantic-level latent vector $q$, $\nu$ is the set of node semantic-specific embedding, $\delta(x)$ is the ReLU function, $W$ is the weight matrix, $b$ is the bias vector. Note that the higher $\exp(\beta_{\Phi_p}^{*})$, the more important meta-path $\phi_p$ is. With the learned weights, we can aggregate these semantic-specific embeddings to obtain the semantic latent embedding of the ROI.

**Prediction Layer.** To this end, we extend user preferences and ROI features in the traditional latent factor model to two components: one component is based on ratings and another one is based on heterogeneous graph information.

In particular, the latent embedding of user and ROI are mapped to a shared hidden space. Given the latent semantic embedding learned from the heterogeneous graph, the interactions between the user $u$ and the ROI $r$ is modeled as

$$s_0 = (p_u + g_u) \odot (e_r + z_r), \tag{8}$$

where $p_u$ and $e_r$ are the basic vectors of $u$ and $r$ in the latent factor model, $g_u$ and $z_r$ represent the latent semantic vectors of $u$ and $r$ obtained from the above attention mechanisms, and $\odot$ denotes the element-wise product of vectors.

The output of this layer is a $s$ dimensional vector, which is passed to the prediction layer to get a real-valued rating as

$$\hat{y}(u_i, r_j) = W_1^T s_0 + b_1, \tag{9}$$

where $W_1 \in R^s$ denotes the edge weights of the prediction layer, and $b_1$ denotes the bias. Clearly, we can add more layers of non-linear transformation between $s_0$ and prediction layer.

## 3 Experiments

### 3.1 Experimental Setup

We evaluate our model on public Foursquare check-in datasets collected from two big cities, namely, New York (NYC) and Tokyo (TKY) [5], during the period covering April 2012 to February 2013. For each observed user-ROI interaction, we treat it as a positive instance.

**Baselines.** We compare MRec with several representative recommendation methods as follows:

- **USG** [6]**:** a CF-based POI recommendation algorithm that considers both geographical influence and users' social correlations.
- **PACE** [4]**:** a deep learning strategy to jointly learn the embeddings of users and POIs for preference prediction.
- **GeoDCF** [1]**:** a neural framework-based non-negative matrix factorization for user and POI embedding.
- **NGCF** [3]**:** a state-of-the-art neural graph CF method, which can effectively inject the collaborative signal into the embedding process in an explicit manner.

For a fair comparison, when we directly use POI recommendation methods for region recommendation, we just regard the user-ROI interactions as the user-POI interactions. Furthermore, we also construct ROI context information from the original user-POI check-in data. To evaluate the effectiveness of top$-K$ recommendation and preference ranking, we adopt two widely-used evaluation protocols: recall@$K$ and ndcg@$K$.

**Table 2.** Performance comparison on two datasets evaluated by Recall and NDCG.

| Method | Recall | NDCG | Recall | NDCG |
|---|---|---|---|---|
| USG | 0.1462 | 0.1633 | 0.1031 | 0.1762 |
| PACE | 0.2187 | 0.2078 | 0.1615 | 0.2425 |
| GeoDCF | 0.2864 | 0.3108 | 0.2146 | 0.3563 |
| NGCF | 0.2541 | 0.28152 | 0.1973 | 0.3242 |
| MRec | **0.3051** | **0.3489** | **0.2372** | **0.3920** |

### 3.2 Recommendation Performance Comparison

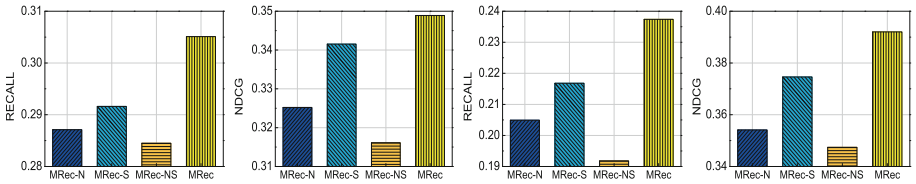
Table 2 shows the overall rating prediction error of the compared recommendation methods on NYC and TKY datasets. Our proposed method achieves the best performance on both datasets, significantly outperforming the state-of-the-art MF and Hybrid methods (on average, the relative improvement over the baseline GeoDCF is 9.62%).

### 3.3 Effects of Node-level and Semantic-level Attentions

In this subsection, we evaluate the effects of the model components. To better understand of the proposed MRec model, we further evaluate its key components, i.e., the node-level and semantic-level attention mechanisms. There are two different attention mechanisms during aggregation, including node-level attention, semantic-level attention.

We compare MRec with its three variants defined in the following: **MRec-N:** removes the node-level attention and employs the mean-based aggregation function on semantic-specific representation aggregation. **MRec-S:** removes the semantic-level attention and employs the mean-based aggregation on user or ROI latent semantic representation. **MRec-NS:** eliminates two attention mechanisms (node-level attention $\alpha$ and semantic-level attention $\beta$) and replace them with mean-based aggregation functions.

The ablation results of different attention mechanisms on MRec are shown in Fig. 2. We have the following findings: (1.) When the attention mechanism is applied at both node-level and semantic-level, the ROI recommendation performance can be enhanced, validating our assumptions that the characteristics of users or ROIs and heterogenous information are reflected at both levels. (2.) The node-level attention contributions more for MRec when compared to the semantic-level attention. This may because the node-level attention focuses on the representative meta-path based neighbors among all user's interactions, while the semantic-level attention may only work in ROIs with rich complex heterogenous information.




(a) NYC (b) TKY

**Fig. 2.** Recommendation performance of MRec when gradually incorporating meta-paths.

## 4 Conclusions

In this paper, we have proposed a novel meta-path based neural graph attention network for ROI recommendation, dubbed MRec, to address the issue of the rich heterogeneous graph information disregarded by most ROI recommendation methods. MRec firstly leverages the heterogeneous graph neural network with both node-level and semantic-level attentions to learn ROI's latent semantic representation that is enhanced via meta-path embedding. Furthermore, MRec utilizes an embedding layer to capture and extend the rich contextual information to incorporate heterogeneous information for ROI recommendation in a unified manner. An extensive experiment on two real-world dataset validates the superior recommendation performance of MRec.

**Acknowledgements.** This work was supported in part by the National Natural Science Foundation of China under Grant No. 62106091, the Doctoral Foundation of Tianjin Normal University under Grant No. 52XB2104, and the Shandong Provincial Natural Science Foundation under Grant Nos. ZR2021MF054 and ZR2019MF062.

## References

1. Rafailidis, D., Crestani, F.: GeoDCF: deep collaborative filtering with multifaceted contextual information in location-based social networks. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11052, pp. 709–724. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10928-8_42
2. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks (2018)
3. Wang, X., He, X., Wang, M., Feng, F., Chua, T.S.: Neural graph collaborative filtering. arXiv preprint arXiv:1905.08108 (2019)
4. Yang, C., Bai, L., Zhang, C., Yuan, Q., Han, J.: Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1245–1254. ACM (2017)

5. Yang, D., Zhang, D., Zheng, V.W., Yu, Z.: Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. IEEE Trans. Syst. Man Cybern. **45**(1), 129–142 (2015)
6. Ye, M., Yin, P., Lee, W.C., Lee, D.L.: Exploiting geographical influence for collaborative point-of-interest recommendation. In: International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR), pp. 325–334 (2011)

# Emotion Recognition of EEG Signals Based on Channel Attention Convolution Neural Network

Xiu Zhang, Xun Pei, and Xin Zhang(✉)

Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
ecemark@tjnu.edu.cn

**Abstract.** Electroencephalography (EEG) is more likely to respond to emotional changes compared to facial expressions and voice because it is not subject to external interference and is not easily disguised. By classifying the emotions of EEG signals, it can provide an aid for future treatment of depression, epilepsy and other diseases. Therefore, this article uses the publicly available EEG emotion dataset and uses a ratio of 80%–20% to divide the training set and test set. The features of EEG data are extracted using Fast fourier transform (FFT), and the Convolutional neural network (CNN) in deep learning is used as the basic premise to incorporate the channel attention mechanism. An EEG emotion recognition model combining CNN and channel attention mechanism is designed, which includes three convolutional layers, three channel attention blocks, three maximum pooling layers, three fully connected layers, a dense layer and a softmax layer, and batch normalization is used to suppress the overfitting of the model. Experimental results show that the sentiment recognition accuracy of the DEAP dataset reaches 90%, which achieves a significant improvement over existing emotion recognition models.

**Keywords:** Convolutional neural networks · Deep learning · EEG · Emotion recognition

## 1 Introduction

Nowadays, as the research of artificial intelligence continues to advance, more and more researchers are focusing on human emotional intelligence. As an advanced stage in the field of artificial intelligence, emotion analysis is an important research object in the fields of human-computer interaction and anthropomorphic control theory [1]. It has been found that positive emotions reflect pleasant psychological states and contribute to human health and attitudes [2]. With the growth of information through social channels, emotion computing emerged under the need for deeper understanding and rational use of emotions [3, 4]. Emotion recognition research has not only improved human-computer interaction systems, but is also an important topic in clinical research, such as the treatment of psychological disorders and epilepsy [5, 6]. Therefore, emotion recognition based on


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 90–97, 2023.
https://doi.org/10.1007/978-981-99-2362-5_12

EEG signals has received wide attention from an increasing number of researchers in the field of affective computing.

Deep learning has achieved remarkable results in the fields of computer vision and natural language processing, and researchers have started to apply deep learning to the field of EEG emotion recognition. The literature [7] proposed a convolutional recurrent neural network combining CNN and Recurrent neural network (RNN), and its binary recognition rate in DEAP dataset [8] was 72.06% and 74.12% in validity and arousal dimension, respectively. The literature [9] compared the performance of deep learning models, i.e., Long Short-Term Memory (LSTM) and CNN on training and test set splits of 80–20 and 75–25. The EEG emotion recognition accuracy of LSTM and CNN was 88.6% and 87.72%, respectively. The literature [10] proposed the application of Hierarchical convolutional neural networks (HCNN) to classify emotions as positive, neutral and negative. Their method preserves the spatial topology of the electrodes. The EEG signal is converted into an image by interpolation of the electrode positions obtained from the 2D projection technique and the frequency bands of the EEG signal. Their results have an accuracy of 86.2% in the $\beta$-band and 88.2% in the $\gamma$-band. In the literature [11] CNN are built based on convolutional operations to extract information features by fusing spatial and channel information within the local sensory field. They focused on channel relations and proposed a new architectural unit, which we call "Squeeze-and-Excitation" (SE) block, to improve the performance of the CNN model significantly.

Some approaches also use brain activity for subject identification or authentication. For example, a cascaded deep learning using a combination of CNN and RNN was proposed in the literature [12] and evaluated on the sentiment dataset DEAP. The results show that CNN-Gate recurrent unit (GRU) and CNN-LSTM can perform person recognition from different sentiment states with an average correct recognition rate of 99.90–100%. The literature [13] proposed imagined speech for user recognition, achieving 99.76% accuracy in six different subjects.

The use of brain activity for sleep monitoring is an emerging field and can be very helpful in the adjunctive treatment of apnea. The literature [14] proposed a single-channel EEG sleep staging model, SleepStageNet, which extracts sleep EEG features through a multiscale CNN and then infers the type of sleep stage by capturing contextual information between adjacent stages using RNN and Conditional random fields (CRF). Their model was validated on Obstructive sleep apnea (OSA) patients with mean accuracies of (F4-M1, 0.80) and (F4-M1, 0.67). To address the problem that existing methods rarely consider local features in feature extraction and cannot distinguish the importance of critical and non-critical local features.

## 2 DEAP Dataset

The DEAP dataset [8] is multichannel data collected experimentally by Koelstra et al. to study human emotional states. The dataset is based on physiological signals generated by music video material evoked stimuli, recorded by 32 subjects watching 40 min of music videos (1 min per music video). The original 512 Hz EEG signal was downsampled to remove artifacts and pre-processed to 128 Hz. Of the 40 physiological signal channels collected, the first 32 channels were collected as EEG signals. EEG channels were

selected according to the international system of 32 channels, the locations of which are shown in Fig. 1.

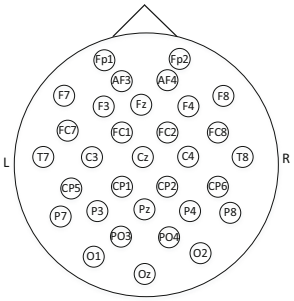



**Fig. 1.** 32 electrode channel positions

## 3 Method

### 3.1 Feature Extraction

FFT is performed on the feature extraction to reduce the dimensionality, thus having a faster training speed as well as better accuracy. These extracted features include five frequency bands. Delta-δ (1–4 Hz), theta-θ (4–8 Hz), alpha-α (8–14 Hz), beta-β (14–31 Hz), and gamma-γ (31–50 Hz). The FFT used in this paper extracts by time, thus changing a signal from the time domain to the frequency domain.

The DFT is given by:

$$X(k) = DFT[x(n)] = \sum_{n=0}^{N-1} x(n) W_N^{kn}, k = 0, 1, \ldots, N-1 \tag{1}$$

in which,

$$W_N = e^{-j\frac{2\pi}{N}} \tag{2}$$

Using the periodicity and symmetry of $W_N^{kn}$ in DFT, the whole DFT calculation becomes a series of iterative operations, which can greatly improve the operation process and the number of operations, which is the FFT algorithm.

As shown in Table 1, 14 channels and 5 frequency bands are selected for our model. The window size chosen is 256, averaging the band power over 2 s. The step size is 16, which means that every 0.125 s the update.

### 3.2 Channel Attention Mechanism

Adding attentional mechanisms to the channel dimension, the traditional channel selection applied to EEG emotion recognition usually manually picks the EEG channels

**Table 1.** Parameter settings of FFT

| Parameters | Values |
|---|---|
| channel | 1, 2, 3, 4, 6, 11, 13, 17, 19, 20, 21, 25, 29, 31 |
| Bands | 4, 8, 12, 16, 25, 45 |
| Window size | 256 |
| Step size | 16 |

associated with emotional information. However, this approach loses some important EEG information and requires the role of more manual factors. In this paper, we combine CNN with channel attention mechanism, which can automatically assign weights to the feature channels obtained by CNN extraction. That is, another new neural network is used to get to the importance of each channel, and then it is used to assign a weight value to each feature so that the neural network focuses on certain feature channels. Boosting feature channels that are useful for the current task and suppressing feature channels that are less useful for the current task.

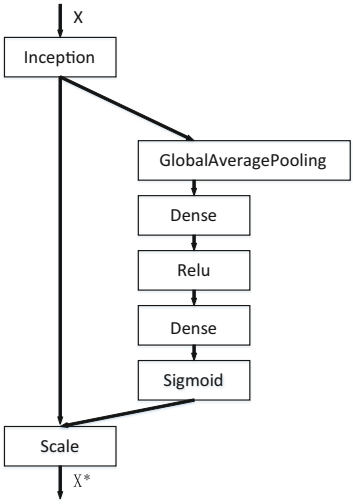



**Fig. 2.** Block diagram of channel attention mechanism

As shown in Fig. 2, the right-hand operation is to generate a weight value for each feature channel, where correlations between channels are constructed through two fully connected layers, with the same number of output weight values as the number of input channels.

Scale operation: The normalized weights obtained earlier are weighted to the features of each channel, and the Scale operation uses a multiplication method, multiplying the weight coefficients channel by channel.

### 3.3 Model Architecture

As shown in the model architecture in Fig. 3 below, there are three conv1D layers, three channel attention blocks, three fully connected layers and one dense layer, and finally softmax activation with 10 classes.

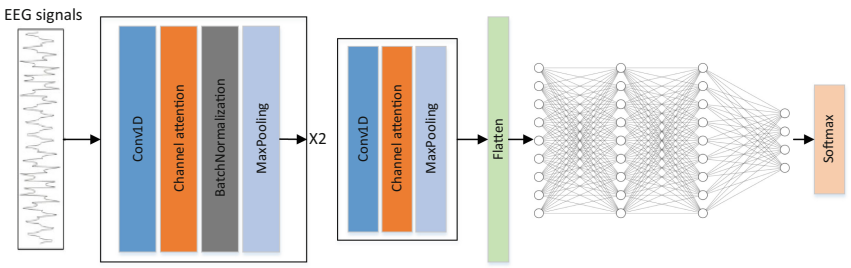



**Fig. 3.** Block diagram of EEG emotion recognition based on one-dimensional convolutional neural network and channel attention mechanism

The first convolutional layer uses Rectified linear unit (ReLU) as the activation function and 128 filters with a kernel size of 3. The input to the first layer passed to conv1D is the shape (70, 1) and uses the same padding in steps of 1. Afterwards, it is passed to the channel attention layer, which generates weights for each channel of the input and then weights the features of each channel.

Normalize the output of the first layer, i.e., use a batch normalization layer to make its mean 0 and variance 1. The next layer is maxpooling with a pool size of 2 and a window size of 2. Take the maximum value to downsample the input.

The next convolutional layer is the same as the first layer, followed by a batch normalization and maxpooling layer. The shape is then spread to form a one-dimensional layer and fed to a fully connected layer consisting of 64 neurons and hyperbolic tangent (tanh) as the activation function. A dropout on the output of the dense layer is used to reduce overfitting of the network with a dropout rate of 0.2. This is followed by a dense layer consisting of 32 neurons with tanh activation and a dropout rate of 0.2, and another consisting of 16 neurons with ReLU as the activation function and a dropout rate of 0.2.

Finally, a dense layer of 10 neurons (with an activation function of softmax) is used to give the output of the network. The softmax formula is shown in (3).

$$Softmax(x) = \frac{e^{x_i}}{\sum_i e^{x_i}} \tag{3}$$

The softmax function is different from the normal max function: while the max function only outputs the maximum value, the softmax function ensures that smaller values have a smaller probability and are not discarded outright. This means that the various probabilities it obtains are related to each other. Therefore it is more suitable for the multiclassification problem in this paper.

## 4 Overall Performance

The accuracy, loss and confusion matrices of the model are shown in Fig. 4, 5 and 6, respectively.

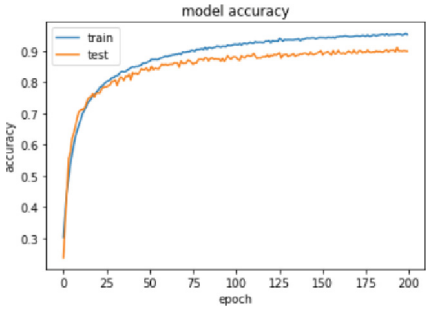



**Fig. 4.** Performance of the model

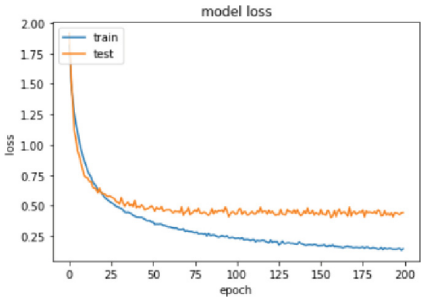



**Fig. 5.** Loss of model

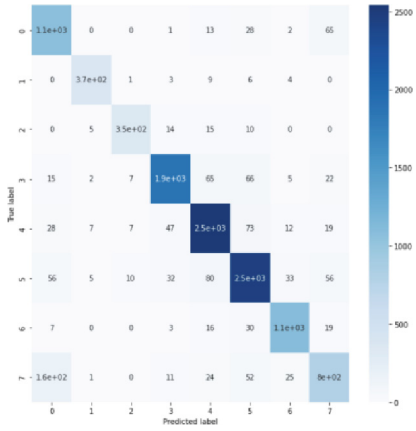



**Fig. 6.** Confusion matrix for Model

Combining the above model validation results, the one-dimensional channel attention CNN model was compared and analyzed with other models, and the results are shown in Table 2. it can be seen that the accuracy of our model is better than most EEG emotion recognition models.

**Table 2.** Comparison with the existing studies

| Method | Accuracy |
|---|---|
| C-RNN [15] | 74.12% |
| CNN [16] | 81.41% |
| RNN [17] | 84.00% |
| DBN [18] | 88.33% |
| **Ours** | **90.00%** |

## 5 Conclusion

In this paper, an EEG emotion recognition method based on channel attention convolution is proposed. Experimental results on the DEAP dataset show that the CNN incorporating the channel attention mechanism has a significant performance improvement over the traditional CNN model, with a test accuracy of 90%. The improved model proposed in this paper outperforms most current schemes. In future work, we intend to extract more EEG signal features and fuse them to further improve the performance of EEG emotion recognition through feature complementation.

## References

1. Zhu, Y.H., Wei, J.H., Mao, J.X.: A review of artificial emotion research. J. Jiangnan Univ. (Nat. Sci. Ed.) **11**(04), 497–504 (2012)
2. Pressman, S.D., Cohen, S.: Does positive affect influence health? Psychol Bull. **131**(6), 925–971 (2005)
3. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
4. Calvo, R.A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)
5. Ozdemir, M.A., Degirmenci, M., Guren, O., Akan, A.: EEG based emotional state estimation using 2-D deep learning technique. In: 2019 Medical Technologies Congress (TIPTEKNO), pp. 1–4. IEEE (2019). https://doi.org/10.1109/tiptekno.2019.8895158
6. Othman, M., Wahab, A., Karim, I., Dzulkifli, M.A., Alshaikli, I.F.T.: EEG emotion recognition based on the dimensional models of emotions. Procedia Soc Behav Sci **97**, 30–37 (2013)
7. Li, X., Song, D., Zhang, P., Yu, G., Hou, Y., Hu, B.: Emotion recognition from multi-channel EEG data through convolutional recurrent neural network. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 352–359 (2016). https://doi.org/10.1109/BIBM.2016.7822545
8. Koelstra, S., et al.: DEAP: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2012). https://doi.org/10.1109/T-AFFC.2011.15
9. Acharya, D., et al.: Multi-class emotion classification using EEG signals. In: Garg, D., Wong, K., Sarangapani, J., Gupta, S.K. (eds.) IACC 2020. CCIS, vol. 1367, pp. 474–491. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-0401-0_38
10. Kuanar, S., Athitsos, V., Pradhan, N., Mishra, A., Rao, K.R.: Cognitive analysis of working memory load from EEG, by a deeprecurrent neural network. In: 2018 IEEE International Conferenceon Acoustics, Speech and Signal Processing (ICASSP), pp. 2576–2580. IEEE (2018)

11. Jie, H., Shen, L., Albanie, S., Sun, G., Enhua, W.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**(8), 2011–2023 (2020). https://doi.org/10.1109/TPAMI.2019.2913372
12. Wilaiprasitporn, T., Ditthapron, A., Matchaparn, K., Tongbuasirilai, T., Banluesombatkul, N., Chuangsuwanich, E.: Affective EEG-based person identification using the deep learning approach. IEEE Trans. Cogn. Dev. Syst. **12**(3), 486–496 (2020). https://doi.org/10.1109/TCDS.2019.2924648
13. Brigham, K., Kumar, B.V. K. V.: Subject identification from electroencephalogram (EEG) signals during imagined speech. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8 (2010). https://doi.org/10.1109/BTAS.2010.5634515
14. Chen, K., Zhang, C., Ma, J., Wang, G., Zhang, J.: Sleep staging from single-channel EEG with multi-scale feature and contextual information. Sleep Breath. **23**(4), 1159–1167 (2019). https://doi.org/10.1007/s11325-019-01789-4
15. Li, X., Song, D., Zhang, P., Yu, G., Hou, Y., Hu, B.: Emotion recognition from multi-channel EEG data through convolutional recurrent neural network. In: IEEE International Conference on Bioinformatics and Biomedical (BIBM), Guang-dong, China, pp. 352–359 (2016)
16. Tripathi, S., Acharya, S., Sharma, R.D., Mittal, S., Bhattacharya, S.: Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. In: AAAI Conference on Innovative Applications, San Francisco, California, USA, pp. 4746–4752 (2017)
17. Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Souly, N., Shah, M.: Deep learning human mind for automated visual classification. In: Proceedings - IEEE Conference on Computer Vision Pattern Recognitin, CVPR, pp. 4503–4511 (2017)
18. Xu, H., Plataniotis, K.N.: Affective states classification using EEG and semi-supervised deep learning approaches. In: IEEE Workshop Multimedia and Signal Processing, Montreal, Canada, USA, pp. 1–6 (2016)

# The Study of Multi-channel SAR Sub-band Receiver and Pulse Compression Technology

Feng Xu(✉), Wen Pan, Defeng Shen, Fan Xu, Jun Lu, Bin Xu, Haixia Li, Xinguang Zhang, Wen Jin, and Min Zhu

Beijing Institute of Space Long March Vehicle, No.1, Nan Da Hong Men Road, Beijing 100076, China
qitianzi@163.com

**Abstract.** The demand of high-resolution SAR system is the wide bandwidth and the wide swath. The direct sampling is impossible for so high speed data streaming. The common solution is parallel receiving echo data by multi-channel. According to the character of synthetic aperture radar, this paper proposes the technology of the multi-channel receiver and sub-band pulse compression. Finally, the computer simulation confirms the validity of the method successfully.

**Keywords:** SAR · Multi-channel sub-band receiver · Sub-band pulse compression

## 1 Introduction

Synthetic aperture radar is the wide bandwidth and high-resolution radar. With the increasing demand of the image resolution, the signal bandwidth is becoming wider and wider. For example, the resolution of 0.1 m needs the radar signal bandwidth of 1.8 GHz [1]. Then generating these problems: first is that there is no the AD sample chip that can sample directly for those wide bandwidth. Secondly, this high-speed data streaming must use multi-channel parallel processing. Thirdly, the implementation of large point FFT is difficult in the real-time processing system. So sub-band receiver and sub-band pulse compression are the ideal method now. This paper mainly studies the implementation method of sub-band receiver and sub-band pulse compression.

## 2 Theoretical Foundations

### 2.1 Sub-band Receiver Technology

Sub-band receiver technology mainly has two types: one is the time-domain sub-band receiver and the other is frequency-domain sub-band receiver [2]. The time-domain sub-band receiver must do the signal synthetic then do pulse compression because it divides into groups in time-domain and can not process the data in parallel. So the frequency-domain sub-band receiver is accepted in the design commonly. The frequency-domain sub-band receiver divides the data to some sub-bands according to the frequency


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 98–105, 2023.
https://doi.org/10.1007/978-981-99-2362-5_13

bandwidth of the signal firstly, then samples the narrow band-width signal in the low data speed [3]. Finally it composes the sub-band signals to a wide-band signal. This is the frequency divider and frequency synthetic of LFM chirp echoes.

The whole wide band is divided into many narrowband signals with filter group. Before band divided, the wide-band signal need a frequency shift. For simplifying the filter design, sub-band sampling method is illustrated in Fig. 1. Shifting the frequency before sampling and then dividing the band by low-pass filter H(w).

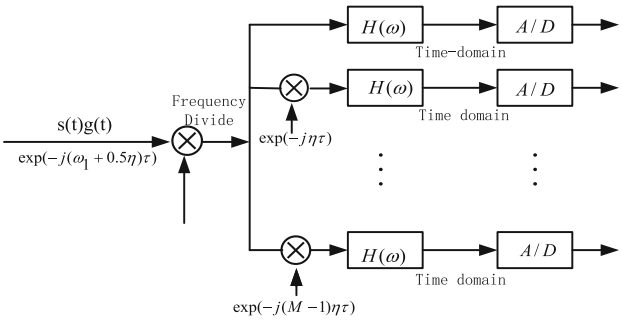



**Fig. 1.** Sub-band receiver in frequency domain

According to the Fig. 1, the process step is: firstly exploit the techniques of intersection of spectrum, secondly filter the signal with the same low-pass filter for decreasing the error of signal synthesized. The error of sub-band sample mainly produces by the multi-channel amplitude and phase error and the distortion of filter response function. The distinct difference between the sub-band sample and the time-shift sample is the front one can do multi-channel synthesis after accomplished the pulse compression and the latter one must do multi-channel synthesis before pulse compression. The special benefit of sub-band sample is that it can accomplish the signal parallel processing by combining the sampling with the pulse compression.

## 2.2 Sub-band Pulse Compression Technology

Sub-band pulse compression is the process of pulse compression after sub-band sampling. The different method of sub-band pulse compression is the same on the mathematics. If the imaging algorithm has azimuth processing before sub-band synthesis, the various methods will affect the final image results.

For decreasing the max frequency of the sub-band, the low-pass filter is used in the intersection of spectrum. The benefit of it is decreasing the Nyquist sampling rate without distortion. The sub-band signal with intersection of spectrum can fulfill sub-band pulse compression by matched filter. The multi-channel synthesis is fulfilled after the interpolation of pulse compressed signal [4, 5]. On the theory, the interpolation of band-limited signal uses the Sinc function, but it can not be fulfilled by FIR filter. Another method that the cost of hardware is huge is the time-interleaved sampling. The two methods often are combined to use in practice.

## 3 Implementation of Pulse Compression

The inner noise and outer random noise of the actual receiver is similar with Gaussian white noise because the width of power spectrum is larger than the width of system transmission band. The output of maximum SNR is emphasized in the distance equation of radar. So the method is needed can make the noise power of radar receiver output minimum. The matched filter is the best linear filter can make the SNR of the output maximum at some time. The two methods often be combined to use in practice.

The transfer function is thus:

$$\mathrm{H(w)} = \mathrm{cS}^*(w)e^{-jwt_0} \tag{1.1}$$

This filter is named by matched filter because it is the conjugate type of the transmitting signal. The transfer function of the filter H(w) is direct ratio to the conjugate of the signal spectrum $S^*(w)$ and $c$ is the constant ratio parameter. The transfer function of the filter $|\mathrm{H}(w)|$ is the operator which is the amplitude of spectrum for the input procedure. It makes the signal pass in the frequency unit [w, w + dw].

The phase-frequency characteristics of |H(w)| are thus:

$$\arg H(w) = -\arg S(w) - wt_0 \tag{1.2}$$

where the first term of the right part is inverse with the phase-frequency characteristics and the second term is linear ratio with the signal frequency. The $\mathrm{argH}(\omega)$ compensate the phase of input signal and make the phase of the output signal same direction at the moment of $t = t_o$. The output signal's amplitude is maximum at the moment $t = t_o$. The phase of noise is random and can not be affected by transfer function. So the SNR of the output signal is maximum at the moment $t = t_o$.

A compression network matched with transmitting signal spectrum is installed in the receiver. It makes the wide pulse to the narrow pulse for obtaining good range resolution.

The essence of the matched filter is the procedure that makes the time-domain convolution mapping into the frequency-domain multiplication by FFT arithmetic. The technological process of pulse compression has four steps. The first step is calculating the FFT results of radar echoes. The second step is generating the reference function of the transmitting signal. The third step is multiplying the FFT results with the reference function. The last step is IFFT of the data multiplied.

Matched filter is used in pulse compression, defined by G(W),the result of frequency-domain after compression is thus:

$$R(w) = F(w) \cdot G(w) \tag{1.3}$$

After sampling, this is given by:

$$r(t) = IDFT(R(w)) \tag{1.4}$$

According to the character of IFFT(Inverse Fast Fourier Transform), there are three various type equations. Three various process flowing that is the same results on theory is established.

The first type is thus:

$$r(t) = IDFT(G(w) \cdot \sum_{i=1}^{N} F_i(w) * \delta(w + w_i)) \tag{1.5}$$

This method obtains the high-speed sampling results through synthesizing the various channel signals before sampling. Then the pulse compression is executed. The signal synthesized can be done in the time-domain and also can be done in the frequency-domain.

The second type is thus:

$$r(t) = IDFT(\sum_{i=1}^{N} (G(w - w_i)F_i(w)) * \delta(w + w_i)) \tag{1.6}$$

This method first runs the multiplication between sub-band signal and the relevant reference function, then obtain a large bandwidth signal by combine the sub-band spectrum in the frequency-domain. Finally, the signal combined maps into time-domain by IFFT operation.

The third type is thus:

$$r(t) = \sum_{i=1}^{N} IDFT(G(w - \omega_i)F_i(w)) \cdot \exp(-jw_i t) \tag{1.7}$$

This method makes sub-band signals mapping into time-domain after multiplying matched and then does spectrum move and spectral overlay in time-domain. As for existing the problem of frequency mixing after spectrum move, the synthesis of the results of compression in time-domain need a process of over-sampling. Zero-padding in frequency-domain can be used as for over-sampling before IFFT operating.

The distinct difference above the three methods is the accuracy and calculating amount. Only at the aspect of calculating amount, the difference is little. The first method is useful for all kinds of image algorithm. The other methods are only useful for the algorithm that makes range compression separate. Otherwise, the error will be taken into the imaging process.

Assuming the NO. i sub-band signal is $F_i(w)$ ,the matched compression function is $G(w - w_i)$. As for the spectrum of sub-band signal is narrow than that of the whole matched function, the actual matched compression function can be written as:

$$G_i(w) = G(w - w_i)Ga(w) \tag{1.8}$$

where $G_a(w)$ is a ideal window-function, also can be regarded as a ideal low-pass filter. The function of it is extracting the overlap part with $F_i(w)$.So the bandwidth of $G_a(w)$ must contain all the frequency information of $F_i(w)$.If $H(w)$ is overlapped then $G_i(w)$ is the same status as for the cutoff frequency of $G_i(w)$ is the same as that of $H(w)$.

The third method is shown in Fig. 2 where the process of IDFT is different from FFT. The implementation procedure of IDFT is IFFT the data after zero-filled and the number of FFT become large.

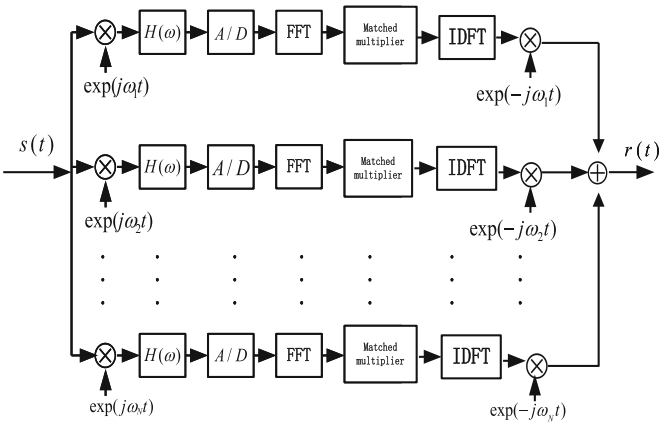



**Fig. 2.** The flow of sub-band pulse compression

The sampling interval of sub-band pulse compression is integer factor by the synthetic pulse compression during the process of multi-channel synthesized. So before multi-channel synthesized, the interpolation with the results of sub-band pulse compression must implemented for over-sampling. The method of interpolation has polynomial interpolation and Sinc function interpolation.

There is the overlap between adjacent sub-bands through segmentation based on sub-bands as for the actual filter is not ideal. This is the factor that affects the performance of synthesized pulse compression. The solution of overcome it is improving the sample speed of AD appropriately. The other factor affects the performance of synthesized pulse compression is the range side-lobe and grating lobe which generated by interpolation. In the condition of constant filter length, the main lobe is not affected by the number of sub-band. When the number of the sub-band is bigger, the first side-lobe of synthesized pulse compression is little and the other side-lobe is big more.

## 4 Simulation and Analysis

In order to test and validate the performance of synthesized pulse compression obtained by multi-channel, simulation experiments can be designed in MATLAB.

In simulation, the wideband signal is divided into three sub-band signals. The signal is standard linear frequency modulation, such as:

$$\mathrm{s(t)} = \mathrm{rect}(\frac{\mathrm{t}}{\tau})\mathrm{exp}(\mathrm{j}2\,\pi\,\mathrm{f_o t} + \mathrm{j}\pi\,\mathrm{kt}^2) \tag{1.9}$$

where $\tau = 2\,\mathrm{us}, B = 300\,\mathrm{MHz}, f_o = 0$,the direct sampling rate is 600 MHz, the sub-band width is 100 MHz and the sampling rate of sub-band is 120 MHz. The interpolation of Sinc function is an ideal filter and the cut-off frequency is 50 MHz.

Figure 3 shows the original linear frequency modulation(LFM) signal in frequency domain. This is the standard spectrum graphic of LFM.

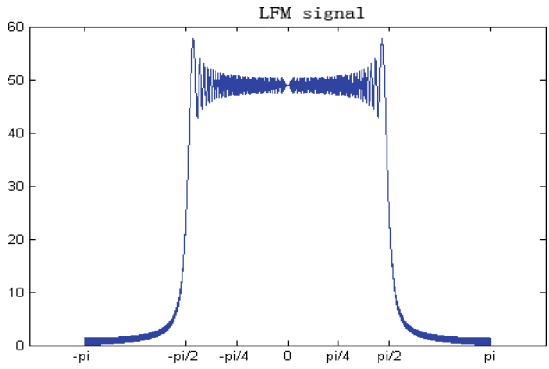



**Fig. 3.** LFM signal

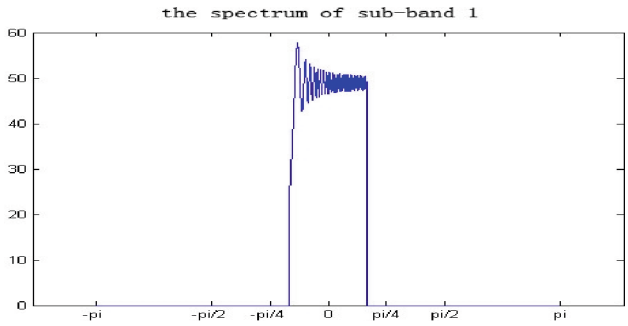



**Fig. 4.** Sub-band signal

Figure 4 shows the sub-band signal of LFM in frequency domain. This graphic shows that the sub-band signal's bandwidth is one part of full spectrum.

Figure 5 shows the result of sub-band signal pulse compression. It means that the sub-band signal is also a coherent signal and can pulse compress.

The graphic of partial enlargement of synthesized sub-band pulse compression is illustrated in Fig. 6. The main lobe of synthesized pulse compression is a little wider than the direct pulse compression. At the same time, the side lobe level of synthesized sub-band pulse compression is higher than the direct pulse compression. Fortunately, the performance of synthesized sub-band pulse compression is approximate to the direct pulse compression. It indicates that multi-channel sub-band pulse compression is effective.

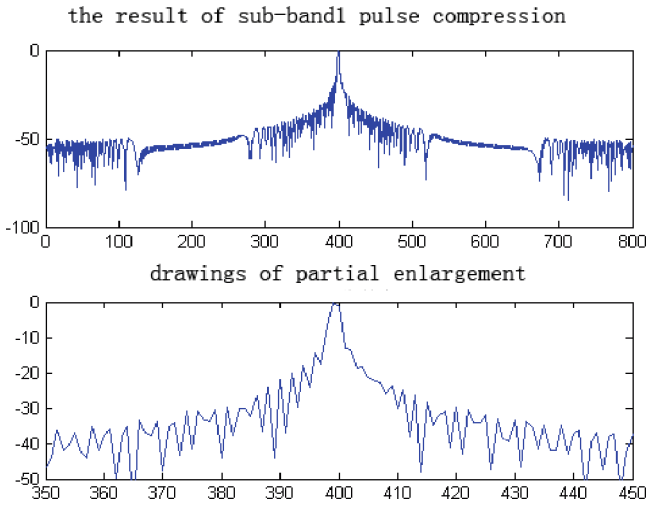



**Fig. 5.** Sub-band pulse compression

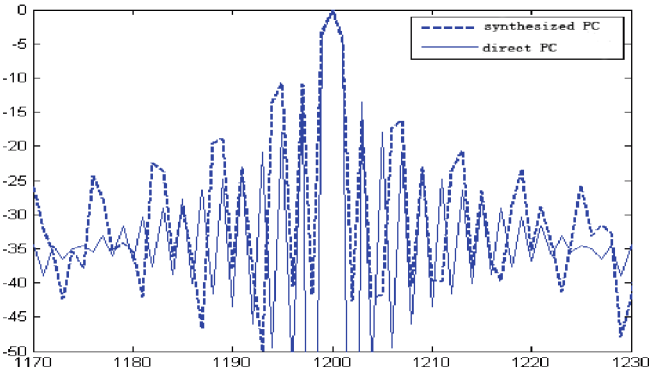



**Fig. 6.** Comparison of pulse compression

## 5 Conclusion

The sub-band receiver and pulse compression technology that this paper proposes can improve the ability of processing high wide-width signal of high-resolution SAR radar. It can reduce the AD sample rate. Simulation experiments confirm the validity of the method of the sub-band pulse compression. It is significant for the real-time process on engineering of the wide-swath and high-resolution SAR.

## References

1. Wang, Y., Chang, L., Li, H., Jia, Y.: An airborne SAR with 0.1 m resolution using multi-channel synthetic bandwidth. J. Electron. Inf. Technol. **35**(1), 29–35 (2013)

2. Cumming, I.G., Wong, F.H.: Digital Processing of Synthetic Aperture Radar Data, p. 140. Artech House, Boston (2005)
3. Ender, J.H.G.: Signal processing for multi-channel SAR applied to the experimental SAR system AER. In: Proceedings of International Radar Conference, Paris (1994)
4. Song, X.F., Wang, F., Yu, W.D.: Over-sampling subband pulse compression system analysis and its application in SAR. J. Grad. School Chin. Acad. Sci. **25**(5), 620–625 (2008)
5. Deng, Y., Chen, Q., Qi, H., Zheng, H., Liu, Y.: A High-resolution imaging algorithm for MIMO SAR based on the sub-band synthesis in frequency domain. J. Electron. Inf. Technol. **33**(5), 1082–1087 (2011)

# Depth Completion Using Infinity Laplacian Based on Steering Positive Definite Metric Operator Plus Convolutional Stage

Vanel Lazcano(✉) and Felipe Calderero

Núcleo de Matemática, Física y Estadística, Universidad Mayor,
Avda. Manuel Montt 316, Providencia, Santiago, Chile
vanel.lazcano@umayor.cl
http://www.umayor.cl

**Abstract.** Depth completion is an important task for autonomous vehicles, video games, and many others applications. Frequently depth maps present holes or data with low confidence level. This paper is devoted to the guided depth completion problem using a scene color reference image to guide the completion of a sparse depth map. In this paper we propose a model to complete the depth map based on an interpolator model and convolution stages enforcing color features of the reference image. In one hand, the interpolator we used is the infinity Laplacian (which is the simplest interpolator that holds a set of axioms). In the other hand, The convolutional stage is constructed using either Gabor filters (GF) or steering filters (SF). We tested different configuration of our proposal to find the best performance: i) SF Stage-infinity Laplacian-GF Stage. ii) SF Stage-Steering kernel-GF Stage. iii) SF Stage-infinity Laplacian (based on Steering Positive Definite Operator)-GF Stage. These models were assessed in the publicly available KITTI Depth Completion Suite. Best performance was obtained by the model iii) outperforming our model previous version, and others contemporaneous models.

## 1 Introduction

Depth completion is an important task for autonomous vehicles, video games and many others applications. Depth sensor such as Kinect sensor or LiDar acquires low resolution images that present lack of information or data with low confidence level. A low cost solution to complete the lack of data is to interpolate the available data using an algorithm. Many strategies has been proposed to solve this problem from simple interpolation, non-local means or convolutional neural networks. This work proposed a hybrid model based on the infinity Laplacian. The infinity Laplacian is a very simple interpolator that satisfy a set of axioms [1]. We applied a convolution stage to a reference color image selecting color features (CF). The interpolator completing the depth data performs anisotropic diffusion guided by the CF image. We tested three different pipelines to improve the interpolator performance in the depth completion task.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 106–113, 2023.
https://doi.org/10.1007/978-981-99-2362-5_14

### 1.1 Related Works

In [2] a fully connected Conditional Random field (CRF) is used to complete depth maps. Authors incorporating cues from cross-modal stereo between Infrared and RGB Kinect camera sensors, provides even dense maps in transparent objects. The authors presented an evaluation of their dataset containing transparent objects, discounting the completion of the depth map in edges, their proposal outperforms other well-known models.

In [3] a model to complete the depth based on Steering Kernel is used in combination with a refinement neural network. It was evaluated in a database created from depth maps and color reference images extracted from a public dataset. They downsampled the available depth covering less than 1% of the image size. Obtained results show that the model outperforms other proposals.

## 2 Stated Model

We implemented the following pipeline to complete sparse depth maps.

### 2.1 Steering Filter

A steering filter (similar in structure to [4]) is a filter whose parameters are data dependent, and the filter "steers" according to the dominant gradient direction of the image. Let us consider $\mathcal{N}_P(\mathbf{x})$ be a neighborhood of radius $P$ around $\mathbf{x}$ [5]. The filtered data is given by,

$$z = \frac{\sum\limits_{x_i \in \mathcal{N}_P(\mathbf{x})} K_H(\mathbf{x} - \mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i)}{\sum\limits_{x_i \in \mathcal{N}_P(\mathbf{x})} K_H(\mathbf{x} - \mathbf{x}_i)} \tag{1}$$

where $K_H(\mathbf{x} - \mathbf{x}_i)$ is the kernel function given by,

$$K_H(\mathbf{x} - \mathbf{x}_i) = \frac{\sqrt{\det(\hat{C}_i)}}{2\pi h^2 \mu_i^2} \exp\left( -\frac{(\mathbf{x} - \mathbf{x}_i)^T \hat{C}_i (\mathbf{x} - \mathbf{x}_i)}{2\, h^2 \mu_i^2} \right), \tag{2}$$

where $\hat{C}_i$ is an approximation to the local covariance matrix ($\hat{C}_i \approx C_i$),

$$C_i = \sum_{x_i \in \mathcal{N}(\mathbf{x})} \left( \nabla I_r(\mathbf{x}_i) \nabla I_r^T(\mathbf{x}_i) + \nabla I_g(\mathbf{x}_i) \nabla I_g^T(\mathbf{x}_i) + \nabla I_b(\mathbf{x}_i) \nabla I_b^T(\mathbf{x}_i) \right), \tag{3}$$

where $\nabla I_k(\mathbf{x})$ with $k = r, g, b$, is the gradient of each color component of the reference image. We remark that the computed covariance matrix could be singular.

To avoid directly work with this covariance matrix, a singular values decomposition (SVD) is performed, thus $\hat{C}_i = \gamma_i U_{\theta_i} \Lambda_i U_{\theta_i}^T$, where $U_{\theta_i}$ is a rotation matrix, $\Lambda_i$ is an elongation matrix, $\gamma_i$ scale parameter and,

$$\Lambda_i = \begin{pmatrix} \sigma_i & 0 \\ 0 & \sigma_i^{-1} \end{pmatrix} \text{ and } U_{\theta_i} = \begin{pmatrix} \cos(\theta_i) & \sin(\theta_i) \\ -\sin(\theta_i) & \cos(\theta_i) \end{pmatrix}. \tag{4}$$

Given the decomposition in singular values $U_i S_i V_i^T$, the dominant orientation of the gradient data in $\mathcal{N}_P(\mathbf{x})$ is given by, $\theta_i = \arctan\left(\frac{v_1}{v_2}\right)$, where $(v_1, v_2)^T$ is the eigenvector associated to the largest singular value of $C_i$. Finally $\sigma_i$ and $\gamma_i$ are defined as [5]: $\gamma_i = \sqrt{\frac{s_1 s_2 + \lambda_2}{2(2P+1)^2}}$, and $\sigma_i = \frac{s_1 + \lambda_1}{s_2 + \lambda_1}$, $s_1$, $s_2$ are the singular values of $C_i$, $\lambda_1$, $\lambda_2 \in \mathbb{R}$ are regularization parameters that ensure that $\hat{C}_i$ be non-singular. The main idea behind this parametrization is that, in regions with low texture (flat regions of the image) the singular values of $C_i$ are $s_1 \approx s_2 \approx 0$. In the other hand, near edges we have that the singular values of $C_i$ are $s_1 >> s_2$ (an oriented region). As a proof of concept (as in [3]) we present an example of the computed parameters $\gamma_i$, $\sigma_i$, $\theta_i$. In Fig. 1 we show an image and its computed $\theta_i$, $\gamma_i$ and $\sigma_i$ for each point of the image. We present in Fig. 1 (a) an urban reference color image. (b) We present its $\gamma_i$ for each value of the image, which presents large values in flat regions of the image. We use color code hot to represent the values $\gamma_i$ values. In (c) we show the $\theta_i$, and we observe that the region orientations are well estimated. In (d) we present $\sigma_i$, in flat regions is almost 0 but in regions with high gradients, it presents larger values.

(a) (b) (c) (d)

**Fig. 1.** Parametrization of the gradient covariance matrix using and urban image. (a) Color reference image. (b) scale $\gamma_i$, (c) orientation $\theta_i$ and (d) dominant orientation $\sigma_i$.

### 2.2 Interpolation Model

The infinity Laplacian is the simplest interpolator that satisfies a set of suitable axioms [1]. Let us consider a color image $I : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$, the available depth data are located in $O \subset \Omega$. The infinity Laplacian equation is given by: $\Delta_{\infty,g} u(\mathbf{x}) = 0 \in \Omega$, where $g$ is the used metric and $u|_{\partial O} = \theta$ is the available data.

### 2.3 Considered Metric

Let's consider the domain $\Omega$ as a discrete grid. The metric includes a spatial and a photometric term. The distance between two points $\mathbf{x}$ and $\mathbf{y}$ is defined as,

$$d_{\mathbf{xy}} = (\kappa_x \delta^p(\mathbf{x}, \mathbf{y}) + \kappa_c \gamma^q(\mathbf{x}, \mathbf{y}))^r, \tag{5}$$

where $\delta(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T A(\mathbf{x} - \mathbf{y})$ and $\gamma(\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x})|I(\mathbf{x}) - I(\mathbf{y})|^2 + (1 - \alpha(\mathbf{x}))(\nabla I(\mathbf{x}) - \nabla I(\mathbf{y}))^T B(\nabla I(\mathbf{x}) - \nabla I(\mathbf{y}))$, $\alpha(\mathbf{x})$ is a balance term between intensities and gradients of the image as explained in [6], [7], $p$, $q$, and $r \in \mathbb{Q}^+$, $A$ and $B$ positive definite matrices, $\nabla I : \Omega \rightarrow \mathbb{R}^2$ is the gradient of the image $I$. We have included the Positive Definite Metric Operator $A$ and $B$ following ideas in [4].

## 3 Practical Implementation

In this section, we describe the numerical solution of the infinity Laplacian, one approximation used, and the complete pipeline of the model.

### 3.1 Numerical Solution

The solution of Infinity Laplacian is given by: $u(\mathbf{x}) = \dfrac{d_{\mathbf{xz}} u(\mathbf{y}) + d_{\mathbf{xy}} u(\mathbf{z})}{d_{\mathbf{xz}} + d_{\mathbf{xy}}}$. Its iterated version is, $u^{k+1}(\mathbf{x}) = \dfrac{d_{\mathbf{xz}} u^k(\mathbf{y}) + d_{\mathbf{xy}} u^k(\mathbf{z})}{d_{\mathbf{xz}} + d_{\mathbf{xy}}}$, with $k \in \mathbb{N} \cup \{0\}$. We observe that the numerical solution of the infinity Laplacian model is just a weighted average, which makes it easy and fast to implement [8].

### 3.2 Convolutional Pipeline

The complete pipeline of our proposal contains different stages as in [9] as we show in Fig. 2. First, the color reference image is decomposed into its color components and each color component is processed with convolutional filters. For our evaluation, we used the Gabor filter or the Steering filter, which we will explain in a further section. Second, the convolved component is max-pooled to extract selective color features. Each component is composed to a color image and used as an input for the interpolator. The depth map is filtered to eliminate outliers as in [10]. The filtered depth map is used as a second input of the interpolator. The interpolator uses the processed color reference image to guide the completion of the depth map. The interpolator can be either the infinity Laplacian or the Steering kernel, depending on the experiment. The interpolated depth is filtered with a convolutional stage to eliminate noise in the final estimation. In Fig. 2 we show our pipeline, where red, green, and light blue blocks convolve each color component with a filter.

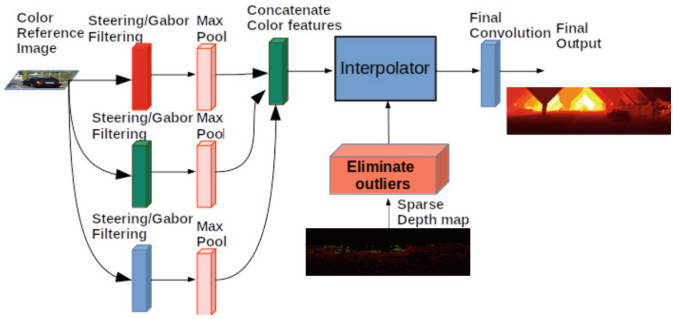



**Fig. 2.** Implemented pipeline for depth completion.

## 4 Parameter Estimation

### 4.1 Data Set and Training Set

Considering the publicly available KITTI Depth Completion Suite [11], we extracted five color reference images and their corresponding ground truth to construct a training set to estimate parameters, as we show in Fig. 3. In Fig. 3 we present five color reference images used to train our proposal, and we present a (color-coded) sparse depth map corresponding to each color reference image. Black color means no data, and depth values are color-coded between small (red) and large (yellow). The KITTI Depth Completion Suite [11] consists of 1000 urban color reference images, a sparse depth map, and its corresponding ground truth. We leave 995 images to validate the model. The depth data were acquired by a LiDar mounted on a vehicle. The dataset is publicly available [11]. We present a list of the estimated model's parameters in Table 1.

**Fig. 3.** Color reference images and their corresponding sparse depth map.

### 4.2 Training the Model

To estimate the parameters we used the PSO algorithm, considering 100 individuals, 30 iterations, and in each iteration, we minimized the depth estimation

**Table 1.** Parameters of the model.

| Number of parameters | Parameter | Description |
|---|---|---|
| 1 | radius | Size of the neighborhood $\mathcal{N}_P(\mathbf{x})$ for used metric |
| 5 | $\kappa_x$, $\kappa_c$, $p$, $q$, $r$ | Metric's parameters |
| 1 | $n_{iter}$ | Number of iterations |
| 2 | $\beta_\alpha$, $\beta_\tau$ | Parameter of $\alpha(\mathbf{x})$ map |
| 2 | $a_{11}$, $a_{12}$, $b_{11}$, $b_{12}$ | Coefficients of $A$ and $B$ matrix |
| 2 | $\tau_m$, $P$ | Prefilter threshold, Patch size |
| $2+3(N-2P)\times(M-2P)$ | $h_i$, $\mu_i$, $\gamma_i$, $\sigma_i$ | Parameters of the steering filter |

error $J$ given by:

$$J = \sum_{i=1}^{5} MAE_i + MSE_i, \tag{6}$$

where $MAE$ are the mean absolute error and $MSE$ is the mean square error. Given a set of the parameters we computed $J$ for each individual. In Fig. 4 we show the learning curve of 100 individuals and the evolution of the best individual in each iteration. As we observe in Fig. 4 (b) the final $J$ value for the best individual es 1.4693.

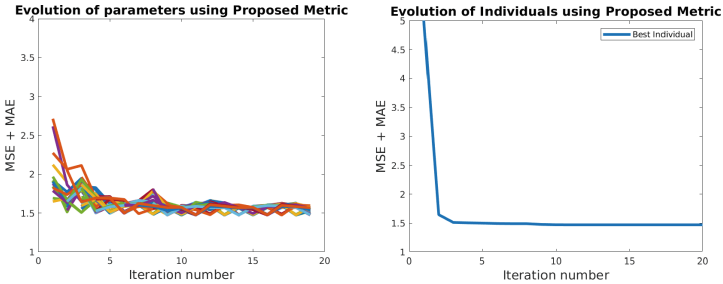



**Fig. 4.** Evolution of 100 individuals in 20 iterations. (a) Evolution of performance of 100 individuals. (b) The learning curve of the Best individual in each iteration.

### 4.3 Considered Models

We constructed three models based on the infinity Laplacian, Steering Filters, and Positive Definite Metric Operator:

i) Pipeline considering Steering Filter-AMLE-Gabor filter.
ii) Pipeline considering Steering Filter - Steering Kernel - Gabor Filter.
iii) Pipeline Steering Filter-Infinity Laplacian based on Positive Definite Metric Operator-Gabor Filter.

We show in Table 2 the final values of the performance for the best individual in three experiments. As we observe in Table 2 the model that presents worst performance is the model ii). Model i) and iii) present acceptable performance. Then, we will not consider the model ii) in further evaluations.

**Table 2.** Obtained MSE+MAE in the training stage.

| Model | MSE + MAE |
|---|---|
| SF- Infinity Laplacian - GF model i) | 1.4725 |
| SF- SK - GF model ii) | 1.9303 |
| SF- Infinity Laplacian (based on PDMO) - GF model iii) | 1.4693 |

## 5 Results

In Table 3 we present our final results evaluated in 995 images of the KITTI data set. We present the obtained results by model i) and iii). Also we compare our results with contemporaneous models. Our best results were obtained by model iii). It outperforms our previous version and model i). We included also results obtained by RigNet. That model is more complex than ours and in their implementation, the authors used 4 NVIDIA TITAN RTX GPU. In Fig. 5 we present qualitative results obtained by our model.

**Table 3.** Obtained RMSE and MAE in the validation stage.

| Moldel | RMSE | MAE |
|---|---|---|
| SF- Infinity Laplacian - GF i) | 1.1703 | 0.3770 |
| SF- Infinity Laplacian (based on PDMO) - GF iii) | 1.1267 | 0.3108 |
| PDC [10] | 1.2866 | 0.2932 |
| Lazcano [12] | 1.3420 | 0.4710 |
| RigNet [13] | 0.7520 | – |

**Fig. 5.** Left image, sparse depth map. Center, color reference image. Right, completed depth.

## 6 Conclusions

We have implemented a model that considers a new metric based on the Positive Definite Operator and Steering Filters, which parameters are computed using the local gradient of the image. The proposal outperforms our previous version of our model and other contemporaneous models. Future work will include other interpolation models, and the evaluation on other datasets.

## References


1. Caselles, V., Igual, L., Sander, O.: An axiomatic approach to scalar data interpolation on surfaces. Numer. Math. **102**(3), 383–411 (2006)
2. Saygili, G., Van Der Maaten, L., Hendriks, E.: Hybrid kinect depth map refinement for transparent objects. In: 2014 22nd International Conference on Pattern Recognition, pp. 2751–2756 (2014). https://doi.org/10.1109/ICPR.2014.474
3. Liu, L., Liao, Y., Wang, Y., Geiger, A., Liu, Y.: Learning steering Kernels for guided depth completion. IEEE Trans. Image Process. **30**, 2850–2861 (2021)
4. Minh, H., San Biagio, M., Murino, V.: Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. In: Proceedings of the 27th Advances in Neural Information Processing Systems, pp. 1–12 (2014)
5. Takeda, H., Farsiu, S., Milanfar, P.: Kernel regression for image processing and reconstruction. IEEE Trans. Image Process. **16**(2), 349–366 (2007)
6. Lazcano, V., Calderero, F.: Reconfigurable hybrid model convolutional stage - infinity Laplacian applied to depth completion. In: AICCC 2021: 2021 4th Artificial Intelligence and Cloud Computing Conference, pp. 108–114, December 2021
7. Lazcano, V., Calderero, F., Ballester, C.: Biased-infinity Laplacian applied to depth completion using a balanced anisotropic metric. In: Liang, Q., Wang, W., Liu, X., Na, Z., Zhang, B. (eds.) Communications, Signal Processing, and Systems, CSPS 2021. LNEE, vol. 878. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-0390-8_132
8. Manfredi, J., Oberman, A., Svirodov, A.: Nonlinear elliptic partial differential equations and p-harmonic functions on graphs. Differ. Integr. Eqn. **28**, 12 (2012)
9. Lazcano, V., Calderero, F.: Hybrid pipeline infinity Laplacian plus convolutional stage applied to depth completion. In: Smys, S., Tavares, J.M.R.S., Balas, V.E. (eds.) Computational Vision and Bio-Inspired Computing. AISC, vol. 1420, pp. 119–134. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-9573-5_8
10. Teutscher, D., Mangat, P., Wassermüller, O.: PDC: piecewise depth completion utilizing superpixels. In: IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, USA, pp. 2752–2758, (2021). https://doi.org/10.1109/ITSC48978.2021.9564656
11. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant CNNs. In: International Conference on 3D Vision (3DV), pp. 11–20 (2017)
12. Lazcano, V., Calderero, F., Ballester, C.: Depth image completion using anisotropic operators. In: Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020) (2020)
13. Yan, Z., et al.: RigNet: repetitive image guided network for depth completion. arxiv.org (2021)

# A Low Complexity Decoding Algorithm for Polar Codes with Flexible Path Expansion and List Size

Chenxuan Fu[1], Xiaoyong Wu[2], and Zhilong Zhang[1(✉)]

[1] Beijing Laboratory of Advanced Information Network, Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China
fcx@bupt.edu.cn, zhilong.zhang@outlook.com

[2] Shanghai Radio Equipment Research Institute, Shanghai 201109, China
wuxiaoyong@bupt.cn

**Abstract.** To reduce the complexity of polar codes, an improved multi-segment CRC-aided and partial path expansion successive-cancellation list algorithm (MP-SCL) is proposed in this paper. This algorithm determines decoding path expansion parameters based on the reliability of each polarization splitting sub-channel. At the same time, it dynamically modifies the size of decoding list according to the current CRC decoding results and the previous decoding list size. Simulation results show that when the bit error rate reaches $10^{-6}$, the complexity of our proposed algorithm can be reduced by 69.29% and 43.16% compared with cancellation list decoding of polar codes based on path metric and traditional MP-SCL, respectively.

**Keywords:** Polar codes · Low complexity decoding · Flexible path expansion · List size adjuster

## 1 Introduction

Polar code is an error correction coding scheme proposed by Arikan [1] in 2009. Unlike low-density parity check (LDPC) codes [2] and turbo codes [3], which have been proved to reach the Shannon limit [4] through experiments, polar code is the first coding scheme that can reach the Shannon capacity through theoretical proof and have been widely concerned in both industrial and academic communities.

Successive Cancellation (SC) decoding [1] is the first decoding algorithm for polar code. This algorithm has low decoding complexity, but its disadvantage is also obvious that the decoding performance is limited in finite code length. On the basis of SC, Successive Cancellation List (SCL) algorithm [5] introduces a decoding list $L$, which greatly improves the decoding performance at the cost

This work was supported by the Shanghai Science and Technology Commission Research Project under Grant 20511106700 and Shanghai Aerospace Science and Technology Innovation Fund under Grant SAST2020-0851.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 114–122, 2023.
https://doi.org/10.1007/978-981-99-2362-5_15

of increased decoding complexity. The CRC-aided SCL(CA-SCL) [6] algorithm further improves the decoding performance of polar code by introducing cyclic redundancy check, but the issue of high complexity is still not addressed. A low complexity decoding algorithm based on parity check concatenated polar codes is proposed in [7], which performs SC decoding and SCL decoding based on parity check for information bits with different reliability respectively. The Low-complexity Successive Cancellation List Decoding of Polar Codes Based on Path Metric [8], hereinafter referred to as PM-SCL, effectively reduces the decoding complexity on the premise of achieving the same performance as SCL by performing SCL decoding before the last frozen bit and then SC decoding. Multi-segment CRC-aided and Partial Path Expansion SCL (MP-SCL) [9] algorithm introduces segmented CRC check and partial path expansion based on CA-SCL, which reduces the complexity of the decoding algorithm. However, the MP-SCL algorithm has the defects of fixed selection of partial path expansion parameters and decoding list size. If these parameters can be flexibly determined, the decoding complexity is expected to be further reduced.

Motivated by the above discussion, we propose a low-complexity polar code decoding algorithm called L-$\alpha$-MPSCL, which is characterized by adaptive path expansion and decoding list size. Simulation results show that through the improvement of the above two aspects, our proposed algorithm further reduces the complexity of the MP-SCL algorithm and PM-SCL algorithm on the premise of considerable performance.

## 2 Proposed Algorithm with Flexible Path Expansion and List Size

### 2.1 Partial Path Expansion Parameter Determined Based on Channel Reliability

The MP-SCL algorithm uses a parameter $\alpha$ to represent the scale of decoding path expansion. A smaller $\alpha$ results in a high decoding complexity and degraded bit error rate performance. In the MP-SCL algorithm, $\alpha$ in each round of the decoding process is unique and non-adjustable, which leads to a lack of flexibility in reducing decoding complexity.

Based on the above analysis, we propose a decoding algorithm to determine partial path expansion parameter $\alpha$ according to channel reliability. The error probability of each split sub-channel is obtained by the Gaussian approximation (GA) method [10]. The split sub-channel is divided into message channel set $M$ and frozen channel set $F$. The former is used to transmit message bits and the latter is used to transmit frozen bits. The determined message channel will be further divided into less reliable message sub-channel $M_1$ and more reliable message sub-channel $M_2$ according to the set reliability threshold. Since the error probability of channel set $M_2$ is lower than that of $M_1$, the corresponding information bits on channel set $M_2$ will be decoded more accurately under the same condition. According to the properties of channel polarization principle, the decoding performance of channel set $M_2$ will be better than that of channel set $M_1$.

Based on the MP-SCL algorithm [9], we can select the more reliable part $M_2$ in the information bit channel set during each polarization code construction. This part applies a smaller $\alpha_1$ for MP-SCL decoding, while the remaining part $M_1$ in the information bit channel applies a larger $\alpha_2$ for MP-SCL decoding. Therefore, the decoding complexity will be reduced. Meanwhile the impact of using lower $\alpha_1$ on the decoding performance can be ignored, because $\alpha_1$ is applied to the most reliable information channel. The decoding process of MP-SCL with adaptive $\alpha$ adjustment based on channel reliability is introduced, as shown in Fig. 1 below.

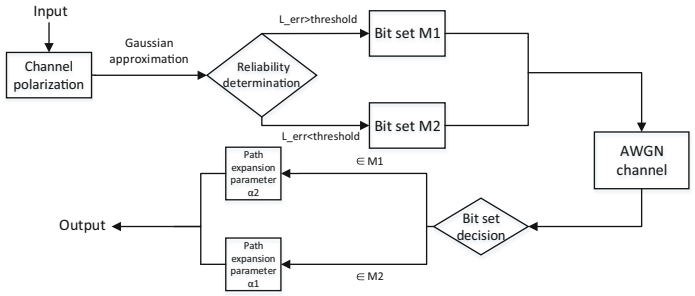



**Fig. 1.** Partial Path Expansion Parameters Determined Based On Channel Reliability

## 2.2 Decoding List Size Adjusted According to the Segment CRC Results

MP-SCL algorithm segments the whole decoding process of CA-SCL according to the number of CRC segments, but the size of the decoding path $L$ during the decoding of each segment remains unchanged, and there is no interaction for decoding information between segments. Therefore, we propose a method to dynamically adjust the size of the decoding path list based on the segment CRC results in the process of MP-SCL decoding. Specifically, let $L_{upper}$ and $L_{lower}$ respectively denote the upper and the lower limits of the decoding list size. $L_{start}$ is an initial decoding list size of the decoding algorithm, where $L_{lower} < L_{start} < L_{upper}$. After each segment decoding, a cyclic redundancy check result can be obtained.

Then, the current size of the segment decoding list $L$ is dynamically adjusted based on the CRC result and the previous $L$. The operation rules of dynamic $L$ adjustment can be summarized as follows: if at least one decoding path passes CRC check after the current segment decoding, compare the current decoding list size $L_{cur}$ and the lower limit of decoding list $L_{lower}$. If $L_{cur} > L_{lower}$, we reduce $L_{cur}$ by $T$ times, where the specific value of T can be pre-set. If $L_{cur} = L_{lower}$, $L_{cur}$ will remain unchanged in the next decoding round. If the current segment decoding obtains a surviving decoding path and no path can pass CRC successfully, then compare the current decoding list $L$ and $L_{upper}$. If $L_{cur} < L_{upper}$, we enlarge the size of $L_{cur}$ by $T$ times. If $L_{cur} = L_{upper}$, the size of $L_{cur}$ remains unchanged.

### 2.3 Overall Decoding Algorithm

Firstly, the Gaussian approximation method is adopted to obtain the error probability of each split sub-channel when constructing the polar code. We use a threshold to distinguish the reliability of the message channels. The channels with reliability values greater than the threshold are used to transmit the bits in $M_2$. Through simulation results, it is suggested to set the threshold to [0.6,0.8]. Moreover, $\alpha_2$ should be consistent with $\alpha$ of the MP-SCL algorithm, and it is verified that setting $\alpha_1$ to 0.1 is a reasonable choice. Secondly, the decoding list size adjuster determines the list size of the next segment based on the CRC result. Finally, the decoding result is obtained by integrating the decoding results of each segment. Our proposed algorithm is summarized in Algorithm 1.

**Algorithm 1.** ADAPTIVE MP-SCL ALGORITHM

**Input:** $bitErr, threshold$
**for** $1 \leq i \leq N$, **do**
  **if** $i$ is an information bit **then**
    **if** $bitErr > threshold$ **then**
      Set $i \in M_1$, by higher $\alpha_2$
    **else**
      Set $i \in M_2$, by lower $\alpha_1$
    **end if**
  **else**
    Decode $i$ to 0
  **end if**
  **Input:** $Err, L_{lower}, L_{upper}, L_{start}$ to $L$ size adjuster
  **if** i is the last bit of segment CRC **then**
    **if** $err == 0$ **then**
      **if** $L_{cur} > L_{lower}$ **then**
        reduce the size of $L$
      **else**
        $L$ remains unchanged
      **end if**
    **else**
      **if** $L_{cur} < L_{upper}$ **then**
        enlarge the size of $L$
      **else**
        $L$ remains unchanged
      **end if**
    **end if**
  **else**
    Perform conventional MP-SCL decoding
  **end if**
**end for**

## 3 Simulation Results

In order to evaluate the performance of our proposed decoding algorithm, We use 4 existing algorithms as comparison baselines, i.e., SCL, CA-SCL, PM-SCL and MP-SCL. Let $\alpha$-MPSCL represent our proposed decoding algorithm which determines partial path expansion parameters based on channel reliability, and L-$\alpha$-MPSCL represent the algorithm that adds the decoding list size adjuster on the basis of $\alpha$-MPSCL. Some simulation parameters are shown in Table 1.

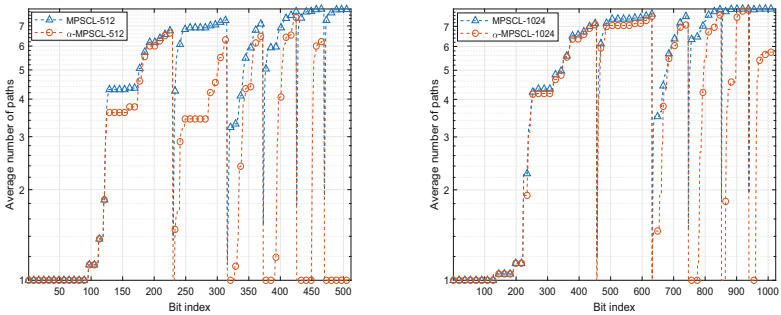
**Table 1.** Simulation Parameters

| Parameter | Value |
|---|---|
| N | 512/1024 |
| R, Frame | 0.5,10000 |
| $\alpha$ | 0.1/0.6/1.0 |
| $L_{lower}$,$L_{upper}$,$L_{start}$ | 4,16,8 |
| $T$ | 2 |
| Channel Reliability Threshold | 0.75 |

### 3.1 Decoding Complexity Comparison

The complexity of different decoding algorithms is compared in terms of the decoding paths corresponding to each bit in the decoding process.

Firstly, we simulate the decoding complexity of MP-SCL (MPSCL in simulation diagram) with partial path expansion parameter of 1.0 and $\alpha$-MP-SCL when the signal-to-noise ratio is equal to 2.5 dB. The corresponding simulation results with $\alpha_1 = 1.0$ and $\alpha_2 = 0.1$ are shown in Fig. 2 (a) and Fig. 2 (b), respectively. It can be seen that $\alpha$-MPSCL can reduce the complexity to a certain extent compared with the conventional MP-SCL, and the complexity is reduced more obviously when the code length is short. The reason is that: the MP-SCL decoding path corresponding to each bit increases slowly when the code length




(a) N=512 (b) N=1024

**Fig. 2.** Comparison of the complexity of the two decoding algorithms

**Table 2.** Proportion of the number of surviving decoding paths

| code length | 3 segs, 8 bits | 4 segs, 6 bits | 6 segs, 4 bits |
|---|---|---|---|
| 512 | 84.45% | 82.08% | 61.07% |
| 1024 | 93.80% | 86.67% | 82.15% |

is 1024 compared with 512. After adaptive $\alpha$ adjustment is introduced, once the $\alpha$ becomes smaller, the average path of each bit decreases more sharply when the code length is 512 compared with 1024.

Table 2 lists the proportion of the total decoding paths of the corresponding $\alpha$-MP-SCL algorithm to the MP-SCL algorithm under different CRC segmentation strategies when the signal-to-noise ratio is 2.5 dB. It can be concluded that the $\alpha$-MP-SCL algorithm will significantly reduce the decoding complexity with the increase of the number of CRC segments.

**Table 3.** complexity comparison

| | 1 dB | 1.5 dB | 2 dB | 2.5 dB | 3 dB |
|---|---|---|---|---|---|
| MP-SCL | 2968.739 | 3178.018 | 3095.604 | 2495.619 | 1583.714 |
| $\alpha$-MPSCL | 2621.595 | 2746.830 | 2594.242 | 2048.399 | 1247.042 |
| Reduce proportion | 11.69% | 13.57% | 16.20% | 17.92% | 21.26% |

Table 3 shows the comparison of the total decoding paths of $\alpha$-MPSCL and MP-SCL under different signal-to-noise ratios when the code length is 512 and the segmentation strategy is 4 segments, 6 bits. Under the same code length and CRC segmentation strategy, the complexity decreases significantly with the increase of SNR.

When the code length is 1024 and the signal-to-noise ratio is changed from 1 dB to 3 dB, we compare the total number of surviving paths decoded by SCL, PM-SCL, MPSCL, $\alpha$-MPSCL and L-$\alpha$-MPSCL in the simulation. The results are shown in Fig. 3. The key idea of PM-SCL is to perform SCL decoding before

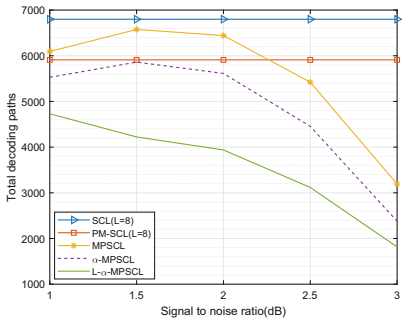



**Fig. 3.** Comparison of total survival paths of five decoding algorithms

the last frozen bit, and then perform SC decoding. Through the simulation results, it can be concluded that the decoding complexity of the proposed L-$\alpha$-MPSCL decoding algorithm is significantly lower than that of other decoding algorithms in the given signal-to-noise ratio range. Specifically, corresponding to the simulation diagram from top to bottom, when the signal-to-noise ratio is 3 dB, the complexity of L-$\alpha$-MPSCL is reduced by 73.30%, 69.29%, 43.16%, and 23.67% compared with the other four decoding algorithms.

### 3.2 Bit Error Rate Performance

After the simulation experiment proves that L-$\alpha$-MPSCL decoding algorithm can achieve significant effect in reducing complexity, the following simulation proves that L-$\alpha$-MPSCL has the same guarantee in bit error performance compared with other decoding algorithms. Figure 4 shows the simulation diagram of bit error rate of SCL, PM-SCL, MPSCL, $\alpha$-MPSCL and L-$\alpha$-MPSCL decoding algorithms under different signal-to-noise ratios.

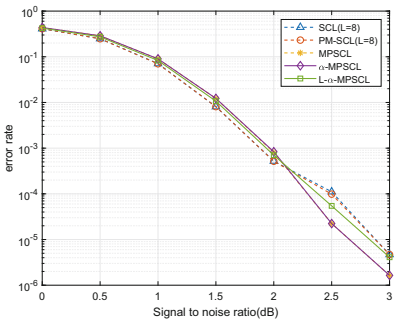



**Fig. 4.** Comparison of error rate of five decoding methods

As shown in Fig. 5, the performance of bit error rate is compared between L-$\alpha$-MPSCL and CA-SCL compare bit error rate under code length 512 or 1024. Firstly, the bit error rate of L-$\alpha$-MPSCL follows the principle that the longer the polarization code length is, the lower the bit error rate is when the signal-to-noise ratio is greater than 1 dB. Secondly, L-$\alpha$-MPSCL algorithm has better BER performance than CA-SCL algorithm under two code lengths. From the above analysis, it can be seen that the decoding complexity of L-$\alpha$-MPSCL is much lower than that of CA-SCL under the same coding condition. This shows that L-$\alpha$-MPSCL realizes the simultaneous optimization of CA-SCL algorithm in both complexity and performance. Finally, when the code length is 512, the bit error performance of L-$\alpha$-MPSCL is better than CA-SCL under some signal-to-noise ratios. This shows that L-$\alpha$-MPSCL can improve the performance degradation of polar code under short code conditions.

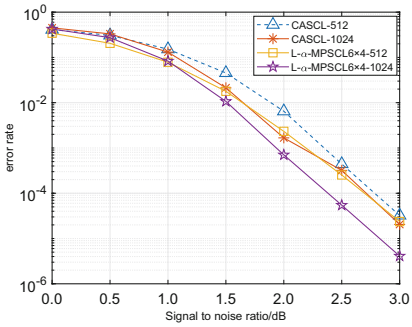



**Fig. 5.** Comparison of bit error rate between CA-SCL and L-$\alpha$-MPSCL

## 4 Conclusion

Aiming at the defect of fixed selection of partial path expansion parameters in MP-SCL, we propose a polarization code decoding algorithm to determine partial path expansion parameters based on channel reliability. The algorithm determines the partial path expansion parameters of each bit through the channel reliability obtained by Gaussian approximation method, and provides the function of adaptive $\alpha$ selection. In view of the lack of dynamics in the size of MP-SCL decoding list, the $L$ size adjuster is introduced: after each CRC test, the verification result and $L_{cur}$, $L_{upper}$ and $L_{lower}$ are introduced into the $L$ size adjuster to dynamically adjust the $L$ size. Simulation results show that the proposed decoding algorithm can further reduce the decoding complexity on the basis of MP-SCL while ensuring the bit error rate. Our proposed L-$\alpha$-MPSCL decoding algorithm provides a tradeoff between complexity and code error rate.

## References

1. Arikan, E.: Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Trans. Inf. **55**(7), 3051–3073 (2009)
2. Gallager, R.G.: Low-density parity-check codes. IRE Trans. Inf. Theory **8**(1), 21–28 (1962)
3. Berrou, C., Glavieux, A., Thitimaishima, P.: Near Shannon limit error-correcting coding and decoding: turbo-codes. 1. In: IEEE International Conference on Communications 1993, pp. 1064–1070. IEEE, Geneva (1993)
4. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
5. Tal, I., Vardy, A.: List decoding of polar of polar codes. IEEE Trans. Inf. **61**(5), 2213–2226 (2015)
6. Niu, K., Chen, K., Feng, B., Wang, Y., Wu, S., Zhang, Q.: CRC-aided decoding of polar codes. IEEE Commun. Lett. **16**(10), 1668–1671 (2012)
7. Liu, S., Wang, Y.: A low-complexity decoding algorithm based on parity-check-concatenated polar codes. J. Electron. Inf. Technol. **44**(2), 637–645 (2022)

8. Liu, X., Sun, C., Tang, W.: Low-complexity successive cancellation list decoding of polar codes based on path metric. In: 2021 IEEE 21st International Conference on Communication Technology, vol. 42(4), pp. 132–137 (2021)
9. Fu, C., Wu, X., Fei, J., Lin, J., Zhang, Z.: Decoding algorithm of polarization code based on multi-segment CRC and partial path expansion. Guidance & Fuze **42**(4), 45–51 (2021)
10. Trifonov, P.: Efficient design and decoding of Polar codes. IEEE Trans. Commun. **60**(11), 3221–3227 (2012)

# Fall Detection for Surveillance Video Based on Deep Learning

Hongwei Liu, Jiasong Mu(✉), and Zhao Zhang

Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
mjiasong@aliyun.com

**Abstract.** Accidental falls constantly threaten the lives of the elderly, and failing to detect them in time after a fall can cause the person to miss the best time to rescue, causing severe injury or even death. This paper proposes a video-based fall detection method to solve this problem. This method first performs motion detection of inter-frame difference on the video, performs feature extraction through deep learning, and finally completes classification by support vector machine to determine whether a fall occurs. The main application scene of this method is the surveillance video of the elderly living alone. The experimental results show that the proposed fall detection method has high accuracy and recall rate and can complete the fall detection task.

**Keywords:** Fall detection · Deep learning · Support vector machine

## 1 Introduction

With the continuous improvement of medical levels and living conditions in recent years, humans are living longer, and the problem of population aging has become more significant. With the increase of age, the physical function, balance ability, and coordination ability of the elderly are declining [1], which increases the probability of falls in the elderly. Falls have become a significant cause of injury and even death in the elderly. However, timely treatment after a fall can significantly reduce the mortality rate caused by a fall, and the more timely treatment is obtained, the lower the risk.

The current fall recognition methods are divided into three categories: methods based on environmental sensors, methods based on wearable devices, and strategies based on computer vision [2]. The way based on computer vision is different from the first two types. It does not require sensors to collect various parameter information. It mainly contains images or videos of human daily behavior and fall behavior and analyzes the obtained data to identify falls.

Transformer [3] has been making good progress and breakthroughs in Natural Language Processing since its introduction. The Vit [4] model completely replaces the convolution with a Transformer in the image processing task, cuts the picture into several small blocks, compares them to words in sentences, and inputs them into the model.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 123–129, 2023.
https://doi.org/10.1007/978-981-99-2362-5_16

TimeSformer [5] extends the Vit model to a video understanding model, which divides the input video data into time series of image blocks extracted from each frame and similarly obtains feature information. The video feature extraction in this paper is mainly based on the Timeformer model.

## 2 Fall Detection Algorithm

The fall detection algorithm flow is shown in Fig. 1, including motion detection, video feature extraction, and fall detection in three stages.

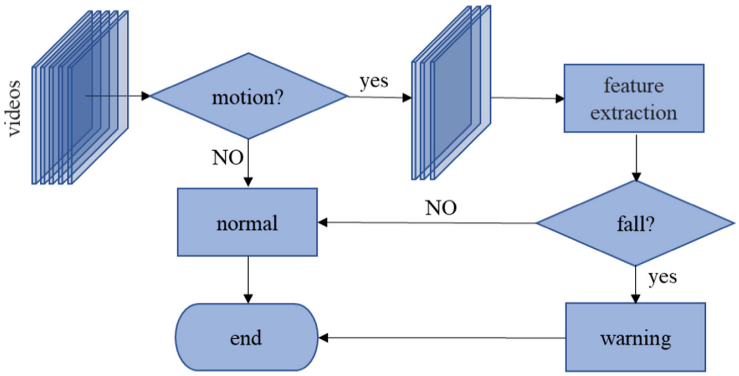



**Fig. 1.** Overall flow of fall detection algorithm

### 2.1 Motion Detection

This paper mainly uses the method of frame difference to detect motion. The difference between the two image frames can be obtained in the video by comparing the pixels. In the case of continuous frame illumination conditions, the pixel point of the difference image is not zero, which indicates that the target moves here. The primary process is as follows:

Firstly, read the video sequence frame, each frame image into a grayscale image, and differential calculation. Due to the slow movement of the elderly, the difference between adjacent frames is slight if the use of conventional two-frame difference is likely to cause misjudgment. Therefore, this paper uses the four-frame difference to take the average to reduce misjudgment. The specific method is to make the difference between the current frame and the last three frames, respectively, and take the mean value of the three different images to form the final difference image. The calculation formula is as follows:

$$D(x,y) = \frac{1}{3}(|F_t(x,y) - F_{t-1}(x,y)| + |F_t(x,y) - F_{t-2}(x,y)| + |F_t(x,y) - F_{t-3}(x,y)|) \quad (1)$$

Suppose that the image of the current time t is Ft, and the last three images are Ft − 1, Ft − 2, Ft − 3, and D representing the different images. The left of Fig. 2 is the image at time t, the middle of Fig. 2 is the image at time t − 3, and the right of Fig. 2 is the image after averaging the four-frame difference.

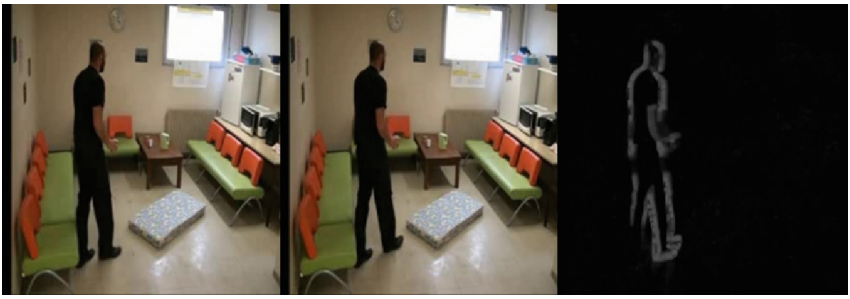


**Fig. 2.** Image at a different time and its mean difference image

Secondly, select an appropriate threshold T1, set the pixels greater than or equal to the threshold T1 as foreground pixels, set the pixels less than the threshold T1 as background pixels, and obtain the average gray value of foreground (avg1) and background pixels. The middle gray value of the pixel (avg2). d(x, y) is the value of the current differential image D at the (x, y) coordinate pixel point, and the calculation formula is as follows:

$$avg1 = \frac{1}{n} \sum_{x=1,y=1}^{n} d(x,y) \quad d(x,y) \geq T1 \tag{2}$$

$$avg2 = \frac{1}{m} \sum_{x=1,y=1}^{n} d(x,y) \quad d(x,y) < T1 \tag{3}$$

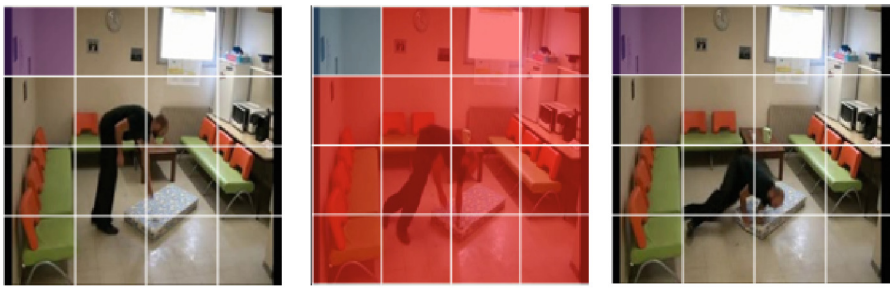
The threshold value T1 is determined by a threshold iteration method, which belongs to an adaptive determination method, and different images have different thresholds. It mainly uses multiple iterations to approximate the optimal threshold for the graph. Initialize a threshold T, calculate avg1, avg2, and Tn(Tn is the average of avg1 and avg2) under the current threshold, and use it as the initial threshold for the next round. If the difference between Tn and Tn-1 is slight, the iteration process ends, and Tn is determined as the threshold T1.

Finally, another threshold, T2, must be determined to calculate the ratio of the foreground pixel's average gray value to the background pixel's average gray value. If the percentage is greater than the threshold T2, the target moves; otherwise, the target does not move. Through the determination of the threshold, it can effectively reduce the misjudgment caused by random noise and improve the accuracy of motion detection.

### 2.2 Video Feature Extraction

Through the first stage of motion detection, the original video has been segmented into video clips containing motion information. In this stage, the video clips need input into

the TimeSformer, and the video features are extracted through the model to avoid complex and cumbersome manual feature extraction. The model first decomposes the input video to form a set of non-overlapping small image modules; to reduce the computational cost, time and space segmentation methods are adopted, and timely attention and spatial attention are used in turn [5]. Temporal attention is to compare the image modules in the same position of each frame image (the blue module in Fig. 3(b) and the purple module in Fig. 3(a) and Fig. 3(c)), and the input video has N frames. N temporal comparisons; spatial attention compares each imaging module with all other modules in the current frame (the blue module in Fig. 3(b) and all other red modules in Fig. 3(b)).



**(a)** **(b)** **(c)**

**Fig. 3.** **(a)** frame t − n. **(b)** frame n. **(c)** frame t + n

Extracting frames from the input video is adopted to extract features to improve the speed of extracting video features and reduce the waste of unnecessary computing resources. The extraction method is the average frame extraction method. For example, if F frames of video are extracted, one frame of image is extracted for every F/f frame of the video. Finally, the extracted video frame is input into the pre-trained TimeSformer model to obtain a high-dimensional feature vector.

### 2.3 Classification of Behavior States

After the video feature information is extracted from the second stage, a classifier is required to discriminate the video content. The support vector machine (SVM) [6] is a supervised learning algorithm. It maps the vectors of the data set into a high-dimensional feature space by the kernel function. Select a hyperplane in the area to classify the dataset so that vectors with different features are divided into both sides of the plane. If there are multiple such planes, the maximum distance between the two sides of the hyperplane is selected as the optimal plane.

SVM can effectively solve the small sample and binary classification problem, consistent with the fall detection to be achieved: the video data set selected in this paper is relatively small. The final classification can be regarded as a binary classification (fall and normal), so this paper chooses SVM as the classifier.

## 3 Experiments and Results Analysis

The experiment uses python3.8, OpenCV, PyTorch, and Sklearn in the PyCharm compiler environment. The experimental video size is 224 × 224, the frame rate is 24fps, the threshold T2 is 8, the penalty coefficient of the support vector machine is 0.9, and the kernel function is radial basis kernel function ( RBF).

### 3.1 Data Sets and Processing

This paper selects two public datasets, the Le2i fall dataset [7], the Multiple cameras fall dataset [8], and the video collected in this experiment. Le2i dataset contains 191 videos from four different scenarios; the Multiple datasets include 24 scenes, each recorded from different angles using eight cameras; The collected data sets mainly include some common daily behaviors (falls, walking, etc.). The final dataset has a total of 700 videos (350 positive samples and 350 negative samples). The ratio of the training set and the test set is 4:1. Due to the uncertainty of the random partitioning of the dataset, the final model predictions may be questionable. Therefore, the cross-validation method combines the segmented data sets into different training sets and test sets. This paper is divided into ten groups.

### 3.2 Performance Index

The binary classification problem can divide the sample into four cases according to the combination of the actual category and the predicted category: true positive (TP, fall detection is fall); false positive (FP, non-fall detection is fall), true negative (TN, non-fall detection is non-fall) and false negative (FN, fall detection is non-fall). This paper selects the commonly used indicators of classification and evaluation: precision, recall rate, and curacy are used for performance measurement; the F1 measure is used to balance precision and recall. The calculation formula is as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{6}$$

$$F1 = \frac{2\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

### 3.3 Experimental Process and Results

In this 2.2, we need to extract the feature vector of video information by frame extraction. For video tasks, the video is generally extracted 8,16,24,32 frames. As the number of frames increases, the calculation speed of the model will be slower. This paper extracts the video data set into several standard frames and inputs it into the pre-trained TimeSformer model to extract high-dimensional feature vectors (768 dimensions). Each video corresponds to a feature vector to form a new vector data set. Finally, the support vector machine is used for training modeling and testing. Table 1 shows the classification results and indicators after extracting different frames.

**Table 1.** Test results of each frame number

| Frame | Precision rate | Recall rate | Accuracy rate | F1 |
|---|---|---|---|---|
| 8 frames | 91.96% | 94.28% | 93.01% | 93.11% |
| 16 frames | 92.91% | 95.43% | 94.07% | 94.15% |
| 24 frames | 94.23% | 94.36% | 94.29% | 94.29% |
| 32 frames | 93.02% | 91.42% | 92.28% | 92.21% |

Although the overall effect of the model is the best when 24 frames of image data are extracted, 16 frames have a higher accuracy rate, and the accuracy rate is minimal because it should maintain a higher recall rate while meeting the accuracy rate. Because the two types of errors in detecting a fall as a fall and a fall as a fall are both misjudgments in statistics, but in practical applications, it is more desirable to have fewer occurrences of the latter category, and a higher recall rate is required. The reasoning speed of 16 frames is faster than that of 24 frames, so the method of extracting 16 frames is finally selected. In addition, training and testing experiments are performed only on the Le2i fall dataset and the Multiple cameras fall dataset. Table 2 shows the results, and Fig. 4 shows an example of overall model checking.

**Table 2.** Test results of each dataset

| The data set | Precision rate | Recall rate | Accuracy rate | F1 |
|---|---|---|---|---|
| Le2i | 96.37% | 98.52% | 97.36% | 97.44% |
| Multiple | 91.89% | 96.26% | 93.88% | 94.02% |

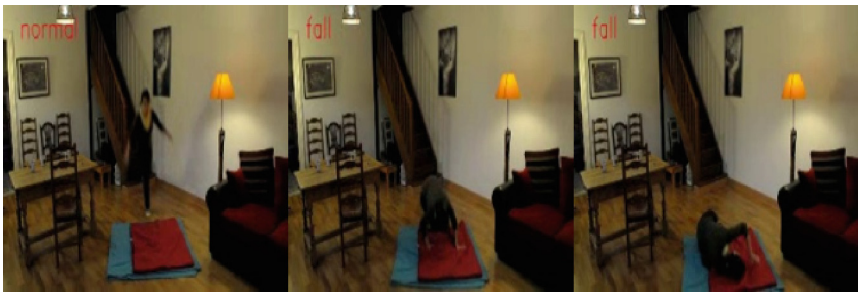



**Fig. 4.** Fall detection process

## 4 Conclusions

Fall detection is a way to monitor the lives of the elderly efficiently. This paper proposes a fall detection method based on video data input. Fall detection mainly involves motion detection, video feature extraction, and classification. The combination of neural network and support vector machine is used to analyze and discriminate the input video data containing motion. The experimental results show that the method can achieve better results. However, the model's generalization ability is affected due to the limited samples of the overall video data set. Therefore, it is necessary to construct a richer data set to improve the detection ability of the whole model.

## References

1. Suqingqing, M., et al.: Analysis of fall status and influencing factors of the elderly in the senile apartment. J. Nurs. **37**(12), 16–18 (2022)
2. Ramachandran, A., Karuppiah, A., Gigantesco, A.: A survey on recent advances in wearable fall detection systems. BioMed Res. Int. (2020)
3. Vaswani, A., Shazier, N., Parmar, N., et al.: Attention Is All You Need (2017)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? (2021)
6. Zhihua, Z.: Machine learning. Tsinghua University Press, Beijing (2016)
7. Charfi, I., Miteran, J., et al.: Optimized Spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification. J. Electron. Image. **22**(4), 041106 (2013)
8. Adventinet, E., Rougier, C., et al.: Multiple cameras fall dataset, Technical report 1350, DIRO-Université de Montréal, July 2010

# An Improved Grey Wolf Optimizer for Numerical Optimization

Linyun Ma[1,2], Ying Tong[1,2(✉)], Baozhu Han[1,2], and Xing Zhang[1,2]

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
tongying2334@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** Aiming at the problems of Grey Wolf Optimizer (GWO), such as slow convergence rate in the late iteration, lack of mechanism to jump out of local optimum. In this paper, an improved Grey Wolf Optimizer (IGWO) is proposed. In this algorithm, a new update equation is used to improve the search speed of the algorithm. Meanwhile, the forgetting mechanism is introduced, which improves the diversification of the population and enables the algorithm to jump out of local optimum. Simulation experiments on the CEC2017 test set show that compared with standard GWO, the results obtained by IGWO have higher accuracy, which proves that the algorithm has better performance in dealing with complex optimization problems.

**Keywords:** global optimization · meta-heuristic algorithm · grey wolf optimizer · forgetting mechanism

## 1 Introduction

Meta-heuristic algorithms are inspired by nature, compared with traditional optimization algorithms, meta-heuristic algorithms have a wide range of applicability, insensitivity to initial values, high search efficiency, and ease of implementation. Among them, the more representative algorithms are Simulated Annealing (SA) [1] Particle Swarm Optimization (PSO) [2], Ant Colony Optimization (ACO) [3], Artificial Bee Colony Algorithm (ABC) [4], and Salp Swarm Algorithm (SSA) [5], etc. Due to the excellent performance and flexibility of metaheuristic optimization algorithms, they have been widely used in various practical problems, such as path planning problem [6], PV model parameter optimization [7] and BP neural network parameter optimization [8].

Gray wolves are excellent predators in nature, they are good at teamwork and possess efficient hunting efficiency. Inspired by this, Mirjalili et al. [9] proposed the Grey Wolf Optimizer (GWO) algorithm for solving optimization problems. In the literature [9], GWO was compared with PSO, GSA, DE and other optimization algorithms in terms of performance and achieved good results. Since then, GWO has also been used for


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 130–137, 2023.
https://doi.org/10.1007/978-981-99-2362-5_17

optimization in several application areas, such as image fusion [10], kernel limit learning machine problems [11], and battery parameter problems [12].

Despite the excellent performance of GWO, there is still room for improvement. The search process of GWO is one-way and lacks the ability to jump out of the local optimum, so it is easy to fall into the local optimum when dealing with multimodal problems. To this end, this paper proposes an improved Grey Wolf Optimizer (IGWO). IGWO has a new update equation and perturbation strategies. To verify the performance of IGWO, this paper will test IGWO using the CEC2017 test set [13].

## 2 Origin Grey Wolf Optimizer

Grey Wolf Optimizer (GWO) is a metaheuristic optimization algorithm proposed by Mirjalili et al. [9] that mimics the social hierarchy and hunting behavior of wolves. Similar to most metaheuristic algorithms, GWO will also represent the solutions distributed in different locations in the space with each individual (also called wolf in GWO). And GWO is characterized by denoting the current optimal solution, the second optimal solution, and the third optimal solution by $\alpha$, $\beta$, $\Omega$, respectively, and the remaining gray wolf individuals by $\Omega$.

The algorithm will first simulate the hunting behavior of a wolf pack in a roundup, as described by the following Eq. (1)–(2).

$$X(t+1) = X_p(t) - A \times B \tag{1}$$

$$B = |D \times X_P(t) - X(t)| \tag{2}$$

where $t$ denotes the number of current iterations, and $X_p$ and $X$ denote the position of the prey and the position of the gray wolf, respectively, $A = a(2r_1 - 1)$ and $D = 2ar_2$ are coefficient vectors, where $r_1$ and $r_2$ are both random numbers of [0, 1] and $a$ is a linearly decreasing coefficient from 2 to 0, as follows Eq. (3).

$$a = 2 - (2t/T) \tag{3}$$

where $T$ denote the maximum number of iterations.

In the hunting process, the wolves are then guided by $\alpha$, $\beta$, and $\Omega$ to complete the hunt as described by the following Eq. (4)–(6). Finally, the final hunt is completed by Eq. (7).

$$X_1(t+1) = X_\alpha(t) - A_1 \times |D_1 \times X_\alpha - X| \tag{4}$$

$$X_2(t+1) = X_\beta(t) - A_2 \times \left|D_2 \times X_\beta - X\right| \tag{5}$$

$$X_3(t+1) = X_\Omega(t) - A_3 \times |D_3 \times X_\Omega - X| \tag{6}$$

$$X(t+1) = \frac{X_1'(t) + X_2'(t) + X_3'(t)}{3} \tag{7}$$

where, $X_\alpha$, $X_\beta$, $X_\Omega$ are the positions corresponding to $\alpha$, $\beta$, $\Omega$.

## 3 Improved Grey Wolf Optimizer

### 3.1 Enhance the Update Equation

Since the update equation of the original GWO is jointly guided by three high-quality individuals in $\alpha$, $\beta$ and $\Omega$, making full use of the information of the population, the equation has excellent exploratory ability. Therefore, this paper designs a new update equation to enhance the exploitation of GWO as follows:

$$X(t+1) = \frac{X_1'(t) + X_2'(t) + X_3'(t)}{3} + G \tag{8}$$

$$G = l_1 \times rand \times (X_\alpha - X(t)) + l_2 \times rand \times (P - X(t)) \tag{9}$$

$$l_{1,2} = \frac{3}{2} \times \cos\left(\frac{1}{2}\pi\left(\frac{t}{T}\right)^2\right) + \frac{1}{2} \tag{10}$$

where $G$ as a guide vector, $P$ is individual historical optimum, $l_1$ and $l_2$ and are adaptive coefficients as shown in Eq. (10).

### 3.2 Forgetting Mechanism

Although the improved strategy in the previous section enhances the exploitation ability of the algorithm, it may also cause the wolves to cluster together too early, thus losing enough search space and increasing the possibility of falling into local optimum. Therefore, a hybrid strategy of forgetting mechanism and contrastive learning is introduced in this paper.

The forgetting mechanism is related to the distance of each wolf to the global optimum, so it is necessary to calculate their distances $Dst_i$ to the global optimum according to Eq. (11) beforehand and rank them in ascending order. Secondly, it is determined how many individuals in the population need to produce the forgetting phenomenon, and this number is denoted by $FN$. Finally, it is necessary to determine the number of dimensions in which each individual needs to produce forgetting, and this number is denoted by $FD$. $FN$ and $FD$ are shown in Eq. (12) and Eq. (13).

$$Dst_i = \sqrt{\sum_{j=1}^{D} \left(X_{i,j} - X_{\alpha,j}\right)^2} \tag{11}$$

$$FN = \left\lfloor N\left(Rmax - (Rmax - Rmin) \cdot \frac{t}{T}\right)\right\rfloor \tag{12}$$

$$FD = \left\lfloor \frac{2}{3} \cdot \frac{(f_i - f_\alpha) \cdot D}{f_i} \right\rfloor \tag{13}$$

where $D$ is the total dimension of the objective function, and $X_{i,j}$ and $X_{\alpha,j}$ are the $j$-$th$ dimension of the $i$-$th$ individual and the current global optimum, respectively; $N$ is the

population size, and *Rmax* and *Rmin* and are both parameters controlling the oblivion range; and $f_i$ and $f_\alpha$ are the values of the fitness functions of the *i-th* individual and the current global optimum, respectively.

Next, it is necessary to generate perturbations using contrastive learning. Unlike the operation of conventional contrastive learning [21, 22], in this paper, contrastive learning does not directly make changes to the wolf population, but acts as an intermediate variable to guide the corresponding wolves in their search. The scheme of dyadic learning is shown in Eq. (14).

$$G_j = C \times (U + L) - X_{\alpha,j} \tag{14}$$

where $G_j$ is the *j-th* dimension of the bootstrap factor, and $U$ and $L$ are the maximum and minimum values in the jth dimension of the current population, respectively. $C$ is the scaling factor.

Since the purpose of the strategy in this subsection is to generate perturbations, it can be regarded as a kind of mutation, which is not conducive to the balance of algorithm development capability and exploration capability if it is too active. Thus, to avoid such an imbalance, this paper also sets a threshold *Hmax* as a trigger condition in IGWO.

### 3.3 Framework of IGWO

In this paper, IGWO is proposed by introducing a strategy of forgetting mechanism mixed with contrastive learning after improving the GWO update equation with enhanced exploitation capability, and the pseudo code is shown Algorithm 1.

**Algorithm 1** IGWO

```
Initialize the population Xi(i = 1, 2, 3, ⋯, N) and select Xα, Xβ, XΩ
WHILE t<T DO
    Calculate Dsti according to equation (10) and Sort Xi according to Dsti
    FOR i=1 to N DO
        Calculate FN according to equation (11)
        IF i<N-FN DO
            IF H<Hmax DO
                Calculate FD according to equation (12)
                Calculate O according to equation (13)
            END IF
        END IF
        Update Xi according to equation (8)
    END FOR
    Update Xα, Xβ, Xδ
    IF Xα changes DO
        H=0
    ELSE DO
        H=H+1
    END IF
    t=t+1
END WHILE
Output Xα
```

# 4 Experiment and Results Analysis

## 4.1 Test Function and Parameters Setting

To verify the performance of IGWO, this paper uses the CEC2017 test set [1] to test IGWO. Compared with the conventional benchmark test functions, CEC2017 is more difficult and closer to the target function of the actual problem.

There are 30 test functions in CEC2017, but since F2 is considered unstable [13, 14], this experiment does not include F2 and only the remaining 29 functions are used for testing. Among them, F1, F3 are unimodal functions, which mainly test the development ability of the algorithm; F4-F10 are multimodal functions, which mainly test the exploration ability of the algorithm; F11-F20 are combinatorial functions and F21-F30 are composite functions, which mainly test the balance of the development ability and exploration ability of the algorithm. The details of each function are described in the literature [13].

The parameters of the five algorithms involved in the experiments are shown in Table 1.

**Table 1.** Parameters setting of algorithms

| Algorithm | Parameters setting |
|---|---|
| GWO | $a = 2 - (2t/T)$ |
| NGWO | $a = 2 - (2t/T)$, $Rmin = 0.7$, $Rmax = 0.9$, $Hmax = 10$, $C = 1$ |

## 4.2 Compared with GWO

In this section, IGWO is compared with GWO. The maximum number of iterations is fixed at 10000, the population size is fixed at 50, and the number of dimensions is fixed at 50. To verify the improvement effect of IGWO on the original algorithm in terms of statistical results, each algorithm is run independently on the test function 30 times and the Wilcoxon rank sum test is added. Where the significance level is fixed at 0.05, with "+" indicates that IGWO is significantly better than the corresponding comparison algorithm, "−" indicates that the corresponding algorithm is significantly better than IGWO, and " = " indicates that there is no significant difference between IGWO and the corresponding algorithm. The mean values of the run results and the Wilcoxon rank sum test are shown in Table 2.

In Table 2, we can see that IGWO outperforms GWO for most of the functions. Unimodal functions (F1, F3), IGWO outperforms GWO for all of them; multimodal functions (F4-F10), IGWO outperforms GWO for functions F4-F9; combined functions (F11-F20) and composite functions (F21-F30), IGWO outperforms GWO for functions F11-F14, F15, F18, F19, F21, F23, F24-F28, F30. IGWO obtains better results than GWO on functions F11-F14, F15, F18, F19, F21, F23, F24-F28, and F30. The Wilcoxon rank

**Table 2.** Results of IGWO and GWO on CEC2017 through 30 independent runs on CEC2017.

| Fun | IGWO | GWO | |
|---|---|---|---|
| F1 | **6.6297E+07** | 6.8283E+09 | + |
| F2 | N/A | N/A | |
| F3 | **4.9536E+04** | 7.8154E+04 | + |
| F4 | **6.0188E+02** | 1.0296E+03 | + |
| F5 | **6.3705E+02** | 7.1367E+02 | + |
| F6 | **6.0182E+02** | 6.1512E+02 | + |
| F7 | **9.0224E+02** | 1.0456E+03 | + |
| F8 | **9.4190E+02** | 1.0149E+03 | + |
| F9 | **1.6668E+03** | 6.1976E+03 | + |
| F10 | 1.4300E+04 | **6.8183E+03** | - |
| F11 | **1.4356E+03** | 3.9405E+03 | + |
| F12 | **4.3633E+07** | 5.8369E+08 | + |
| F13 | **4.5941E+05** | 1.6717E+08 | + |
| F14 | **2.7965E+05** | 4.2571E+05 | = |
| F15 | **3.4784E+05** | 4.7096E+06 | = |
| F16 | 3.1646E+03 | **2.9830E+03** | = |
| F17 | 2.9915E+03 | **2.6558E+03** | - |
| F18 | **1.6710E+06** | 3.1266E+06 | = |
| F19 | **3.9636E+05** | 2.6634E+06 | + |
| F20 | 3.4868E+03 | **2.7782E+03** | - |
| F21 | **2.4314E+03** | 2.4913E+03 | + |
| F22 | 1.5825E+04 | **8.4827E+03** | - |
| F23 | **2.8642E+03** | 2.9615E+03 | + |
| F24 | **2.9921E+03** | 3.1471E+03 | + |
| F25 | **3.0703E+03** | 3.4649E+03 | + |
| F26 | **5.2092E+03** | 6.0892E+03 | + |
| F27 | **3.3069E+03** | 3.5173E+03 | + |
| F28 | **3.3908E+03** | 4.2374E+03 | + |
| F29 | 4.2865E+03 | **4.2793E+03** | = |
| F30 | **1.9580E+07** | 8.0075E+07 | + |

sum test result "+/ = /−" is "20/5/4", indicating that IGWO significantly outperforms GWO on most of the tested functions.

In summary, IGWO achieves more satisfactory results than GWO. Therefore, it is proved that the strategy introduced in this paper not only improves the algorithm development ability and exploration ability, but also achieves a balance of these two abilities.

## 5 Conclusion

In this paper, an improved grey wolf optimization algorithm (IGWO) is proposed to address the shortcomings of GWO. The algorithm proposes two strategies for improving the performance of GWO. The first strategy is an improved update equation, which strengthens the algorithm's ability for local search by introducing a new bootstrap factor. The second strategy is a hybrid strategy of forgetting mechanism and contrastive learning, which first determines how many individuals need to undergo perturbation by forgetting mechanism, and then uses contrastive learning to make them generate perturbation in order to increase population diversity and jump out of local optimum. The test results of IGWO on the CEC2017 test set show that the algorithm strikes a good balance between exploration ability and exploitation ability. The experimental results show that IGWO has a significant performance improvement relative to GWO. In the future, IGWO is expected to be used for optimization of more practical problems, e.g., path planning, feature extraction, training of neural networks, and various types of multi-objective optimization problems.

## References

1. Franzin, A., Stützle, T.: Revisiting simulated annealing: a component-based analysis. Comput. Oper. Res. **1**, 191–206 (2019). https://doi.org/10.1016/j.cor.2018.12.015
2. Jordehi, A.R.: Particle swarm optimisation (PSO) for allocation of FACTS devices in electric transmission systems: a review. Renew. Sustain. Energy Rev. **52**, 1260–1267 (2015). https://doi.org/10.1016/j.rser.2015.08.007
3. Dorigo, M., Birattari, M., Stutzle, T.: Ant colony optimization. IEEE Comput. Intell. Mag. **1**(4), 28–39 (2006). https://doi.org/10.1109/MCI.2006.329691
4. Karaboga, D.: An idea based on honey bee swarm for numerical optimization, Technical report - TR06, Technical Report, Erciyes University (2005)
5. Mirjalili, S., Gandomi, A.H., Mirjalili, S.Z., et al.: Salp swarm algorithm: a bio-inspired optimizer for engineering design problems. Adv. Eng. Softw. **114**, 163–191 (2017). https://doi.org/10.1016/j.advengsoft.2017.07.002
6. Jia, Y., Qu, L., Li, X.: A double-layer coding model with a rotation-based particle swarm algorithm for unmanned combat aerial vehicle path planning. Eng. Appl. Artif. Intell. **116**, 105410 (2022). https://doi.org/10.1016/j.engappai.2022.105410
7. Chen, H., Jiao, S., Heidari, A.A., et al.: An opposition-based sine cosine approach with local search for parameter estimation of photovoltaic models. Energy Convers. Manage. **195**, 927–942 (2019). https://doi.org/10.1016/j.enconman.2019.05.057
8. Ren, C., An, N., Wang, J., et al.: Optimal parameters selection for BP neural network based on particle swarm optimization: a case study of wind speed forecasting. Knowl. Based Syst. **56**, 226–239 (2014). https://doi.org/10.1016/j.knosys.2013.11.015

9. Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. Adv. Eng. Softw. **69**, 46–61 (2014). https://doi.org/10.1016/j.advengsoft.2013.12.007
10. Daniel, E., Anitha, J., Kamaleshwaran, K.K., et al.: Optimum spectrum mask based medical image fusion using Gray Wolf optimization. Biomed. Signal Process. Control **34**, 36–43 (2017). https://doi.org/10.1016/j.bspc.2017.01.003
11. Cai, Z., Gu, J., Luo, J., et al.: Evolving an optimal kernel extreme learning machine by using an enhanced grey wolf optimization strategy. Exp. Syst. Appl. **138** (2019). https://doi.org/10.1016/j.eswa.2019.07.031
12. Ali, M., El-Hameed, M.A., Farahat, M.A.: Effective parameters' identification for polymer electrolyte membrane fuel cell models using grey wolf optimizer. Renew. Energy **111**, 455–462 (2017). https://doi.org/10.1016/j.renene.2017.04.036
13. Wu, G., Mallipeddi, R., Suganthan, P.: Problem definitions and evaluation criteria for the CEC 2017 competition and special session on constrained single objective real-parameter optimization (2016)
14. Gupta, S., Deep, K.: Improved sine cosine algorithm with crossover scheme for global optimization. Knowl. Based Syst. **165**, 374–406 (2019). https://doi.org/10.1016/j.knosys.2018.12.008

# A Fault Diagnosis Method for Power Transformer Using Canonical Variate Analysis and Support Vector Machine

Long Luo[1], Yan Li[1], Yan Shi[1], Ting Han[1], Wencui Yang[1], Xiaojun Jin[1], and Di Han[2(✉)]

[1] Extra High Voltage Company of State Grid Qinghai Electric Power Company, Beijing, China

[2] Advanced Smart Electric Information Technology, Hangzhou 310000, China

handi202111@163.com

**Abstract.** The transformer is one of the most important units in the power grid. Due to the potential failures and costs of the power system, it is necessary to pay attention to the fault diagnosis of power transformers. This paper proposes a fault diagnosis method based on Canonical Variate Analysis and Support Vector Machine (CVA-SVM). As a system identification method, CVA is widely used for fault detection because of its ability to identify multivariate state space models using experimental data. The support vector machine is a new machine learning method and is a powerful tool for solving problems with nonlinear and non-Gaussian distributed data. Dissolved gas analysis (DGA) has shown great potential for detecting faults in power transformers. For fault diagnosis based on DGA, a CVA model is first constructed for the process variables to generate a series of feature vectors, and then the fault types are classified using SVM. A real power transformer process is employed to verify the effectiveness of the proposed method.

**Keywords:** Canonical Variate Analysis · Support Vector Machine · Power transformer · fault diagnosis

## 1 Introduction

As an important piece of equipment in a power system, an interruption of power supply can be caused by any failure in the power transformer. Therefore, it is vital to detect transformer faults [1–3]. Dissolved gas analysis (DGA) has been widely recognized as an effective diagnostic technique for fault detection in power transformers. The analysis of specific dissolved gas concentrations in transformer insulating oil yields knowledge about the condition of the transformer and allows necessary preventive measures to be taken based on the results of the process [4–6]. However, due to the variability of gas data and the nature of the operation, fault detection by conventional methods is not always an easy task.

This work was supported by the key technology project of the State Grid Corporation of China under Grant 522821200090.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 138–146, 2023.
https://doi.org/10.1007/978-981-99-2362-5_18

To develop more accurate diagnostic tools based on DGA data, scholars have developed a number of artificial intelligence methods [7,8]. With the development of machine learning, the fault diagnosis of power transformers has also been enhanced. To cope with the uncertainty in fault diagnosis, Huang et al. proposed a fuzzy logic-based fault diagnosis method for power transformers, where the technique can diagnose multiple faults in a transformer and quantitatively indicate the severity of each fault [9]. To reduce the redundant information of the data, Kari et al. proposed to reduce the dimensionality of the data with principal component analysis and detect power transformer faults using fuzzy C-means method [10].

To develop more accurate diagnostic tools based on DGA data, scholars have developed many artificial intelligence methods [7,8]. With the development of machine learning, the fault diagnosis of power transformers has also been enhanced. To cope with the uncertainty in fault diagnosis, Huang et al. proposed a fuzzy logic-based fault diagnosis method for power transformers, where the technique can diagnose multiple faults in a transformer and quantitatively indicate the severity of each fault [9]. To reduce the redundant information of the data, Kari et al. proposed to reduce the dimensionality of the data with principal component analysis and detect power transformer faults using the fuzzy C-means method [10].

However, the above methods do not take into account the dynamic nature of power transformer data, and for continuously operating systems, may fail to explore valuable dynamic information for the process and lead to some misleading monitoring results [11]. CVA is widely used in a dynamic process to generate a state-space model from data by maximizing the correlation between the constructed "past" and "future" matrices [12]. To the best of the authors' knowledge, the CVA method has not been used in the data processing of power transformers for fault diagnosis.

Motivated by the above discussion, considering the characteristics of CVA and SVM, a new fault diagnosis method is proposed by combining CVA and SVM for the power transformer process. First, CVA extracts the dynamic features of the process data. Based on the extracted features. And then SVM is employed to classify the fault types to address the issues of no-Gaussian assumption and nonlinearity. For the parameter optimization problem in SVM, this paper uses the random grid search cross-validation method to improve the accuracy of the model.

This paper is organized as follows. Section 2 briefly reviews the CVA and SVM. Section 3 is devoted to describing the proposed CVA-SVM method. Section 4 presents the application of the proposed method in the real power transformer data. Finally, conclusions are given in Sect. 5.

## 2 Review of CVA and SVM

### 2.1 CVA

CVA is based on the so-called subspace identification, where process measurements are stacked to form the past and future spaces [13]. Denote $\mathbf{x_k} \in \mathbb{R}^m$ $(k = 1, 2, \ldots, N)$ as the normalized stacked vector at time instant $k$. For each $k$, the past data vector $\mathbf{x}_{p,k}$ and future data vector $\mathbf{x}_{f,k}$ are collected as

$$\mathbf{x}_{p,k} = \begin{bmatrix} \mathbf{x}_{k-1} \\ \mathbf{x}_{k-2} \\ \vdots \\ \mathbf{x}_{k-l} \end{bmatrix} \in \mathbb{R}^{ml}, \mathbf{x}_{f,k} = \begin{bmatrix} \mathbf{x}_{k} \\ \mathbf{x}_{k+1} \\ \vdots \\ \mathbf{x}_{k+l-1} \end{bmatrix} \in \mathbb{R}^{ml} \tag{1}$$

where $l$ is the number of time lag. For a finite sequence with $N$ samples, the past and future Hankel matrices $\mathbf{X}_p$ and $\mathbf{X}_f$ are constructed,

$$\begin{aligned} \mathbf{X}_p &= [\mathbf{x}_{p,l+1}, \mathbf{x}_{p,l+2}, \ldots, \mathbf{x}_{p,l+M}] \in \mathbb{R}^{ml \times M} \\ \mathbf{X}_f &= [\mathbf{x}_{f,l+1}, \mathbf{x}_{f,l+2}, \ldots, \mathbf{x}_{f,l+M}] \in \mathbb{R}^{ml \times M} \end{aligned} \tag{2}$$

where $M = N - 2l + 1$. The estimates of the sample covariance and cross-covariance of the past and future vector are expressed below,

$$\begin{bmatrix} \Sigma_{pp} & \Sigma_{pf} \\ \Sigma_{fp} & \Sigma_{ff} \end{bmatrix} = \frac{1}{M-1} \begin{bmatrix} \mathbf{X}_p \mathbf{X}_p^\top & \mathbf{X}_p \mathbf{X}_f^\top \\ \mathbf{X}_f \mathbf{X}_p^\top & \mathbf{X}_f \mathbf{X}_f^\top \end{bmatrix} \tag{3}$$

In CVA, the projection matrices $\mathbf{J}$ and $\mathbf{L}$ can be computed through performing singular value decomposition (SVD) on the Hankel matrix $\mathbf{H}$,

$$\mathbf{H} = \Sigma_{ff}^{-1/2} \Sigma_{fp} \Sigma_{pp}^{-1/2} = \mathbf{U} \Lambda \mathbf{V}^\top \tag{4}$$

Here, $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular matrices of the matrix $\mathbf{H}$, respectively. $\Lambda = diag[\sigma_1, \sigma_2, \ldots, \sigma_q]$ is the diagonal matrix containing all singular values, and $q$ is the rank of $\mathbf{H}$.

From the result of SVD, the projection matrices $\mathbf{J}$ and $\mathbf{L}$ can be calculated. The first $r$ columns of $\mathbf{V}$ can be considered to have the highest pairwise correlation with the first $r$ columns of $\mathbf{U}$ [14]. It produces a pair of new matrices $\mathbf{U}_r \in \mathbb{R}^{ml \times r}$ and $\mathbf{V}_r \in \mathbb{R}^{ml \times r}$ with smaller dimensionality.

$$\begin{aligned} \mathbf{J} &= \mathbf{V}_r^\top \Sigma_{pp}^{-1/2} \in \mathbb{R}^{r \times ml} \\ \mathbf{L} &= (\mathbf{I} - \mathbf{V}_r \mathbf{V}_r^\top) \Sigma_{pp}^{-1/2} \in \mathbb{R}^{ml \times ml} \end{aligned} \tag{5}$$

Finally, two matrices contain the state and residual vectors are derived below,

$$\begin{aligned} \mathbf{Z} &= \mathbf{J} \mathbf{X}_p \in \mathbb{R}^{r \times \mathbf{M}} \\ \mathbf{E} &= \mathbf{L} \mathbf{X}_p \in \mathbb{R}^{ml \times \mathbf{M}} \end{aligned} \tag{6}$$

### 2.2 SVM

As illustrated in Fig. 1, a Support Vector Machine aims to find an optimal hyperplane by maximally separating the margins between the hyperplane and the data [15,16].

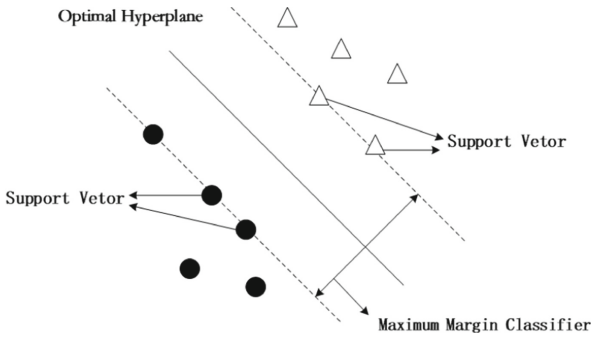



**Fig. 1.** Separation of two classes by SVM.

Given a data set $F = \{x_i, y_i\}_{i=1}^{m}$, where $m$ is the sample number, $x_i \in R^n$ stands for the input vectors, $y_i \in +1, -1$ denotes two classes. The hyperplane $f(x) = 0$ that separates the given data can be determined when the two classes are linearly differentiable.

$$f(x) = w \cdot x + b = \sum_{i=1}^{m} w_k \cdot x_k + b = 0 \tag{7}$$

where $w$ denotes the weight vector and $b$ denotes the bias term. The separation hyperplane should satisfy the following constraints,

$$y_i f(x_i) = y_i(w \cdot x_i + b) \geq 1, i = 1, 2, \ldots m \tag{8}$$

For solving the linear indivisibility problem, the relaxation variable $\zeta_i$ is introduced, and thus the constraint becomes as,

$$\begin{array}{ll} Min & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\zeta_i, i = 1, \ldots, m \\ S.t. & \begin{cases} y_i(w \cdot x_i + b) \geq 1 - \zeta_i) \\ \zeta \geq 0 \end{cases} \end{array} \tag{9}$$

where $C$ is the error penalty.

The above optimization problem is transformed into a pairwise quadratic optimization problem by introducing the Lagrange multiplier $\alpha_i$ , i.e.

$$\begin{array}{ll} Max & L(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i, x_j) \\ S.t. & \sum_{i=1}^{m}\alpha_i y_i = 0, \alpha_i \geq 0, i = 1, \cdots, m \end{array} \tag{10}$$

The linear decision function is therefore created by solving a pairwise optimization problem defined as,

$$f(x) = sign(\sum_{i,j=1}^{m} \alpha_i y_i (x_i, x_j) + b) \tag{11}$$

SVM can be used for nonlinear classification. By using a nonlinear mapping function, the original data $x$ is mapped to a high-dimensional feature space in which linear classification can be performed. Then the decision function is transformed into,

$$f(x) = sign(\sum_{i,j=1}^{m} \alpha_i y_i K(x_i, x_j) + b) \tag{12}$$

In this paper, the Gaussian Kernel is selected as kernel function,

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)) = exp(- \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2 / h) \tag{13}$$

where $\phi$ is a nonlinear mapping that maps data points to the high-dimensional feature space. To obtain a tighter boundary, an appropriate width parameter $h$ of the Gaussian kernel function is selected.

## 3 CVA-SVM Based Fault Diagnosis

In the proposed CVA-SVM method, the space of canonical variables can be divided into the state space and the residual space. The state space is then used as target objects for developing SVM hypersphere layers. Finally, The SVM faults classification is performed. The procedure of the CVA-SVM based fault detection method is depicted (Table 1).

As shown in Fig. 2, two phases are included, offline training and online diagnosis. Specifically, the procedure of the CVA-SVM based fault diagnosis is described in detail as follows,

**Offline training:**

- Step 1. Standardize the collected faulty measurements.
- Step 2. Construct the Hankel matrices $\mathbf{X}_f$ and $\mathbf{X}_p$ with the determined time-lag $l$.
- Step 3. Obtain the projection matrices $\mathbf{J}$ and $\mathbf{L}$ according to Eq.(5)
- Step 4. Determine the state and residual matrices $\mathbf{Z}$ and $\mathbf{E}$ using Eq.(6).
- Step 5. Build SVM model for $\mathbf{Z}$ with the determined $C$ and $h$.
- Step 6. The SVM classifier is trained using the appropriate values of parameters.

**Online diagnosis:**

- Step 1. Obtain and Standardize the test sample $\mathbf{x}_k^t$.

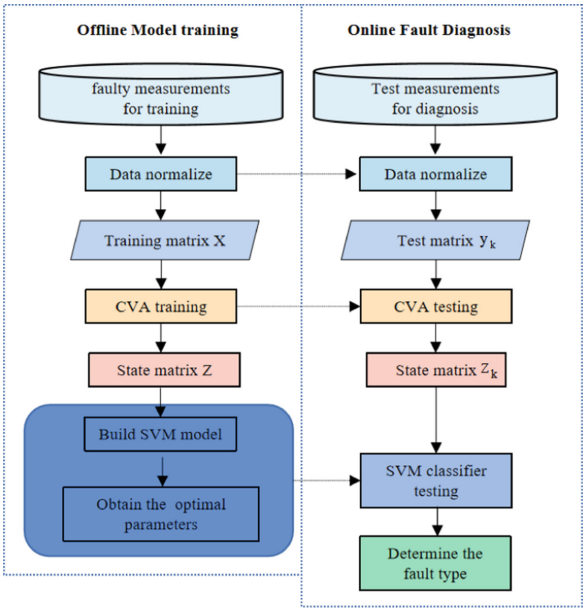



**Fig. 2.** Main steps of the CVA-SVM based fault diagnosis.

– Step 2. Construct stacked vectors and calculate the state and residual vectors from **J** and **L**,

$$\begin{aligned} \mathbf{z}_k &= \mathbf{J}\mathbf{x}_{p,k}^t \\ \mathbf{e}_k &= \mathbf{L}\mathbf{x}_{p,k}^t \end{aligned} \tag{14}$$

– Step 3.Input the state space $\mathbf{z}_k$ into the SVM classifier.
– Step 4. Obtain the diagnostic results.

## 4 Case Study

To verify the effectiveness of the CVA-SVM method proposed in this paper, 188 power transformer oil dissolved gas content faulty data were collected for the experiment. This data has 6 types of fault states and 5 components of dissolved gas content in oil, some of which are shown in Table 2. For computational convenience, we coded and labeled the fault types of the dataset, as shown in Table 3. After that, the data were divided into a training set and a test set, and the data in the test set and the training set each accounted for 50% of the original data.

By subjecting the gas data to the CVA algorithm analysis, the five data variables of transformer faults can be dimensionally reduced to four data variables. This will facilitate the linear partitioning of the data by the SVM classifier and

can also improve the computational speed of the fault diagnosis system. In addition, to compare the superior performance of the proposed methods, we compare the traditional SVM method, PCA and SVM combined algorithms. The classification result plots of the three methods are shown in Fig. 3.

The final classification accuracy of each model is summarized in Table 4. From Fig. 3, it is obtained that the SVM algorithm is less effective in identifying normal samples and has a lower detection rate for the medium to low temperature overheating and the high temperature overheating faults. The SVM method also has some false detection. The detection effectiveness of the PCA-SVM method is improved. Table 3 quantifies the detection effect of each model, and from the Table, we get that CVA-SVM has the highest accuracy for both the test set and the training set, and achieves the best classification effect.

**Table 1.** Comparison of fault diagnosis results.

| Method | Training accuracy (%) | Test accuracy(%) |
|---|---|---|
| SVM | 90.43 | 74.19 |
| PCA-SVM | 92.64 | 81.72 |
| CVA-SVM | 94.68 | 82.80 |

**Table 2.** Description of power transformer fault data.

| No | $H_2$ | $CH_4$ | $C_2H_6$ | $C_2H_4$ | $C_2H_2$ | Fault type |
|---|---|---|---|---|---|---|
| 1 | 568 | 26.5 | 6.9 | 2.1 | 0 | Partial discharge |
| 2 | 3433 | 180 | 34.4 | 3.6 | 0.4 | Partial discharge |
| 3 | 2083 | 85.6 | 18.4 | 2.6 | 0 | Partial discharge |
| 4 | 568 | 26.5 | 6.9 | 2.1 | 0.1 | Partial discharge |
| ... | ... | ... | ... | ... | ... | ... |
| 93 | 32 | 41.6 | 10 | 120 | 2.6 | High Temperature Overheating |
| ... | ... | ... | ... | ... | ... | ... |

**Table 3.** Codes for power transformer fault type.

| Fault type | Partial discharge | Low energy discharge | High energy discharge | Normal state | Medium to low temperature overheating | High temperature overheating |
|---|---|---|---|---|---|---|
| Fault code | 0 | 1 | 2 | 3 | 4 | 5 |

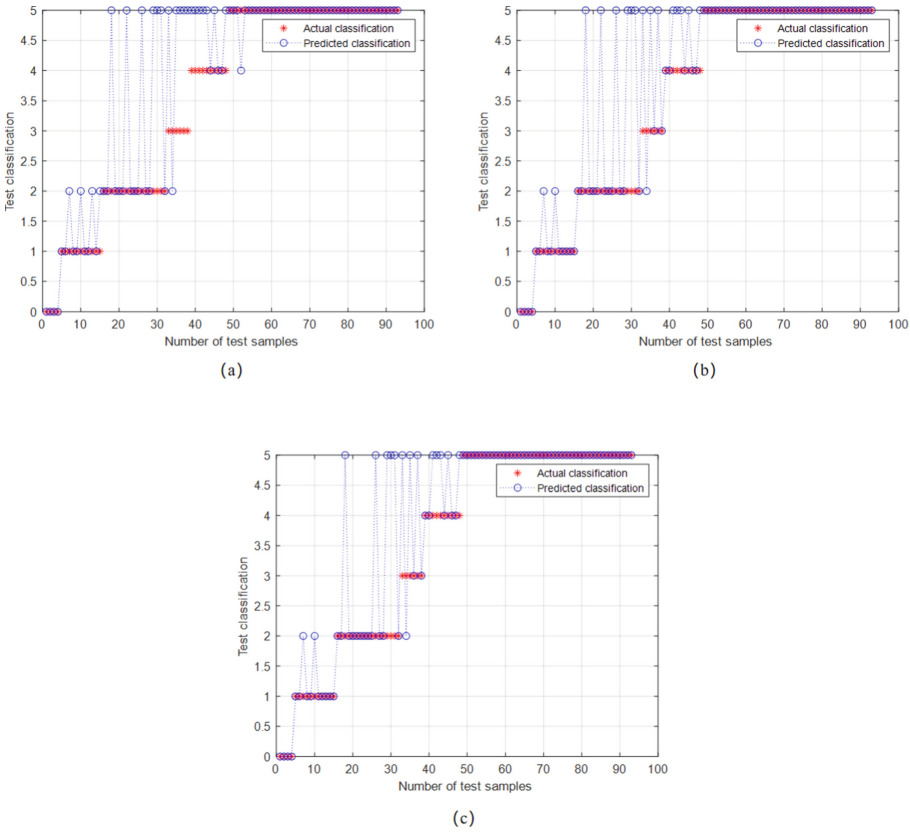



**Fig. 3.** Diagnostic results of power transformer faults based on three methods (a). SVM, (b) PCA-SVM, (c) CVA-SVM.

## 5 Conclusion

This paper proposes a design of a power transformer fault diagnosis system based on the optimized SVM kernel function model algorithm and optimizes the SVM model by collecting the data of five gases generated by oil fission when a transformer fault occurs. Compared with the traditional SVM and PCA-SVM methods, it can be seen that the CVA-SVM method can significantly improve the accuracy of transformer fault diagnosis. CVA can extract dynamic information from the data, so the optimized SVM model algorithm is more suitable for transformer fault diagnosis systems with high practicality. Further investigation is strongly recommended to extend power transformer fault identification methods.

## References


1. Fei, S.-W., Zhang, X.-B.: Fault diagnosis of power transformer based on support vector machine with genetic algorithm. Expert Syst. Appl. **36**(8), 11352–11357 (2009)
2. Wang, M.-H.: A novel extension method for transformer fault diagnosis. IEEE Trans. Power Delivery **18**(1), 164–169 (2003)
3. Kari, T., et al.: Hybrid feature selection approach for power transformer fault diagnosis based on support vector machine and genetic algorithm. IET Gen. Transm. Distrib. **12**(21), 5672–5680 (2018)
4. Bacha, K., Souahlia, S., Gossa, M.: Power transformer fault diagnosis based on dissolved gas analysis by support vector machine. Electric Power Syst. Res. **83**(1), 73–79 (2012)
5. Fang, J., Zheng, H., Liu, J., Zhao, J., Zhang, Y., Wang, K.: A transformer fault diagnosis model using an optimal hybrid dissolved gas analysis features subset with improved social group optimization-support vector machine classifier. Energies **11**(8), 1922 (2018)
6. Zhang, Y., et al.: A fault diagnosis model of power transformers based on dissolved gas analysis features selection and improved Krill Herd Algorithm optimized support vector machine. IEEE Access **7**, 102803–102811 (2019)
7. Sun, H.-C., Huang, Y.-C., Huang, C.-M.: Fault diagnosis of power transformers using computational intelligence: a review. Energy Procedia **14**, 1226–1231 (2012)
8. Bakar, N.A., Abu-Siada, A., Islam, S.: A review of dissolved gas analysis measurement and interpretation techniques. IEEE Electr. Insul. Mag. **30**(3), 39–49 (2014)
9. Huang, Y.-C., Sun, H.-C.: Dissolved gas analysis of mineral oil for power transformer fault diagnosis using fuzzy logic. IEEE Trans. Dielectr. Electr. Insul. **20**(3), 974–981 (2013)
10. Kari, T., Gao, W.: Power transformer fault diagnosis using FCM and improved PCA. J. Eng. **2017**(14), 2605–2608 (2017)
11. Zheng, J., Zhao, C.: Enhanced canonical variate analysis with slow feature for dynamic process status analytics. J. Process Control **95**, 10–31 (2020)
12. Jiang, B., Braatz, R.D.: Fault detection of process correlation structure using canonical variate analysis-based correlation features. J. Process Control **58**, 131–138 (2017)
13. Samuel, R.T., Cao, Y.: Kernel canonical variate analysis for nonlinear dynamic process monitoring. IFAC-PapersOnLine **48**(8), 605–610 (2015)
14. Odiowei, P.-E.P., Cao, Y.: Nonlinear dynamic process monitoring using canonical variate analysis and Kernel density estimations. IEEE Trans. Industr. Inform. **6**(1), 36–45 (2009)
15. Widodo, A., Yang, B.-S.: Support vector machine in machine condition monitoring and fault diagnosis. Mech. Syst. Signal Process. **21**(6), 2560–2574 (2007)
16. Zhang, X., Chen, W., Wang, B., Chen, X.: Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization. Neurocomputing **167**, 260–279 (2015)

# Research on Blockchain-Based Mobile Edge Computing System in Smart City

MingCi Hai(✉), Cheng Yang, Jun Liu, and XinQian Huang

Research Institute of China Telecom Corporation Limited, Beijing 100081, China
haimc@chinatelecom.cn

**Abstract.** Massive data produced by a growing number of smart devices accessed to the communication network gives a great transmission burden to the network, when implementing smart city service. Along with the security problem, Mobile Edge Computing(MEC) and Blockchain techniques are applied to relieve these problems. This paper proposes a MEC system based on Blockchain, and models the resource allocation problem taking into account the mining time and mining reward. The modeled problem is optimized using Asynchronous Advanced Actor-Critic(A3C) algorithm and Sychronous Advanced Actor-Critic(SA2C) algrorithm. The simulation results show the effectiveness of the methods and compare the efficiency between two algorithms.

**Keywords:** smart city · mobile edge computing · blockchain · resource allocation

## 1 Introduction

The Smart City utilizes of the Internet of Things(IoT), Cloud Computing(CC) and other new technologies of Information Technology(IT) to promote the intelligence of urban planning, construction, management and service [1,2]. However, in smart city service, with the bloom in the number of sensor devices producing massive data, the traditional cloud computing will generate great communication burden to the transmission network and huge processing delay [3].

With the support by Mobile Edge Computing(MEC) in smart city, where the position of service providers for storage and computing is set on the edge of the network near to the smart devices, so that it can relieve the burden and reduce the time delay. However, privacy leakage and other problems [4] still exist in the process of traditional MEC architecture. In order to improve the robustness of MEC system, blockchain technology is used to solve some security problems [5].

Nonetheless, The resource allocation problem in the blockchain-based mobile edge computing system will be more complicated, because the system needs to allocate certain computing resources to solve the mining task in blockchain. Some researchers have carried out works related to the resource allocation problems within blockchain-based MEC systems. An artificial intelligence method was applied by Alia Asheralieva in [6] to cope with mining task based on MEC. But the resource allocation tasks were omitted. He Y. et al. [7] designed a MEC


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 147–154, 2023.
https://doi.org/10.1007/978-981-99-2362-5_19

system supported by blockchain, however regardless of mining time and mining reward. While Qiu X.Y. et al. [8] considered mining reward, the items in reward were just simply added with same weights.

The paper studies the resource allocation problem of mobile edge computing system architecture based on blockchain in smart city. A blockchain-based Mobile Edge Computing service system is designed, which can maintain high efficiency of block generation and safety. The Asynchronous Advanced Actor-Critic(A3C) algorithm in smart contract is proposed to optimize the resource allocation policies of service nodes. Meanwhile the delay, cost and mining reward are taken into consideration, setting different weightings to adjust the service nodes' tendencies of service policy. The efficiency between A3C algorithm and its improved version, Synchronous Advanced Actor-Critic(A2C) algorithm is compared.

The rest of the article is organized as following: In Part 2, a detailed design of blockchain-based Mobile Edge Computing System is proposed with the procedure. In Part 3, the resource allocation problem is modeled as an optimization problem. Two different reinforcement learning models that are based on A3C and SA2C algorithms are utilized to solve the problem. In Part 4, the performance of algorithms are simulated by evaluations and investigated. In Part 5, a conclusion is presented and future research work directions are given.

## 2 Blockchain-Based Edge Computing Service Model

### 2.1 Overall System Architecture

The proposed system is composed of two components: Client layer and Service layer, what we can see in Fig. 1.

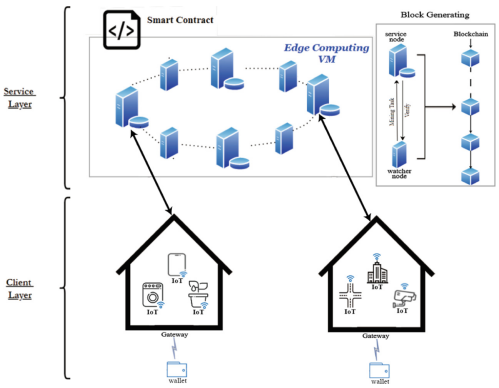



**Fig. 1.** Overall system architecture.

The client layer consists of many smart devices, such as smart vehicles, smart buildings and street lamps. Each device must connect to the blockchain wallet.

Then it sends a trade request to edge computing system. The service layer mainly consists of service nodes and watcher nodes, they all behave as blockchain nodes to stick up for the normal operation of the whole system. Service node takes charge of dealing with the data and generation of blocks. Meanwhile, a watcher node is a specific server which is in charge of monitoring the operation of service nodes. All the nodes are managed through smart contract.

### 2.2 Smart Devices and Nodes

A smart device is the beginning of a certain operation. A client accesses the edge computing system through smart devices. Data processing cost by servers is calculated in GHz.h.

As for the client, there are two parameters that are related to a certain servicing process.

$$t_{delay} = t_q + t_p \tag{1}$$

$$cost = x \text{ GHz.h} \times u \tag{2}$$

where $t_{delay}$ means the time waiting for data processing, $t_q$ means the time queuing, $t_p$ means the time latency of dealing with the data for a server, *cost* means the fee paid by the client, $x$ means CPU cycles in GHz.h that the server consumes, $u$ means the cost for every GHz.h.

The service node provides computing resource and maintain the ledger. Service nodes connect with each other through blockchain to make data in ledger synchronized.

In the designed system, smart contracts mainly helps allocate the computing resource and upload the trade onto blockchain. After clients subscribing data to the service node, the smart contract would allocate the computing resource according to the current status of each node. Its allocation policy would be discussed in later section.

### 2.3 System Procedure

1. The client starts a trade with a service node, providing data and public key of the wallet.
2. After data being uploaded to the service node, the smart contract embedded in the service node conducts it, guiding the node to allocate computing resources based on the policy trained by specific algorithm.
3. The service node processes data and calculates the cost.
4. The service node sends back the processed data to the smart device. The client needs to submit queries about the result and pay the fee.
5. The service node finishes the mining task assigned by watcher node. After that, service node will get its mining reward and a new block is generated by the service node and the watcher node.

## 3 Resource Allocation Problem

### 3.1 Edge Computing System Model

Supposed that $m$ servers compose one service node (SN), they serve $n$ smart devices (SD) all together. Those two are presented by two sets: $\mathbf{SN} = \{E_1, E_2, ..., E_m\}$, $\mathbf{SD} = \{I_1, I_2, ..., I_n\}$.

Different tasks definitely have different requirements in parameters. Those parameters partly determine the latency of tasks. A model to the latency of single task has been set up. Now, it can be extended to a multi-task model. If delay represents the whole-time delay of the system, we have $t_{delay}$ with the same form of (1).

For a certain task, assuming its data size is $d$ bits, an edge computing server needs $M$ CPU cycles to deal with 1 bit data. The frequency of CPU is $f$ GHz.

Moreover, the service time of all the tasks in the target's queue is mentioned, regarding it as the latency of the target task. Moreover, we consider the time of dealing with the tasks in the target's queue as the latency of the target task.

Based on those above, equation (1) can be rewrote as:

$$t_{delay} = \frac{\sum_i d_i \cdot M_i}{f} + d\frac{M}{f} = \frac{b}{f} + d\frac{M}{f} \tag{3}$$

where the first and second terms are respectivly $t_q$ and $t_p$ and we use $b$ to simply denote $\sum_i d_i \cdot M_i$ meaning the total CPU cycles needed by the task in the buffer.

Assuming the data size $d$ follows Gaussian distribution, $d \sim \mathcal{N}(\mu, \sigma^2)$, the arrival rate is $\lambda$. so the cost could be expressed as:

$$cost = h \cdot u, \ h = f \cdot t_{delay} \tag{4}$$

### 3.2 Blockchain Model

The mining process should be considered in the blockchain model when dealing with resource allocation problem, as the process will bring mining reward to service nodes and cause mining delay.

The mining process is the procedure of dealing with a hash problem. The service node is ought to use the address of its last block as the initial string, adding extra string to do hash operation. The watcher node would set a proper degree of difficulty $D$, which means the first $D$ bits of hash operation is zero.

The service node would get rewarded after mining. If one service node takes time $T(s)$ to do one hash operation, the average time of success with $n$ operations is:

$$\begin{aligned} T_n^* =& TP_1 + 2T(1-P_1)P_1 + ... + nT(1-P_1)^{n-1}P_1 \\ =& TP_1 \sum_{i=1}^{n} i(1-P_1)^{i-1} \end{aligned} \tag{5}$$

where the success rate after $n$ times hash collision is $P_n = 1 - \left(1 - \frac{1}{16^D}\right)^n$ [9].

The mining reward (*bonus*) can be calculated through $t_{mining}$ and $\mathbb{E}[T_n^*]$:

$$bonus = \begin{cases} 0.5S, & t_{mining} < \mathbb{E}[T_n^*] \\ 0.5S \times e^{-(t_{mining} - \mathbb{E}[T_n^*])}, & t_{mining} \geq \mathbb{E}[T_n^*] \end{cases} \tag{6}$$

where $\mathbb{E}[T_n^*] = \lim_{n \to \infty} T\left(\frac{1-(1-P_1)^n}{P_1} - \frac{n(1-P_1)^n}{P_1}\right)$, which is the expectation of mining time of service nodes.

Then, assuming the service node accomplishes certification through $n$ times hash operations. Each hash operation takes $h$ cycles, which leads to $t_{mining} = \frac{n \cdot h}{f}$.

### 3.3 Reinforcement Learning Model

In this part, we mainly describe how the service node applies distributed learning through reinforcement learning model.

Two distributed reinforcement learning algorithms are employed: A3C algorithm and SA2C algorithm, to enable the nodes to deal with complex cases. In two algorithms, each worker is actually one service node, but there exists difference between these two algorithms that the workers in A3C algorithm update parameters respectively and the ones in SA2C algorithm do that synchronously.

The server's state has to be recorded for the process from queuing to processing contained in the simulation. Now we use set $\mathbf{S(t)}$ to represent the state at time t:

$$\begin{aligned} \mathbf{S(t)} = \{&d_1(t), M_1(t); d_2(t), M_2(t); ...; d_n(t), M_n(t); \\ &b_1(t), b_2(t), ..., b_m(t); D\} \end{aligned} \tag{7}$$

where $d_j$ denotes data size of device $j$ and $M_j$ means the GPU cycles the task's one bit of device $j$ needs. $b_i$ represents the buffer size in server $i$. Given that there is a upper bound number $n$ of devices accessed to a server at the same time, we can simulate the procedure of the access and exit.

The action is determined by $\mathbf{S(t)}$, i.e. arriving tasks and the server's buffer size. In terms of one service node consisting of $m$ servers, the first step is to separate arriving data into $m$ parts. Moreover, because of the simulating limitation, it will choose two servers to finish the mining task. We express the action as:

$$\begin{aligned} \mathbf{A} = \{a_1, a_2, ..., a_n; c_1, c_2, ..., c_m\}, a_i = \{e_{i1}, e_{i2}, ...e_{im}\}, \\ \sum_{j=1}^{m} e_{ij} = d_i, \sum_{k=1}^{m} c_k = 2_{\{c_k = 0 \ or \ 1\}} \end{aligned} \tag{8}$$

where $a_i$ presents dividing the task for device $i$ and $c_k$ is either 0 or 1. When $c_k = 1$, server $k$ is the miner, else server $k$ is not the miner. $e_{ij}$ means the task of device $i$ that is assigned to server $j$.

We suppose the arrival rate of device $I_j$ is subject to Poisson distribution with parameter $\lambda_j$ so the arrival rate of a service node is $\lambda = \sum_{j=1}^{n} \lambda_j$. $\mathbf{S(t)}$ is updated when task arrives. And the time interval $\tau$ between adjacent states obeys negative exponential distribution with parameter $\lambda$.

When the task arrives at the service node, the possibility that the task belongs to device $I_j$ is $P_{I_j} = \frac{\lambda_j}{\lambda}$.

The reward is composed of three parts: total delay for processing the task, cost paid by clients and the profit of mining. As the reinforcement learning aims to maximize the accumulated reward, we design this reward to lower the latency, bridle the cost and guarantee the profit of mining.

The latency $t^k_{delay}$ (the superscript $k$ denotes the task $k$) is made up of three items: $t_q$, $t_p$ and $t_{mining}$. For having seperated each task into several servers, parameter $t^{ki}_q$ and parameter $t^{ki}_p$ are introduced to indicate waiting delay and latency of dealing with the task in server $i$.

Therefore, we can express $t^k_{delay}$ as:

$$\begin{aligned} t^k_{delay} &= \max_i [t^{ki}_q + t^{ki}_p] + t^k_{mining} \\ &= \max_i [\frac{b_i}{f_i} + e_{ki}\frac{M_k}{f_i}] + t^k_{mining} \end{aligned} \tag{9}$$

Then we calculate the cost of task $k$ as:

$$cost^k = \sum_i h_i \cdot u_i + t^k_{mining} \cdot \sum_{i=1}^{m} (c_i f_i u_i), \quad h_i = e_{ki} \cdot M_k \tag{10}$$

where $h_i$ denotes the resource consumed in server $i$, $u_i$ denotes the resource price of server $i$.

According to analysis, the mining reward would decrease as values of time and cost get bigger. Hence the reward function is given as following:

$$R^k = -\alpha t^k_{delay} - \beta cost^k + \gamma bonus^k \tag{11}$$

where $\alpha, \beta, \gamma$ are weight coefficients of latency, cost and mining reward, respectively.

A3C is a new design of Actor-Critic(AC) algorithm, converging faster, performing more stable [10,11].

*1. Actor-Critic algorithm*

The Critic part is a neural network with parameter $\boldsymbol{w}$ which takes the current state as input and outputs a scalar $V(s;\boldsymbol{w})$ to approximate the value function $V(s)$. When updating $\boldsymbol{w}$, Gradient Descent is applied to update $\boldsymbol{w}$ and the loss function is defined using the multi-step temporal difference(TD) error.

And the Actor part parameterized with $\boldsymbol{\theta}$ is a neural network also taking the current state as input but generating the probability distribition of actions as output to approximate the policy function $\pi(a|s)$.

*2. Asynchronous algorithm*

In order to increase the learning rate of an agent, multiple agents are used to interact with various environments. In our proposed system, the workers can duplicate global network and they interact with independent environments parallelly. After they have explored the environments for a period, some workers calculate their gradients of loss functions respectively based on their own

parameters and loss functions and then they upload their gradients to update the parameters of the global network asynchronously. Finally the workers waiting for local parameter updating duplicate the updated global network and go on with the process mentioned above until the global model converges.

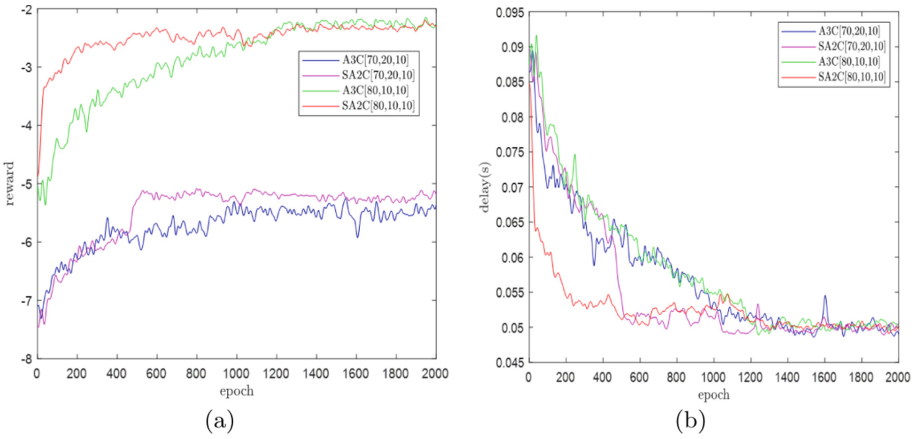
*3. SA2C model*

SA2C algorithm has an evident difference to A3C algorithm: different workers use one same policy at any time, they also update the parameters at the same time [12]. Namely, SA2C has only two networks, one is in charge of updating parameters of global network, the other interacts with environments to gather information. By leveraging multiple environments in parallel, the SA2C model can gather multiple sets of independent experiences for decoupling.

## 4 Simulation and Analysis

In this section, we will simulate the model mentioned above. There are four servers with a computing capability $f_i$ respectively in one service node. They serve five smart devices at the same time.

The simulation is given with an arrival rate of $140s^{-1}$. After testing several groups of coefficients, we set weight coefficients as [70, 20, 10] and [80, 10, 10] (Fig. 2).




(a) (b)

**Fig. 2.** The comparison between A3C and SA2C. (a) Average reward per epoch; (b) Average $t_{delay}$ per epoch.

Obviously, SA2C not only has better astringency, also better policy selection than A3C algorithm. It is more likely to be a well-trained team which explores the environment and decides the best policy by all the members in it. SA2C algorithm can effectively find the policy that makes reward convergence the fastest. Also, it mitigates the possibility to get trapped in local optimum.

## 5 Conclusion

In this article, a Mobile Edge Computing System based on Blockchain is proposed. The system architecture and procedure are described step by step. After that, the resource allocation problem based on this system is modeled, which mainly includes latency, cost and mining reward. Distributed reinforcement learning algorithm is used to solve the problem of resource allocation, which can get a more stable and accurate policy. Finally, the simulation results show the efficiency of the policy acquired by reinforcement learning algorithm. Due to the limitation of computing resource, the large-scale server and user environment can not be simulated, neither have tested the performance of entire system that deploys nodes in the real world environment.

## References

1. Chamoso, P., González-Briones, A., Rodríguez, S., Corchado, J.M.: Tendencies of technologies and platforms in smart cities: a state-of-the-art review. Wirel. Commun. Mob. Comput. **2018**(1), 1–17 (2018)
2. Harrison, C., Eckman, B., Hamilton, R., Hartswick, P., Williams, P.: Foundations for smarter cities. IBM J. Res. Dev. **54**(4), 1–16 (2010)
3. Jan, M.A., Yeh, K.H., Tan, Z., Wu, Y.: Blockchain for edge-enabled smart cities applications. J. Inf. Secur. Appl. **61**, 102937 (2021)
4. Xu, S., Ratazzi, E.P., Du, W.: Security Architecture for Federated Mobile Cloud Computing. Springer, Mobile Cloud Security (2016)
5. Corchado Rodríguez, J.: Artificial intelligence, blockchain and edge computing for smart cities and smart grids (11 2018)
6. Asheralieva, A., Niyato, D.: Bayesian reinforcement learning and bayesian deep learning for blockchains with mobile edge computing. IEEE Trans. Cogn. Commun. Netw. **7**(1), 319–335 (2020)
7. He, Y., Wang, Y., Qiu, C., Lin, Q., Li, J., Ming, Z.: Blockchain-based edge computing resource allocation in IoT: a deep reinforcement learning approach. IEEE Internet Things J. **8**(4), 2226–2237 (2020)
8. Qiu, X., Liu, L., Chen, W., Hong, Z., Zheng, Z.: Online deep reinforcement learning for computation offloading in blockchain-empowered mobile edge computing. IEEE Trans. Veh. Technol. **68**(8), 8050–8062 (2019)
9. Yun, J., Goh, Y., Chung, J.M.: Analysis of mining performance based on mathmatical approach of pow. In: 2019 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1–2. IEEE (2019)
10. Mnih, V., et al.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937. PMLR (2016)
11. Ajay Rao, P., Navaneesh Kumar, B., Cadabam, S., Praveena, T.: Distributed deep reinforcement learning using tensorFlow. In: 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp. 171–174 (2017)
12. Wu, Y., Mansimov, E., Grosse, R.B., Liao, S., Ba, J.: Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. Adv. Neural Inf. Process. Syst. **30**, 5279–5288 (2017)

# Graph Learning-Based Cooperative Spectrum Sensing in Cognitive Radio Networks with Incomplete RSS Measurements

Tao Jiang, Ming Jin(✉), and Juan Liu

Ningbo University, Ningbo 315211, China
{2011082045,jinming,liujuan1}@nbu.edu.cn

**Abstract.** Due to device failure and limited energy reserves, receive signal strength (RSS) measurements in some sensing time slots are missing, which leads to the degraded performance of cooperative spectrum sensing (CSS). To address this issue, we propose a robust graph learning-based CSS (RGL-CSS) framework for a spectrum-heterogeneous cognitive radio network in this work. Specifically, we first formulate a robust graph learning problem to capture the correlation among SUs and solve it by alternating direction method of multipliers (ADMM). Then, a robust graph signal recovery problem is formulated to recover RSSs based on the learned correlation. Finally, the SUs with high correlation perform CSS with the recovered RSSs. Numerical results demonstrate the superiority of the proposed scheme compared to the state-of-the-art schemes.

**Keywords:** Cooperative spectrum sensing · Graph learning · Signal recovery · Incomplete RSS measurements

## 1 Introduction

Spectrum sensing is a critical functionality in cognitive radio (CR) to probe spectrum holes. Traditional spectrum sensing focuses on improving the detection accuracy of the activity of PU under spectrum-homogeneous environments[1]. For non-cooperative spectrum sensing schemes, an SU makes a decision individually. To improve the performance, Many cooperative spectrum sensing schemes integrate the RSS measurements from multiple SUs to make a decision about the presence of primary signal and perform well [1].

However, SUs are located at the spectrum-heterogeneous environments[2] in many realistic scenarios actually [2]. The CSS-based schemes designed for the

[1] In spectrum-homogeneous environments, all SUs are located within the coverage of the PU and have the same access opportunities, which usually correspond to the small-scale cognitive networks, and SUs only have access opportunities when the primary signal is absent at the licensed channel.

[2] In spectrum-heterogeneous environments, SUs at different locations have different access opportunities, some SUs located outside the coverage of the PU could access even primary signal is present because the interference to PU is acceptable, which usually correspond to the large-scale cognitive networks.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 155–164, 2023.
https://doi.org/10.1007/978-981-99-2362-5_20

spectrum-homogeneous environment can detect the activity of PU accurately, but this "correct detection" prevents all SUs from accessing the licensed channel if the primary signal is present for the spectrum-heterogeneous environment. Therefore, a new CSS-based scheme needs to be designed in the spectrum-heterogeneous environment. Significantly, we focus on the selection of SUs participating in cooperation in this paper.

Otherwise, on the one hand, SUs need to listen to primary signals and then process them. On the other hand, SUs need to report processed information to the fusion center (FC), resulting in colossal energy consumption. However, the energy reserves of SUs are limited. Thus, SUs can not always work due to power depletion. Besides, SU may break down at times. Thus some RSS measurements collected by the FC are missing.

### 1.1 Related Work

Some CSS schemes concentrate on the selection of SUs participating in collaboration. In [3], a multi-agent reinforcement learning-based cooperative SUs selection approach was proposed. However, in [3], the spectrum-homogeneous environment is considered. As for the spectrum-heterogeneous environment, in [4], by assuming high correlations of RSSs exists among neighbouring SUs and several continuous sensing slots, a joint spatial and temporal correlation based CSS scheme is proposed . However, the assumption that high correlation of RSSs exist among neighboring SUs may not always hold. To address this issue, a three stage sensing framework was proposed in [5]. However, [4,5] both assume the RSS measurements received at FC are complete. In addition, these methods require the knowledge of the locations of SUs. Besides, studies about CSS under incomplete RSS measurements were proposed [6,7]. A CP-decomposition-based approach was proposed in [6]. In [7], a matrix completion (MC) technique was applied in CSS to recover RSS measurements. However, the CSS-based schemes [6,7] are performed over the spectrum-homogeneous environment.

### 1.2 Motivations and Contributions

As analyzed above, it is necessary to consider the spectrum-heterogeneous environment for probing spectrum holes more comprehensively and realize the performance degradation due to data loss. Interestingly, graph learning is a promising means for capturing the affinity relationships among entities from multiple observed vectors, which has attracted considerable attention and has been applied in many fields, e.g., scholarly citation networks, traffic networks, biological networks.

Motivated by the findings above, we propose a robust graph learning-based cooperative spectrum sensing (RGL-CSS) scheme. Firstly, by exploiting the low-rank property of the matrix constructed with average RSSs, a robust graph learning problem is formulated to capture the correlations of average RSSs of SUs from historical incomplete RSS measurements. After constructing a graph reflecting the correlation among SUs correctly, we design a robust average RSS

recovery optimization problem with the learned graph before data fusion. Then, the SUs with high correlations are collected to perform CSS. It is noted that the RSSs participating in CSS have been recovered to achieve better performance. In addition, as a data-driven approach, the locations of SUs are not required. Numerical simulations are provided to demonstrate the superior performance of the proposed scheme, by comparing to the state-of-the-art schemes.

### 1.3 Notations

Throughout the paper, lowercase normal (e.g., $x$ and $y$), lowercase bold (e.g., $\mathbf{x}$ and $\mathbf{y}$) and uppercase bold (e.g., $\mathbf{X}$ and $\mathbf{Y}$) letters denote scalars, vectors and matrices. $\mathbb{E}[\cdot]$ denotes the expectation operator, $()_{\mathcal{M}}$ denotes the observed subset $\mathcal{M}$. $\|\cdot\|_*$ denotes the nuclear norm, $||\cdot||_F$ denotes the Frobenius norm, Tr() denotes the trace operator, and $\langle\cdot,\cdot\rangle$ represents the inner product operator, $(\cdot)^{\mathrm{H}}$ is the conjugate transpose operator.

## 2 Signal Model

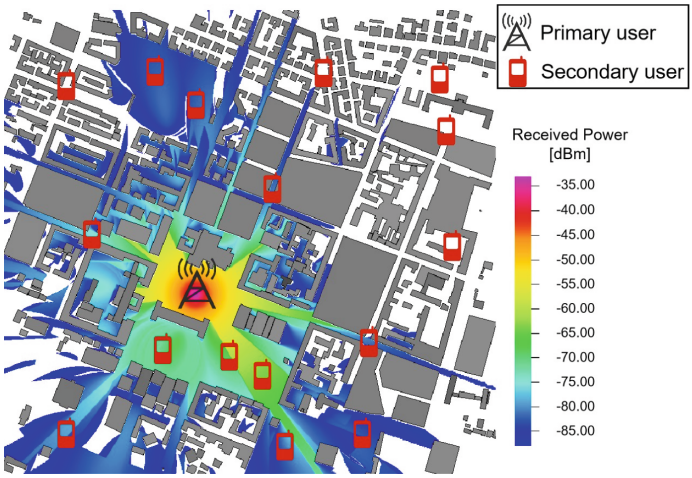



**Fig. 1.** A sketch map example of the coverage of a PU channel.

Assume one PU with $K$ licensed channels and $N$ SUs are distributed in a given area. A sketch map example of the coverage of a PU channel is shown in Fig. 1. Let $\mathcal{H}_1^{(k)}$ and $\mathcal{H}_0^{(k)}$ represent the presence and absence of the PU signal at $k$th channel separately, and denote the regions inside and outside the coverage of the $k$th channel of PU as $\mathcal{S}_1^{(k)}$ and $\mathcal{S}_0^{(k)}$. If the SU is inside the coverage of

the $k$th channel of PU[3] and the PU signal is present, i.e., $\mathcal{O}_1^{(k)} = \mathcal{H}_1^{(k)} \cap \mathcal{S}_1^{(k)}$, there is no spectrum hole can be utilized for a given SU. Otherwise, there is a spectrum hole exists, i.e., $\mathcal{O}_0^{(k)} = \mathcal{H}_0^{(k)} \cup \mathcal{S}_0^{(k)}$. It is noted that the boundary of the coverage is irregular due to shadow fading, especially in electromagnetic propagation environments of urban. Further, the detection of spectrum holes can be modeled as a composite hypothesis testing problem as

$$\begin{aligned} \mathcal{O}_0^{(k)} : r_{k,i}(l) &= \begin{cases} w_{k,i}(l), \\ h_{k,i}\sqrt{g_{k,i}}s_{k,i}(l) + w_{k,i}(l), \mathrm{SU}_i \in \mathcal{S}_0^{(k)} \end{cases} \\ \mathcal{O}_1^{(k)} : r_{k,i}(l) &= h_{k,i}\sqrt{g_{k,i}}s_{k,i}(l) + w_{k,i}(l), \mathrm{SU}_i \in \mathcal{S}_1^{(k)} \end{aligned} \tag{1}$$

where $r_{k,i}(l)$ denotes the $l$th received signal sample at the $i$th SU from the $k$th channel with $i = 1, 2, ..., N$ and $l = 0, 1, ..., L-1$; $h_{k,i}$ denotes the channel coefficient due to small-scale fading and $g_{k,i}$ denotes the power attenuation due to large-scale fading; $s_{k,i}(l)$ denotes the primary signal, without loss of generality, it is assumed that $\{s_{k,i}(l), \forall l\}$ are i.i.d with zero mean and variance $\sigma_s^2$, and $w_{k,i}(l)$ denotes the additive independent and identically distributed (i.i.d.) circularly symmetric complex white Gaussian noise with zero mean and variance $\sigma_w^2$, i.e., $w_{k,i}(l) \sim \mathcal{CN}(0, \sigma_w^2)$; In addition, we assume that $h_{k,i}$ follows the standard complex Gaussian distribution, i.e., $h_{k,i} \sim \mathcal{CN}(0, 1)$, and it keeps unchanged within each sensing time slot but varies independently from one to another.

The complete RSS measurement $x_{k,i}$ of the $i$th SU from the $k$th channel at a sensing time slot is given by

$$x_{k,i} = \frac{1}{L}\sum_{l=0}^{L-1} |r_{k,i}(l)|^2. \tag{2}$$

According to the central limit theorem, with a sufficiently large $L$, $x_{k,i}$ is approximately Gaussian distributed as

$$x_{k,i} \sim \begin{cases} \mathcal{N}\left(\sigma_w^2, \frac{1}{L}\sigma_w^4\right), & \mathcal{H}_0^{(k)} \\ \mathcal{N}\left(\sigma_w^2\left(1+\gamma_{k,i}\right), \frac{1}{L}\sigma_w^4\left(1+\gamma_{k,i}\right)^2\right), & \mathcal{H}_1^{(k)} \end{cases} \tag{3}$$

where the signal-to-noise ratio (SNR) $\gamma_{k,i}$ is defined by $g_{k,i}|h_{k,i}|^2\sigma_s^2/\sigma_w^2$. Moreover, due to limited energy reserves and device failures, SUs can not send RSS measurements to the FC in some sensing time slots. Therefore, the RSSs are missing in some sensing time slots of some SUs. Finally, the received incomplete RSS measurements $x_{k,i}^{ec}$ at the FC can be denoted as

$$x_{k,i}^{ec} = \begin{cases} \text{null}, & \text{if } x_{k,i} \text{ is missing.} \\ x_{k,i}, & \text{otherwise.} \end{cases} \tag{4}$$

[3] The coverage of PU can be understood as the protect range for the PU that depends on the interference tolerance of the PU et.al.

## 3 Graph Learning-Based Cooperative Spectrum Sensing

A cognitive radio network can be naturally associated with the structure of graph. SUs on the network can correspond to nodes on the graph, and the RSS measurements at SUs are regarded as the graph signals on the nodes. In this section, a robust graph learning-based CSS framework is introduced.

### 3.1 Graph Learning

Assume that the matrix for complete RSS measurements is given by

$$\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, ..., \hat{\mathbf{x}}_N] \in \mathbb{R}^{KQ \times N}. \tag{5}$$

where

$$\hat{\mathbf{x}}_i = \left[x_{1,i}^{(1)}, ..., x_{1,i}^{(Q)}, ..., x_{K,i}^{(1)}, ..., x_{K,i}^{(Q)}\right]^{\mathrm{T}} \in \mathbb{R}^{KQ \times 1} \tag{6}$$

where $\hat{\mathbf{x}}_i$ denotes the complete RSS measurements from $K$ channels and $Q$ historical sensing slots at each SU. Considering complete RSSs $x_{k,i}$ still affected by a small perturbation, we decompose $\hat{\mathbf{X}}$ further, let $\mathbf{P} \in \mathbb{R}^{KQ \times N}$ be the expectation of $\hat{\mathbf{X}}$, i.e., $\mathbf{P} = \mathbb{E}[\hat{\mathbf{X}}]$, . it can be obtained that the rows in $\mathbf{P}$ are either $[\sigma_w^2, \cdots, \sigma_w^2]$ or $\sigma_w^2 + [g_{k,1}\sigma_s^2, \cdots, g_{k,N}\sigma_s^2]$ for the $k$th channel. Hence, the rank of $\mathbf{P}$ is not large than $K + 1$.

Let $\mathbf{N}$ be the small perturbation of $\hat{\mathbf{X}}$ whose elements follow the Gaussian distribution with mean zero and take the RSS measurements loss into consider, The observed set of $\hat{\mathbf{X}}$ can be written as

$$\hat{\mathbf{X}}_{\mathcal{M}} = (\mathbf{P} + \mathbf{N})_{\mathcal{M}} \tag{7}$$

which indicates that $\hat{\mathbf{X}}$ contains a low-rank matrix and a Gaussian random matrix of mean zero.

Let $s_{i,j}$ denotes the probability that the $j$th SU is the neighbor of the $i$th SU. The sum of the probabilities between the $i$th SU and its all neighbors is 1, i.e., $\mathbf{S1} = \mathbf{1}$, $\mathbf{1}$ is an all-one column vector. The graph Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - (\mathbf{S} + \mathbf{S}^{\mathrm{T}})/2$, where $\mathbf{D}$ is a diagonal matrix with the $i$th diagonal elements being $d_{i,i} = \sum_j (s_{i,j} + s_{j,i})/2$. To obtain an estimate of the probability matrix $\mathbf{S}$ from incomplete RSS measurements, introduce an auxiliary variable $\mathbf{Z} = \mathbf{P}$ and a residual matrix $\mathbf{R}$, we formulate a robust graph learning problem as

$$\begin{aligned} \min_{\mathbf{P},\mathbf{N},\mathbf{S},\mathbf{Z}} & \|\mathbf{P}\|_* + \alpha||\mathbf{N}||_F^2 + \beta \mathrm{Tr}(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}) + \gamma||\mathbf{S}||_F^2 + \mathcal{I}(\mathbf{R}_{\mathcal{M}}) \\ \text{s.t. } & \hat{\mathbf{X}} = \mathbf{P} + \mathbf{N} + \mathbf{R}, \mathbf{S1} = \mathbf{1},\ 0 \le s_{i,j} \le 1, \forall i, j, \mathbf{Z} = \mathbf{P}. \end{aligned} \tag{8}$$

where $\mathcal{I}$ is a indicator operator as

$$\mathcal{I}(\mathbf{R}_{\mathcal{M}})_{i,j} = \begin{cases} 0, & if\ (\mathbf{R}_{\mathcal{M}})_{i,j} = 0 \\ +\infty, & \text{otherwise} \end{cases} \tag{9}$$

where $\alpha$, $\beta$, $\gamma$ are hyperparameters. and $\|\mathbf{P}\|_*$ is the convex approximation the rank of $\mathbf{P}$ and minimize this term corresponds to the assumption that $\mathbf{P}$ is a low rank matrix, and minimize $||\mathbf{N}||_F^2$ enforces $\mathbf{N}$ is a small perturbation, the term $\mathrm{Tr}\left(\mathbf{PLP}^{\mathrm{T}}\right)$ is the graph Laplacian regularizer and minimize this term means assign weights between SUs according to the similarity of data of SUs, and $||\mathbf{S}||_F^2$ is a penalty. It means that the correlations captured is based on average RSS matrix $\mathbf{P}$.

By incorporating the first two constraints, the augmented Lagrange function of (8) is given by

$$\begin{aligned}\mathcal{L}(\mathbf{P},\mathbf{N},\mathbf{S},\mathbf{Z},\mathbf{R},\mathbf{Y}_1,\mathbf{Y}_2) = \|\mathbf{P}\|_* + \alpha\|\mathbf{N}\|_F^2 + \beta\mathrm{Tr}\left(\mathbf{ZLZ}^{\mathrm{T}}\right) + \gamma\|\mathbf{S}\|_F^2 + \mathcal{I}\left(\mathbf{R}_{\mathcal{M}}\right) \\ + \left\langle \mathbf{Y}_1, \mathbf{P}+\mathbf{N}+\mathbf{R}-\hat{\mathbf{X}}\right\rangle + \langle \mathbf{Y}_2, \mathbf{P}-\mathbf{Z}\rangle \\ + \frac{\rho_1}{2}\left\|\mathbf{P}+\mathbf{N}+\mathbf{R}-\hat{\mathbf{X}}\right\|_F^2 + \frac{\rho_2}{2}\|\mathbf{P}-\mathbf{Z}\|_F^2 \\ \text{s.t.}\quad \mathbf{S1}=\mathbf{1},\ 0\le s_{i,j}\le 1, \forall i,j.\end{aligned} \tag{10}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ denote the Lagrange multiplier matrices, $\rho_1$ and $\rho_2$ are penalty coefficients. Problem (10) can be broken into several subproblems which optimize $\mathbf{P}$, $\mathbf{N}$, $\mathbf{S}$, $\mathbf{Z}$, $\mathbf{R}$, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ alternately. Then ADMM can be employed to solve these subproblems. The solutions to the subproblems at the $(m+1)$th iteration are given as follows beginsmall

$$\mathbf{P}^{(m+1)} = \mathcal{D}_{\frac{1}{\rho_1^{(m)}+\rho_2^{(m)}}}\left(\mathbf{A}^{(m)}\right) \tag{11a}$$

$$\mathbf{N}^{(m+1)} = \frac{1}{2\alpha+\rho_1^{(m)}}\left[\rho_1^{(m)}\left(\hat{\mathbf{X}}-\mathbf{P}^{(m+1)}-\mathbf{R}^{(m)}\right)-\mathbf{Y}_1^{(m)}\right] \tag{11b}$$

$$s_{ij}^{(m+1)} = \frac{\hat{d}_{i,\varrho+1}^{(m)} - \hat{d}_{ij}^{(m)}}{\varrho\hat{d}_{i,\varrho+1}^{(m)} - \sum_{j=1}^{\varrho}\hat{d}_{ij}^{(m)}} \tag{11c}$$

$$\mathbf{Z}^{(m+1)} = \left(\rho_2^{(m)}\mathbf{P}^{(m+1)} + \mathbf{Y}_2^{(m)}\right)\left(2\beta\mathbf{L}^{(m+1)} + \rho_2^{(m)}\mathbf{I}\right)^{-1} \tag{11d}$$

$$\mathbf{R}_{\mathcal{M}}^{(m+1)} = 0 \tag{11e}$$

$$\mathbf{R}_{\mathcal{U}}^{(m+1)} = \left(\hat{\mathbf{X}}-\mathbf{P}^{(m+1)}-\mathbf{N}^{(m+1)}-\left(\mathbf{Y}_1^{(m)}\right)/\rho_1^{(m)}\right)_{\mathcal{U}} \tag{11f}$$

$$\mathbf{Y}_1^{(m+1)} = \mathbf{Y}_1^{(m)} + \rho_1^{(m)}\left(\mathbf{P}^{(m+1)}+\mathbf{N}^{(m+1)}+\mathbf{R}^{(m+1)}-\hat{\mathbf{X}}\right) \tag{11g}$$

$$\mathbf{Y}_2^{(m+1)} = \mathbf{Y}_2^{(m)} + \rho_2^{(m)}\left(\mathbf{P}^{(m+1)}-\mathbf{Z}^{(m+1)}\right) \tag{11h}$$

$$\rho_1^{(m+1)} = \tau\rho_1^{(m)} \tag{11i}$$

$$\rho_2^{(m+1)} = \tau\rho_2^{(m)} \tag{11j}$$

where

$$\mathbf{A}^{(m)} = \frac{1}{\rho_1^{(m)} + \rho_2^{(m)}} \rho_1^{(m)} \left[\hat{\mathbf{X}} - \left(\mathbf{N}^{(m)} + \mathbf{R}^{(m)}\right)\right] + \rho_2^{(m)} \mathbf{Z}^{(m)} - \left(\mathbf{Y}_1^{(m)} + \mathbf{Y}_2^{(m)}\right) \tag{12}$$

$\mathcal{D}_\theta(\cdot)$ denotes singular value shrinkage operator, $\mathcal{S}_\theta(\cdot)$ denotes soft thresholding operator, $d_{i,j}^{(m)} = \|\mathbf{z}_i^{(m)} - \mathbf{z}_j^{(m)}\|_2^2$ with $\mathbf{z}_i^{(m)} \in \mathbb{R}^{KQ\times 1}$ being the $i$th column of $\mathbf{Z}$ and $\mathbf{d}_i^{(m)} = \left[d_{i1}^{(m)}, d_{i2}^{(m)}, \ldots, d_{iN}^{(m)}\right]^{\mathrm{T}} \in \mathbb{R}^N$, sort the elements in $\mathbf{d}_i$ in ascending order and denoted as $\hat{\mathbf{d}}_i$, $\varrho$ denotes the number of nonzero elements in $\mathbf{s}_i$.

### 3.2 Average RSS Recovery with Learned $\hat{\mathbf{S}}$

The incomplete RSS measurements of $k$th channel of all SUs at sensing time slot to be tested $(\mathbf{x}_k)_{\mathcal{M}}$ can be denoted as

$$(\mathbf{x}_k)_{\mathcal{M}} = (\mathbf{f}_k + \mathbf{n}_k)_{\mathcal{M}} \tag{13}$$

where $\mathbf{x}_k = [x_{k,1}, x_{k,2}, ..., x_{k,N}]^{\mathrm{T}}$ denotes the complete RSS measurement vector, $\mathbf{f}_k \in \mathbb{R}^N$ denotes the average RSS form all the SUs on $k$th channel, $\mathbf{n}_k \in \mathbb{R}^N$ denotes small perturbation. Introducing a residual vector $\mathbf{c}_k \in \mathbb{R}^N$ to capture the error introduced by the unobserved part of $\mathbf{x}_k$, we formulate the following optimization problem

$$\begin{aligned} \min_{\mathbf{f}_k, \mathbf{n}_k} \quad & \left\| \mathbf{f}_k - \frac{1}{|\lambda_{\max}|} \hat{\mathbf{S}} \mathbf{f}_k \right\|_2^2 + \zeta \|\mathbf{n}_k\|_2^2 + \mathcal{I}\left((\mathbf{c}_k)_{\mathcal{M}}\right) \\ \text{s.t.} \quad & \mathbf{x}_k = \mathbf{f}_k + \mathbf{n}_k + \mathbf{c}_k \end{aligned} \tag{14}$$

where $\lambda_{\max}$ denotes the largest eigenvalue of $\hat{\mathbf{S}}$. The augmented Lagrange function can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{f}_k, \mathbf{n}_k, \mathbf{c}_k, \mathbf{y}) = & S_2(\mathbf{f}_k) + \zeta \|\mathbf{n}_k\|_2^2 + \mathcal{I}\left((\mathbf{c}_k)_{\mathcal{M}}\right) \\ & + \langle \mathbf{y}, \mathbf{f}_k + \mathbf{n}_k + \mathbf{c}_k - \mathbf{x}_k^{ec} \rangle + \frac{\rho_3}{2} \|\mathbf{f}_k + \mathbf{n}_k + \mathbf{c}_k - \mathbf{x}_k\|_2^2 \end{aligned} \tag{15}$$

where $\zeta$ are hyper-parameters, $\mathbf{y}$ is lagrange multiplier vector, $\rho_3$ is penalty coefficient, the ADMM is employed to update $\mathbf{f}_k$, $\mathbf{n}_k$, $\mathbf{c}_k$ and $\mathbf{y}$ alternately. The solutions to the subproblems at the $(m+1)th$ iteration are given as follows

$$\mathbf{f}_k^{(m+1)} = \left(2\bar{\mathbf{S}} + \rho_3^{(m)} \mathbf{I}\right)^{-1} \left[\rho_3^{(m)} \left(\mathbf{x}_k - \mathbf{n}_k^{(m)} - \mathbf{c}_k^{(m)}\right) - \mathbf{y}^{(m)}\right] \tag{16a}$$

$$\mathbf{n}^{(m+1)} = \frac{1}{2\zeta + \rho_3^{(m)}} \left[\rho_3^{(m)} \left(\mathbf{x}_k - \mathbf{f}_k^{(m+1)} - \mathbf{c}_k^{(m)}\right) - \mathbf{y}^{(m)}\right] \tag{16b}$$

$$\left(\mathbf{c}_k^{(m+1)}\right)_{\mathcal{M}} = 0 \tag{16c}$$

$$\left(\mathbf{c}_k^{(m+1)}\right)_{\mathcal{U}} = \left(\mathbf{x}_k - \mathbf{f}_k^{(m+1)} - \mathbf{n}_k^{(m+1)} - \mathbf{y}^{(m)}/\rho_3^{(m)}\right)_{\mathcal{U}} \tag{16d}$$

$$\mathbf{y}^{(m+1)} = \mathbf{y}^{(m)} + \rho_3^{(m)} \left(\mathbf{f}_k^{(m+1)} + \mathbf{n}_k^{(m+1)} + \mathbf{c}_k^{(m+1)} - \mathbf{x}_k\right) \tag{16e}$$

where $\bar{\mathbf{S}} = \left(\mathbf{I} - \hat{\mathbf{S}}/|\lambda_{max}|\right)^{\mathrm{H}} \left(\mathbf{I} - \hat{\mathbf{S}}/|\lambda_{max}|\right)$.

### 3.3 Cooperative Spectrum Sensing

After obtaining an estimate of the probability matrix $\hat{\mathbf{S}}$ and recovered RSS $\hat{\mathbf{f}}_k$, we can obtain the neighbor set of each SU. Let $\mathbf{v}_{\mathcal{N}_i}$ be the neighbor set of the $i$th SU. The elements in $\mathbf{v}_{\mathcal{N}_i}$ are the indices of the corresponding SUs. The recovered RSSs from the SUs in $\mathbf{v}_{\mathcal{N}_i}$ and the one from the $i$th SU are collected to implement cooperative spectrum sensing. The test-statistic of the detector for probing the spectrum hole of $k$th channel at the $i$th SU is given by

$$T_{k,i} = \sum_{j \in \mathbf{v}_{\mathcal{N}_i} \bigcup \{i\}} \frac{w_{i,j}}{w_i} \hat{f}_{k,i} \tag{17}$$

where $w_{i,j} = (\hat{s}_{i,j} + \hat{s}_{j,i})/2$ denotes the weight for the $j$th SU with $j \in \mathbf{v}_{\mathcal{N}_i}$, $\hat{s}_{i,j}$ is the $(i,j)$th element in $\hat{\mathbf{S}}$. $w_{i,i} = (\sum_{j \in \mathbf{v}_{\mathcal{N}_i}} w_{i,j})/\mathcal{N}_i$, where $\mathcal{N}_i$ denotes the cardinality of the set $\mathbf{v}_{\mathcal{N}_i}$. $w_i = \sum_{j \in \mathbf{v}_{\mathcal{N}_i} \bigcup \{i\}} w_{i,j}$ is the normalization factor of the weights. With a decision threshold $\lambda_k$, the SU makes decision on that a spectrum hole is available when $T_{k,i} < \lambda_k$; Otherwise, no spectrum hole is available.

## 4 Numerical Result

In the simulation, the average RSSs of SUs are generated by domain path model at a realistic block of Munich with the software WinProp and the size is $1421m \times 1345m$. A PU has $K = 10$ channels with carrier frequencies from 2412 MHz MHz to 2457 MHz MHz. The noise power is set to –75dBm. The boundary of the coverage of the PU channel is set to –12dB and the sensing sensitivity of the SUs is set to –108 dBm. we set $\alpha = 100$, $\beta = 1$, $\rho_1^{(0)} = \rho_2^{(0)} = \rho_3^{(0)} = 1$ and $\zeta = 0.2$. Additionally, if the test statistics at $\mathrm{SU}_i$ is missing, the sensing result of $\mathrm{SU}_i$ comes from random guess. Here, the average detection probability $P_d^{(k)}$ and the average false-alarm probability of the $k$th channel $P_f^{(k)}$ are denoted as

$$P_d^{(k)} = \frac{1}{|\mathbb{N}_{\text{in}}^{(k)}|} \sum_{i \in \mathbb{N}_{\text{in}}^{(k)}} P_{d,i}^{(k)} \qquad P_f^{(k)} = \frac{1}{N} \sum_i P_{f,i}^{(k)} \tag{18}$$

where $P_{f,i}^{(k)} = P(\tilde{\mathcal{O}}_{1,i}^{(k)} | \mathcal{O}_{0,i}^{(k)})$, and $P_{d,i}^{(k)} = P(\tilde{\mathcal{O}}_{1,i}^{(k)} | \mathcal{O}_{1,i}^{(k)})$, $\mathbb{N}_{\text{in}}^{(k)}$ denotes the cardinality of the SU sets inside the coverage of the PU on the $k$th channel. we evaluate the performance of the proposed RGL-CSS scheme by comparing it to three the state-of-art schemes: (1) SN-SS: The single node spectrum sensing scheme. (2) U-CSS: The universal cooperative sensing scheme where the measurements from all SUs are fused. (3) SC-CSS: The spatial correlation cooperative sensing scheme [4] where the neighboring SUs are fused.

We first evaluate the performance of various schemes with different missing rate. Figure 2 shows the receiver operating characteristic (ROC) curves of various schemes on all PU channels with 20%, 30%, 40% RSS measurements lost when the active probability of the PU is 0.7 and the density of the SUs is

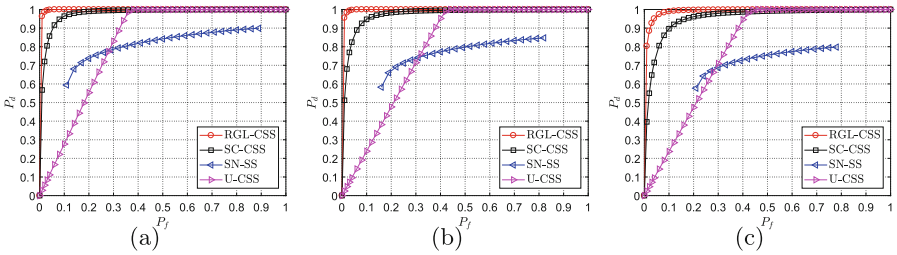



**Fig. 2.** Receiver operating characteristic (ROC) curves when the active probability of the PU is $p_1 = 0.7$ and the missing rate of RSS measurements is (a) 20%, (b) 30%, (c) 40%.

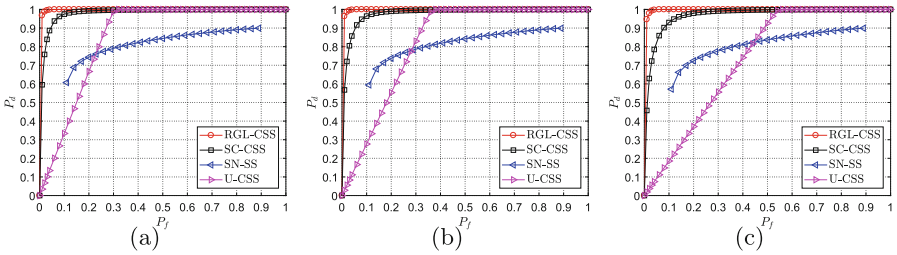



**Fig. 3.** Receiver operating characteristic (ROC) curves when the missing rate of the RSS measurements is 20% and the active probability of the PU is (a) 0.5, (b) 0.7, (c) 0.9.

$1.75 * 10^{-3}/m^2$, It's can be observed that our RGL-CSS scheme outperforms the other schemes at all missing rate tested, because it not only has superior temporal detection, but also reduces the spatial false-alarm probability in general. Specifically, the performance of RGL-CSS is relatively stable when the loss rate is less than 30%, and the $P_d$ of RGL-CSS scheme is 13%, 13.6% and 19.6% higher than SC-CSS scheme when the missing rate are 20%, 30% and 40% separately when $P_f = 0.05$. the performance of U-CSS degrades due to the high false alarm probability of SUs outside the coverage of PU channel, and the $P_d$ of SN-SS degrades with the increase of missing rate.

Then we change the active probability of the PU in Fig. 3 when the density of SU is $1.75 * 10^{-3}/m^2$, the performance of our RGL-CSS scheme achieves the best performance and slightly degrades when the active probability of PU increases, the performance of other schemes, especially the U-CSS scheme, degrades severely when the active probability of the PU increases.

By fixing the missing rate at 20% and active probability of the PU at 0.7, we evaluate the performance of different schemes when the density of SU are $0.375 * 10^{-3}/m^2$, $0.75 * 10^{-3}/m^2$, $1.125 * 10^{-3}/m^2$ and $1.75 * 10^{-3}/m^2$ in Fig. 4 separately, the proposed RGL-CSS still achieves the best performance compared with other detectors.

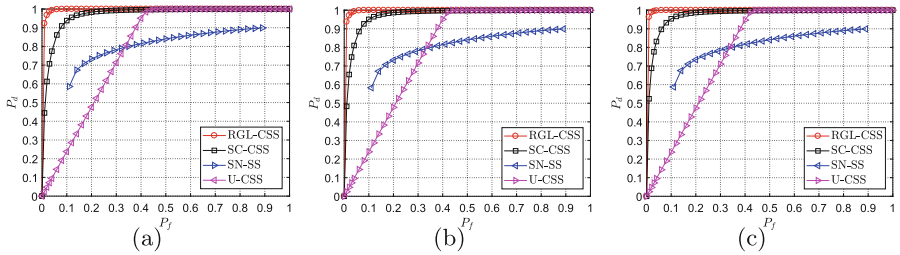



**Fig. 4.** Receiver operating characteristic (ROC) curves when the active probability of PU is $p_1 = 0.7$ and the density of SUs is (a) $0.75 \times 10^{-3}/m^2$, (b) $1.125 \times 10^{-3}/m^2$, (c) $1.75 \times 10^{-3}/m^2$.

## 5 Conclusion

We propose a robust graph learning based cooperative sensing scheme for probing spectrum holes in the spectrum-heterogeneous environment with incomplete RSS measurements. By combining the low-rank property of average RSS matrix, a graph that reflect the correlation of SUs is acquired by solving a robust graph learning problem. Then, the RSSs are recovered by solving a graph-based robust signal recovery problem. Numerical results have shown that the proposed RGL-CSS scheme delivers superior performance by exploiting the correlation of SUs and recovered RSSs. Even the missing rate of RSS measurements is 40%, the $P_d$ of the RGL-CSS scheme is 19.6% higher than SC-CSS when $P_f = 0.05$.

## References

1. Sarikhani, R., Keynia, F.: Cooperative spectrum sensing meets machine learning: deep reinforcement learning approach. IEEE Commun. Lett. **24**(7), 1459–1462 (2020)
2. Tandra, R., Sahai, A., Veeravalli, V.: Unified space-time metrics to evaluate spectrum sensing. IEEE Commun. Mag. **49**(3), 54–61 (2011)
3. Gao, A., Du, C., Ng, S.X., Liang, W.: A cooperative spectrum sensing with multi-agent reinforcement learning approach in cognitive radio networks. IEEE Commun. Lett. **25**(8), 2604–2608 (2021)
4. Wu, Q., Ding, G., Wang, J., Yao, Y.: Spatial-temporal opportunity detection for spectrum-heterogeneous cognitive radio networks: two-dimensional sensing. IEEE Trans. Wirel. Commun. **12**(2), 516–526 (2013)
5. Zhang, Z., Wen, X., Xu, H., Yuan, L.: Sensing nodes selective fusion scheme of spectrum sensing in spectrum-heterogeneous cognitive wireless sensor networks. IEEE Sensors J. **18**(1), 436–445 (2018)
6. Wang, H., Fang, J., Duan, H., Li, H.: Compressive wideband spectrum sensing and signal recovery with unknown multipath channels. IEEE Trans. Wirel. Commun. **21**(7), 5305–5316 (2022)
7. Qin, Z., Gao, Y., Plumbley, M.D.: Malicious user detection based on low-rank matrix completion in wideband spectrum sensing. IEEE Trans. Signal Process. **66**(1), 5–17 (2018)

# Research and Implementation of O-OFDM Transmission System Based on FPGA

Du Wu[1,2], Yupeng Li[1,2(✉)], Qianqian Li[1,2], Xiaoming Ding[1,2], and Xiaocheng Wang[1,2]

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
fx_lyp@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** Orthogonal frequency division multiplexing (OFDM) can effectively eliminate the inter symbol interference (ISI) generated by the signal in the multi-path propagation. Using this feature, design an optical OFDM (O-OFDM) transmission system based on the field programmable gate array (FPGA), and give the system scheme and key technology implementation method. In the design, in order to effectively reduce the inter channel interference (ICI) and inter symbol interference, a cyclic prefix is introduced to avoid the influence of related interference through periodic convolution. At the same time, the training sequence (TS) is introduced in the design process for the relevant synchronous design and equalization processing, to reduce the complexity of the baseband system design.

**Keywords:** OFDM · FPGA · ICI · ISI · Cyclic prefix · TS

## 1 Introduction

OFDM is a modulation technique that can also be seen as a multiplexing technique. It is a special multi-carrier transmission scheme that is well able to combat frequency selective fading and narrowband interference. Its basic principle is to decompose a data stream transmitted in parallel into many low-rate sub-number streams and transmit them in parallel over multiple sub-channels. In 2005, N.E. Jolley and T.M. Tang et al. presented the first application of OFDM technology to fiber optic transmission systems at the OFC2005 conference and experimentally demonstrated that 10 Gb/s signals can be transmitted in multimode fiber transmission [1]. Since then, the research of optical orthogonal frequency division multiplexing technology has been started. O-OFDM has the advantages of high spectral efficiency, strong resistance to dispersion and outstanding effect of nonlinearity. Therefore, using the OFDM technology can effectively improve the anti-dispersion ability of optical fiber transmission system [2].


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 165–171, 2023.
https://doi.org/10.1007/978-981-99-2362-5_21

O-OFDM systems can be divided into two categories according to the different detection methods at the receiver: direct detection optical OFDM systems (DDO-OFDM) and coherent detection optical OFDM systems (CO-OFDM) [3]. Among them, CO-OFDM has higher spectrum utilization and good receiver sensitivity, while DDO-OFDM system has simple structure, low power consumption and low cost. Considering the complexity of CO-OFDM structure, the structure used in this paper is DDO-OFDM system, and the design ideas of DDO-OFDM transmission system and FPGA implementation of baseband signal modulation and demodulation are elaborated.

## 2 The Design of DDO-OFDM Transmission System

A simple DDO-OFDM transmission system implementation is shown in Fig. 1. At the transmitter, the data to be transmitted is processed by FPGA to form OFDM data frames, however, the complex data output from IFFT to achieve optical modulation requires digital upconversion (DUC) processing to get the real number signal, and the digital high frequency signal is output without changing the information of the signal. Finally, the analog signal obtained after digital-to-analog conversion (DAC) is fed to a Mach-Zendel modulator (MZM) for modulation to obtain the optical signal transmitted in the fiber. At the receiver, the radio frequency signal is obtained after photoelectric detection (PD), and then the original binary bit stream is restored according to the inverse process at the transmitter.

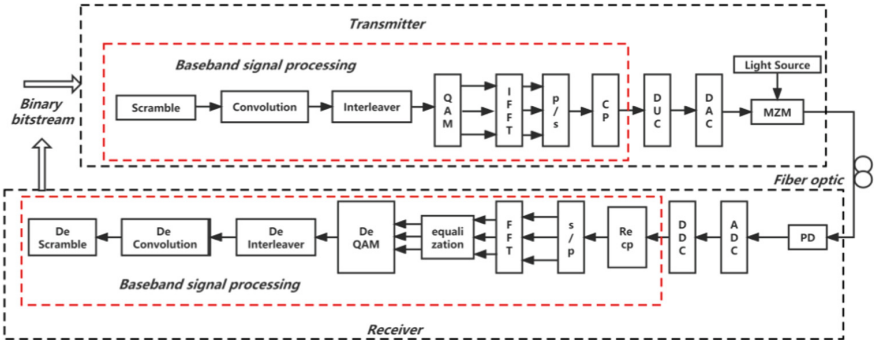



**Fig. 1.** DDO-OFDM transmission system block diagram

### 2.1 Processing of OFDM Baseband Signals

The modulation and demodulation of OFDM baseband signals is mainly implemented using FPGA (Fig. 1, red box). The data to be sent first passes through a scrambler to avoid long "0" or "1" that could affect the establishment and maintenance of synchronization at the receiver. Then the scrambled data is fed to the convolutional encoder with constraint length 7 for encoding to improve the reliability of information code element transmission. In this paper, 16QAM modulation is used, but the coded data need to be interleaved before QAM mapping to ensure that adjacent coded bits are mapped to non-adjacent subcarriers

[4, 5]. The symbol modulation is performed using the IFFT principle [6], and finally a cyclic prefix is added to form the OFDM data frame; The demodulation of baseband signal needs synchronization and equilibrium processing. First, the baseband signal after DDC is symbol synchronized and the cyclic prefix is removed. After FFI operation, the time domain or frequency domain response of the output channel should be estimated to ensure the accuracy of demodulation.

The OFDM data frame formed at the transmitter is preceded by two training sequences, which include a short training sequence (STS) of 10 cycles (16 samples per cycle) and a long training sequence (LST) of two cycles (64 samples per cycle), while the last 32 samples of the LST are copied to the front of the LST to form a cyclic prefix (Fig. 2). STS is mainly responsible for data detection, symbol timing synchronization, automatic gain control, etc. The LTS is mainly responsible for the frequency bias estimation and the channel estimation [7].

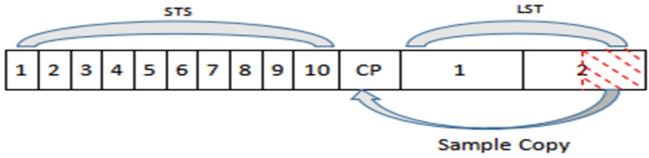



**Fig. 2.** Schematic diagram of training sequence

### 2.2 Symbol Synchronization

The intercorrelation operation of the data can be performed at the receiver using locally known short training symbols to determine the end point of the short training symbols or the start point of the long training symbols and to remove the cyclic prefixes. The module implementation can obtain the interrelation number $C_k$ by multiplying the received data groupings with the conjugate complexes of locally known short training symbols and summing them [8].

$$C_k = \sum_{m=0}^{D-1} r_{k-m} \times S_m^* \tag{1}$$

In the above equation, the superscript * indicates the conjugate value. D is the length of the Correlation coefficient, whose size determines the performance of the symbolic synchronization algorithm. Larger values improve performance but also increase computational effort. For IEEE 802.11a systems, $D = 16$ is the cycle length of short training symbols. When $C_k$ has a peak, this point serves as the end point of a short training sequence. Using this feature, you can find the end point of all short training symbols in an OFDM grouping.

### 2.3 Channel Estimation and Equalization

The purpose of channel estimation and equalization is to estimate the time or frequency domain response of the channel, correct and recover the received data, and obtain the performance gain of coherent detection (3 dB) [9]. The maximum multipath delay of the channel, the movement speed of the user, and the noise of the receiver are important factors that affect the performance of the channel estimation algorithm. The design can be done using two long training symbols for simple and efficient estimation of the frequency response of all subsequent subcarrier channels in that frame [10].

The frequency domain channel estimation and equalization method based on long training sequence symbols is as follows.

$$R_{RLTS} = (R_{1LTs} + R_{2LTS})/2 \tag{2}$$

$$\hat{H} = R_{RLTS}/L_{LTS} \tag{3}$$

$$\hat{R} = R/\hat{H} \tag{4}$$

where $R_{1LTS}$ and $R_{2LTS}$ are the first and second long training symbols received; $L_{LTS}$ is the standard long training symbol, $R_{RLTS}$ is the average of the two training symbols; $\hat{H}$ is the estimated value of the channel frequency response; R is the OFDM symbol received; $\hat{R}$ is the OFDM symbol after the channel equalization.

## 3 Simulation Results and Analysis

The modulation and demodulation of baseband signals are mainly implemented using IFFT/FFT algorithms [11]. The IFFT/FFT function can be implemented by using the IP core provided by Xilinx. The main parameters of the IP core are shown in Table 1, where the IP core performs IFFT operation when FWD_INV is 0 and FFT operation when FWD_INV is 1. From the joint simulation results of IFFT/FFT (Fig. 3), it can be seen that the input data of IFFT and the output data of FFT are basically the same, and a certain error is allowed here because the demodulation of the receiver dynamically selects the bit information according to the size of the judgment threshold.

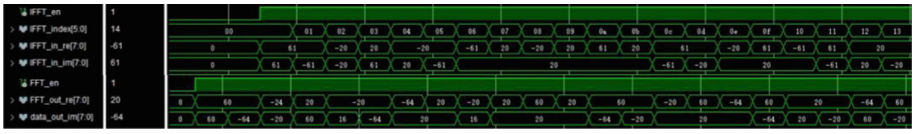



**Fig. 3.** The IFFT /FFT joint simulation diagram

In the baseband signal processing at the transmitter side, the training sequence is required to be transmitted before the modulated OFDM data frame and form a continuous sequence to be transmitted [12]. As shown in Fig. 4, four identical m-sequences (data_in) are input, and after coding, mapping and IFFT processing, a cyclic prefix is inserted to

**Table 1.** Main parameter settings for the FFT IP core

| Parameter | Value/Option |
|---|---|
| Implementation | Pipelined, Streaming I/O |
| Input Data With | 16 |
| Phase Factor Width | 16 |
| Transform Length | 64 |
| Scaling Option | Scaled |
| SCALE_SCH | 10110 |
| FWD_INV | IFFT(0), FFT(1) |

form the final OFDM data frame (cp_outr/cp_outi), and then the data to be transmitted (ofdm_re/ofdm_im) is formed after prepose the training sequence (TS_re/TS_im). In order to verify the correctness of the FPGA implementation of baseband signal modulation and demodulation, the baseband output data at the receiver side is compared with the baseband input number at the transmitter side (Fig. 5), and if the data agree, it shows that the design of OFDM baseband system based on FPGA is feasible.

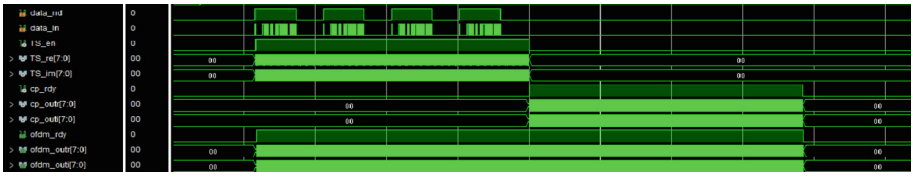



**Fig. 4.** Launch-end baseband simulation diagram

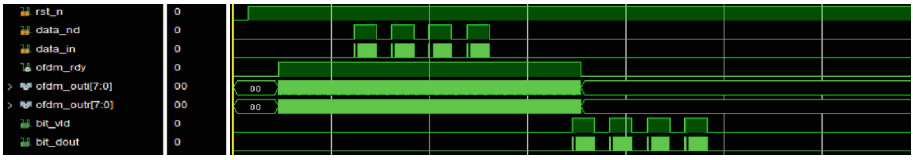



**Fig. 5.** Comprehensive simulation diagram of OFDM baseband system

Finally the paper gives the FPGA chip resource usage for the OFDM baseband (Table 2). Among them, the time-frequency conversion module takes up the most resources, mainly because the FFT ip core consumes a lot of DSP and RAM resources.

**Table 2.** Internal logical resource consumption of FPGA

| Block | LUTs | FFs | Slicess | BRAMs | DSP48s |
|---|---|---|---|---|---|
| ADC interface & Storage | 4343 | 4456 | 2172 | 116 | 0 |
| IFFT & FFT | 14017 | 31578 | 7086 | 100 | 192 |
| Equalizer | 4198 | 9213 | 2008 | 24 | 128 |
| Mapper & Remapper | 519 | 412 | 226 | 0 | 32 |
| Add CP & Remove CP | 2415 | 3823 | 1087 | 0 | 0 |
| Total | 34215 (11.3%) | 55182 (9.1%) | 16578 (21.8%) | 391 (38%) | 352 (13%) |

## 4 Conclusion

The main work of this paper is to design and implement an O-OFDM communication system, and elaborate the design scheme of OFDM baseband system. The channel estimation and equalization algorithm in this paper can accurately estimate the frequency response of the channel, and can effectively correct and recover the received data. The IFFT/FFT operation is implemented using the IP cores provided by Xilinx, which is verified by simulation and can simulate the process of OFDM baseband modulation and demodulation to reduce the design complexity. Finally, the relevant data results are analyzed and the FPGA-based O-OFDM transmission system design is fully feasible.

**Acknowledgment.** This work was supported in part by the Natural Science Foundation of China under Grant 61901301, 62001328, 62001327.

## References

1. Jolley, N.E., Kee, H., Pickard, P.: Generation and propagation of a 1550 nm 10Gbit/s optical orthogonal frequency division multiplexed signal over 1000m of multimode fibre using a directly modulated DFB. In: Optical Fiber Communication Conference, 2005. Technical Digest, pp. 6–11 (2005)
2. Buchali, F., Dischler, R., Liu, X.: Optical OFDM: a promising high-speed optical transport technology. Bell. Labs Tech. **14**, 125–146 (2009)
3. Deng, C.: Research room of real-time processing module in OFDM PON. Beijing University of Posts and Telecommunications (2014)
4. Li, F., Xiao, X., Yu, J.: Realization of real-time 100G 16QAM OFDM signal detection. Proc. SPIE **9388**, 111–121 (2015)
5. Chen, M., He, J., Fan, Q., Dong, Z., Chen, L.: Experimental demonstration of real-time high-level QAM-encoded direct-detection optical OFDM systems. Light. Technol. **33**, 4632–4639 (2015)
6. Bruno, J.S., Almenar, V., Valls, J.: FPGA implementation of a 10 GS/s variable-length FFT for OFDM-based optical communication systems. Microprocesors Microsyst. **64**, 195–204 (2019)

7. Fort, A., Weijers, J.W., Derudder, V., et al.: A performance and complexity comparison of auto-correlation and cross-correlation for OFDM burst synchronization. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. II-341–344 (2013)
8. Laourine, A., Stephenne, A., Affes, S.: A new OFDM synchronization symbol for carrier frequency offset estimation. Sig. Process. Lett IEEE **14**(5), 321–324 (2017)
9. Li, Y.: Pilot symbol aided channel estimation for OFDM in wireless systems. IEEE Trans. Vech. Technol. **49**(04), 1207–1215 (2000)
10. Hu, S.C., Kang, K., Wang, H.F., Qian, H.: Low complexity blind detection in OFDM systems with phase noise. Digit. Sig. Process. **129** (2022)
11. Fernandes, B., Sarmento, H.: FPGA implementation and testing of a 128 FFT for a MB-OFDM receiver. Analog Integr. Circ. Sig. Process **70**(2), 241–248 (2012)
12. Tang, X., Wang, M.: The OFDM time-frequency synchronization algorithm using the train sequence. J. Chongqing Univ. **32**(8), 956–959 (2019)

# Review of Bias Point Stabilization Methods for IQ Modulator

Mingzhu Zhang[1,2], Yupeng Li[1,2(✉)], Xiaoming Ding[1,2], and Xiaocheng Wang[1,2]

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
fx_lyp@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** The main sub-device of In-phase and Quadrature (IQ) modulator is Mach-Zehnder Modulator (MZM) based on $LiNbO_3$ crystals, and the susceptibility of $LiNbO_3$ material to external environment such as temperature and device aging will lead to changes in the intrinsic properties of MZM, which will degrade the quality of optical signal originally biased at a specific voltage. In order to ensure the effective transmission of signals in optical communication systems, this paper summarized the bias point stabilization control method for IQ modulators in optical communication systems.

**Keywords:** IQ modulator · Bias point · Control of stability

## 1 Introduction

Generally speaking, the modulator used to generate higher-order modulated signals is IQ modulator, which consists of two MZMs and one Mach-Zehnder Interferometer (MZI) that is known as phase modulator. The branches of the two MZMs are the In-phase (I) branch and the Quadrature (Q) branch respectively, and both of them work in the push-pull mode. The phase modulator mainly produces a certain phase shift for the optical signal of I/Q channels, which is generally 90° [1, 2].

Using IQ Modulation at the transmitter side, the real and imaginary parts of the amplified RF signal are generally loaded into the I- and Q-way Radio Frequency (RF) ports of the IQ modulator. The basic structure of the IQ modulator is shown in Fig. 1. The signal light enters the IQ modulator input port and is divided into two paths by the coupler into the I/Q two branches, and the two paths will be coupled into one signal at the output for output. And the three modulators are controlled by three bias voltages respectively [3, 4].

The main advantage of IQ modulation is the ability to very easily combine independent signal components into a composite signal, which is later decomposed into independent signal components.IQ modulators can actually be seen as a combination of a phase modulator and an intensity modulator and are widely used in higher order modulation formats such as QAM. Meanwhile, IQ modulator is an important device


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 172–179, 2023.
https://doi.org/10.1007/978-981-99-2362-5_22

choice in ultra-high-speed coherent systems, and stabilizing its bias point is of great significance for communication system stability.

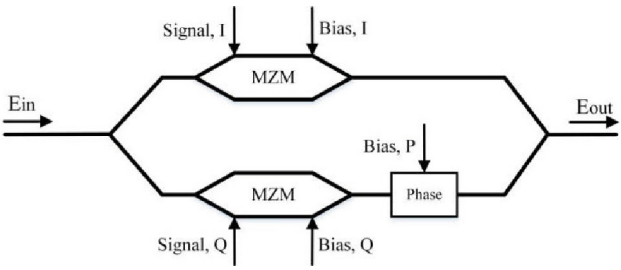



**Fig. 1.** Basic structure of the IQ modulator

## 2 IQ Modulator Bias Point Stabilization Control Method

IQ modulator bias point stabilization control methods can be classified into three main categories according to the essential principles: power monitoring, dither signal and phase monitoring based bias point stabilization control technique.

Figure 2 shows the schematic diagram of the power monitoring based bias point stabilization control technique. The principle of this bias point stabilization control technique is based on finding a special value of the average optical power and RF power spectrum signal. The Coupler separates a small portion of the signal from the IQ light modulation signal into the photodetector, which converts the light signal into an electrical signal. Followed by the electrical signal into the RF detector, then into the analog-to-digital converter, and finally into the control unit. Control unit analyzes the received signal in order to control the three output signals of the digital-to-analog converter, i.e., the three bias voltages of the IQ optical modulator.

The schematic diagram of the dither signal based bias point stabilization control technique is similar to that the schematic diagram of the power-based bias point stabilization control technique, therefore it is no longer drawn. But the difference is that the key to this method is to add the pilot signal to the bias control signal in both I/Q paths, while the phase shift bias voltage contains only the Direct Current (DC) signal. The comparison signal obtained from the clock signal in the oscillator is sent to the bias control unit, where the data is analyzed and the bias point is stabilized [5, 6].

The schematic diagram of the phase monitoring based bias point stabilization control technique is similar to that the schematic diagram of the power-based bias point stabilization control technique, therefore it is no longer drawn. But the difference is that the method incorporates a differential phase detector in the system. The stability of the bias voltage is controlled by monitoring the phase information of the modulating signal. The control unit determines whether the bias voltage drifts according to the difference between the real-time phase information and the ideal phase information, and controls the bias point stability by algorithm controlling the three bias voltages.

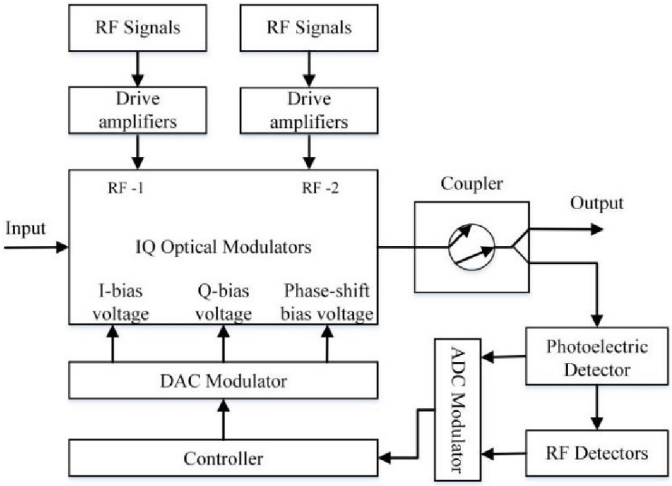



**Fig. 2.** The schematic diagram of the power monitoring based bias point stabilization control technique

### 2.1 Power Monitoring Based Bias Point Stabilization Control Technique

Method 1: The ratio of the first- and second-order bias of the mean power of the output signal is used for direct regulation similar to a look-up table, and the regulation of the bias voltage is subsequently achieved by the corresponding iterative algorithm.

The scheme diagram of ratio of the first- and second-order bias of the mean power of the output signal is shown in Fig. 3. Firstly, for monitoring and regulating the two sub-MZMs, the mean power can be used as a rough monitoring indicator, and then the ratio of the first- and second-order bias of the mean power to the bias can be used as a specific regulation indicator. When the bias point is changed, the corresponding bias ratio can be calculated and then looked up in the table, which saves the corresponding iteration time. After the two MZMs are adjusted, the variance and normalized variance are calculated based on the sampled data, and the phase modulator is iteratively adjusted according to the normalized variance, and the position with the smallest normalized variance is the best bias [7, 8].

Method 2: The bias point stabilization control technique using reverse waves. A wavelength adjustable light source is added at the output of the IQ modulator, and an optical signal is input to the modulator, which propagates in the opposite direction to the modulated optical signal. A power monitoring instrument is placed at the input of the target modulator to minimize the power of the reverse wave by monitoring it, in order to stabilize the bias point at the optimal working position. This scheme increases the power consumption and cost of the modulation system [9, 10]. Figure 4 shows a schematic of the configuration of the monitoring and control scheme.

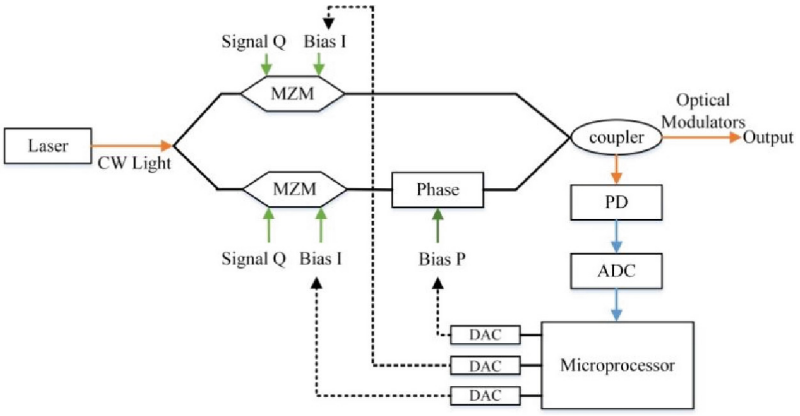



**Fig. 3.** The scheme diagram of ratio of the first- and second-order bias of the mean power of the output signal

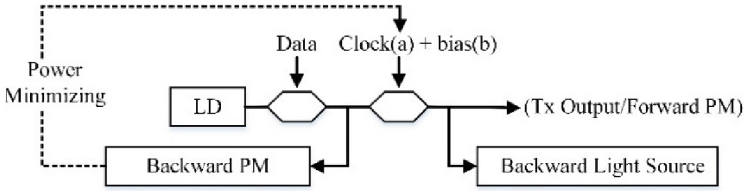



**Fig. 4.** Configuration schematic of the monitoring and control scheme

### 2.2 Dither Signal Based Bias Point Stabilization Control Technique

Method 1: As shown in Fig. 5, the structure of the integrated DP-QPSK modulator is shown, and the lead signals are added to the bias control signals of the I-way, Q-way and the phase shift between the two ways, respectively, setting the frequencies $f_I$, $f_Q$ and $f_P$, so that three beat frequency signals can be obtained with frequencies $f_I - f_Q$, $f_I - f_p$ and $f_Q - f_p$, respectively, and by eliminating these three beat frequency signals, the Mach-Zehnder modulator of the I-way and Q-way can be stabilized [11, 12].

Firstly, Control the outer MZM at Quad by nulling the detected beat signal at frequency of $(f_I - f_Q)$. Secondly Control the inner MZM at Min by nulling the detected beat signal at frequency of $(f_Q - f_p)$. And finally Control the inner MZM at Min by nulling the detected beat signal at frequency of $(f_I - f_p)$.

Mathematically speaking, the magnitudes of the beat terms at the difference frequencies are proportional to the 2nd order partial derivatives of the detected photocurrent with respect to the corresponding bias voltages. Therefore, the three bias conditions derived from the above control technique should satisfy the following system of three equations:

$$\frac{\partial^2 I_{\text{out}}}{\partial v_{I,\text{dc}} \cdot \partial v_{Q,\text{dc}}} = 0 \tag{1}$$

$$\frac{\partial^2 I_{\text{out}}}{\partial v_{I,\text{dc}} \cdot \partial v_P} = 0 \tag{2}$$

$$\frac{\partial^2 I_{\text{out}}}{\partial v_{Q,\text{dc}} \cdot \partial v_P} = 0 \tag{3}$$

In the above equations, $I_{out}$ denotes the current at the output. $V_{I,dc}$, $V_{Q,dc}$ and $V_P$ denote the DC voltages at the I-way MZ modulator, Q-way MZ modulator and phase modulator terminals, respectively.

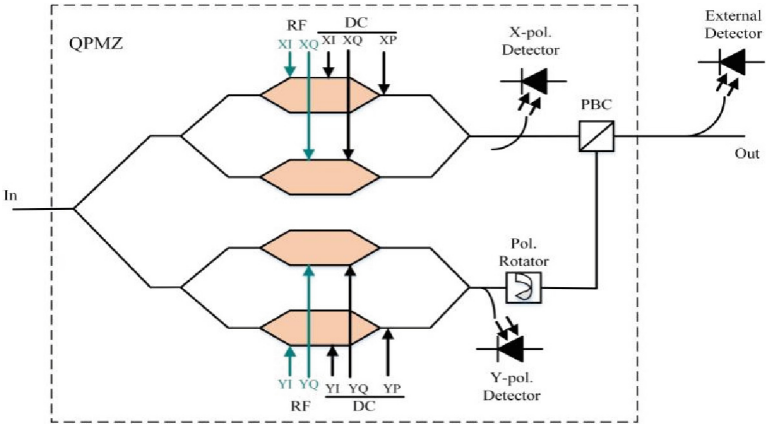



**Fig. 5.** Integrated DP-QPSK modulator structure diagram

Method 2: Figure 6 shows the automatic bias control method based on two different dither signals. The bias signals of the two sub-MZMs (Bias I and Bias Q) are dithered with two sine waves. A microprocessor unit (MPU) is then used to monitor the average optical signal power, the transmitted dither signal, and the correlation integral of the detected dither signal for automatic bias control. The scheme does not require FFT-based spectral analysis and is easy to implement directional bias adjustment. The algorithm has lower complexity and faster convergence [13, 14].

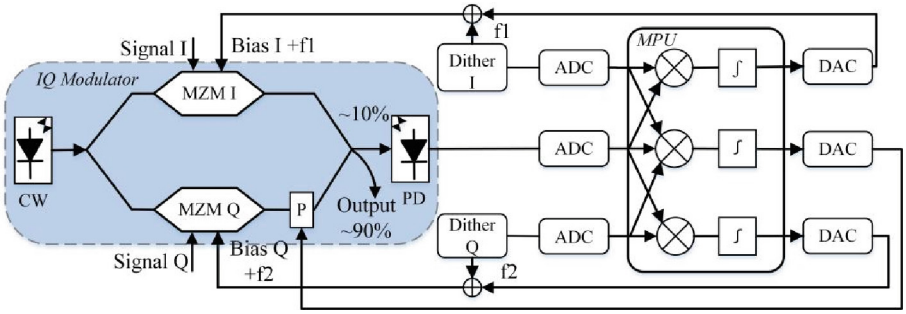



**Fig. 6.** Automatic bias control method based on two different dither signals

Method 3: Stabilization control technique using the ratio of peak to average optical power of modulated optical signal. In I/Q bias control signal to add the guide frequency signal, combined with the modulation coefficient for bias point stability control. The system structure is simple, but the flexibility is not good [15, 16]. Figure 7 shows the configuration of the modulator bias control block. The bias voltages applied to the ports each include a DC bias voltage and a dither signal.

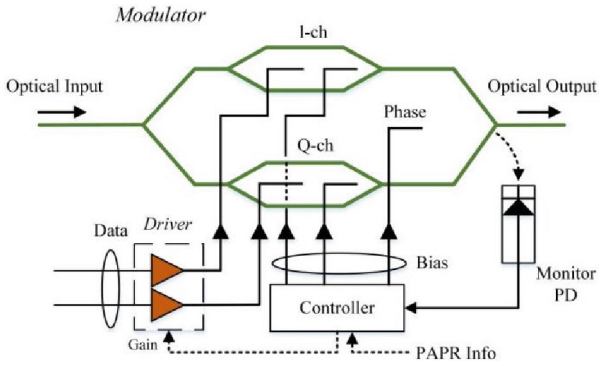



**Fig. 7.** Configuration diagram of the modulator bias control block

### 2.3 Phase Monitoring Based Bias Point Stabilization Control Technique

Figure 8 shows the method of bias point stabilization control based on phase monitoring. The method is not limited by the modulation format and has good flexibility. It is based on the principle that the bias voltage changes after the bias point drift, resulting in the phase of the output optical signal also changes. First, we set all three bias voltages to work at the optimal point to obtain the phase in the ideal situation, and then judge the

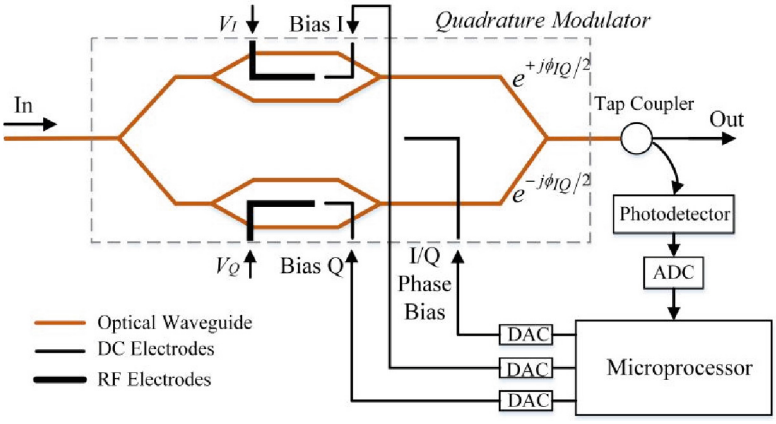



**Fig. 8.** The method of bias point stabilization control based on phase monitoring

drift of the bias point by comparing the real-time phase with the ideal phase, and use the feedback control system to adjust the three bias points [17, 18].

## 3 Conclusion

This paper summarizes the automatic bias point stabilization control techniques for IQ optical modulators. Each of these bias point stabilization control techniques has its own advantages and disadvantages, but their ultimate goal is the same, i.e., to quickly and stably control the bias voltage from the IQ optical modulator at the best possible position. Finally, by summarizing and comparing the advantages and disadvantages of different bias point stability control methods, a theoretical basis for subsequent research is provided.

**Acknowledgment.** This work was supported in part by the Natural Science Foundation of China under Grant 61901301, 62001328, 62001327.

## References

1. Yu, J., Zhang, J.: Recent progress on high-speed optical transmission. Digit. Commun. Netw. **2**(2), 65–76 (2016)
2. Cho, P.S., Khurgin, J.B., Shpantzer, I.: Closed-loop bias control of optical quadrature modulator. IEEE Photonics Technol. Lett. (2006)
3. Peng, W.R., et al.: Compensation for I/Q imbalances and bias deviation of the Mach-Zehnder modulators in direct-detected optical OFDM systems. IEEE Photonics Technol. **21**, 103–105 (2009)
4. Roberts, K., et al.: Performance of dual-polarization QPSK for optical transport systems. IEEE J. Lightw. Technol **16**(27), 3546–3559 (2009)
5. Zhang, W.: The research on technique scheme of IQ modulator auto-bias control. Beijing University of Posts and Telecommunications (2016)
6. Morohashi, I.: 16 QAM synthesis by angular superposition of polarization using dual-polarization QPSK modulator. In: ECOC (2010)
7. Tang, W.: Research on automatic bias control and distortion compensation algorithms for IQ modulation. Huazhong University of Science & Technology (2021)
8. Nelson, L.E., Woodward, S.L.: Capabilities of real-time digital coherent transceivers. In: ECOC (2010)
9. Sekine, K., Hasegawa, C., Kikuchi, N., et al.: A novel bias control technique for MZ modulator with monitoring power of backward light for advanced modulation formats. In: Conference on Optical Fiber Communication and the National Fiber Optic Engineers Conference, OFC/NFOEC 2007, pp. 1–3 (2007)
10. Armstrong, J.: OFDM for optical communications. J. Lightw. Technol. **207**(3), 189–204 (2009)
11. Sotoodeh, M., Beaulieu, Y., Harley, J., et al.: Modulator bias and optical power control of optical complex E-field modulators. J. Lightwave Technol. **29**(15), 2235–2248 (2011)
12. Yamanaka, S.: 11x171Gb/s PDM 16-QAM transmission over 1440 km with a spectral efficiency of 6.4 b/s/Hz using high-speed DAC. In: ECOC (2010)
13. Kawakami, H., Kobayashi, T., Yoshida, E., Miyamoto, Y.: Auto bias control technique for optical 16-QAM transmitter with asymmetric bias dithering (2011)

14. Li, X., Deng, L., Chen, X., Cheng, M., Fu, S., et al.: Modulation-format-free and automatic bias control for optical IQ modulators based on dither-correlation detection (2017)
15. Kawakami, H.: Auto bias control technique for QPSK modulator with asymmetric bias dithering. In: OECC (2010)
16. Yoshida, T., Sugihara, T., Uto, K., et al.: A study on automatic bias control for arbitrary optical signal generation by dual-parallel Mach-Zehnder modulator. In: Optical Communication (ECOC), pp. 1–3 (2010)
17. Kawakami, H., Yoshida, E., Miyamoto, A.Y.: Asymmetric dithering technique for bias condition monitoring in optical QPSK modulator. Electron. Lett. **6**(46), 430–431 (2010)
18. Cho, P.S., Nazarathy, M.: Bias control for optical OFDM transmitters. IEEE Photonics Technol. Lett. (2010)

# Joint 3D Trajectory and Scheduling Design for IRS-Assisted Secure UAV Communication

Xiaoqi Zhai(✉), Juan Liu, Lingfu Xie, and Ming Jin

School of Electrical Engineering and Computer Science, Ningbo University, Zhejiang 315211, China
{2011082155,liujuan1,xielingfu,jinming}@nbu.edu.cn

**Abstract.** In this paper, we study the secure transmission design in an intelligent reflective surface (IRS) assisted unmanned aerial vehicle (UAV) communication system, where the UAV acts as an aerial base station (BS) to serve numerous legitimate users in the presence of an eavesdropper (Eve). In the system, the user scheduling, three-dimensional (3D) trajectory of the UAV, and phase shifts of IRS are jointly optimized to maximize the average secrecy rate subject to diverse data rate requirements of the legitimate users. However, the formulated problem is a mixed integer non-convex optimization problem, which is generally difficult to solve. We propose an iterative algorithm based on alternating optimization (AO) and successive convex approximation (SCA) to obtain the suboptimal solution. Simulation results show that the proposed method can effectively combat eavesdropping while ensuring the quality-of-service (QoS) of legitimate users.

**Keywords:** UAV communication · intelligent reflective surface · average secrecy rate · 3D trajectory optimization

## 1 Introduction

Equipped with communication and computing platforms, unmanned aerial vehicle (UAV) plays a pivotal role in the field of wireless communications thanks to its unique advantages such as flexibility and manipulability [1]. The UAV can fly close to ground terminals and communicate with them via the line-of-sight (LoS) links, thus improving the air-to-ground (A2G) data rates. However, UAV communications over LoS channels are more vulnerable to malicious eavesdropping. In [2], the authors studied a secrecy rate maximization problem by UAV trajectory design and power allocation. In [3], authors summarized several potential schemes for mitigating adversarial eavesdropping in UAV-assisted networks.

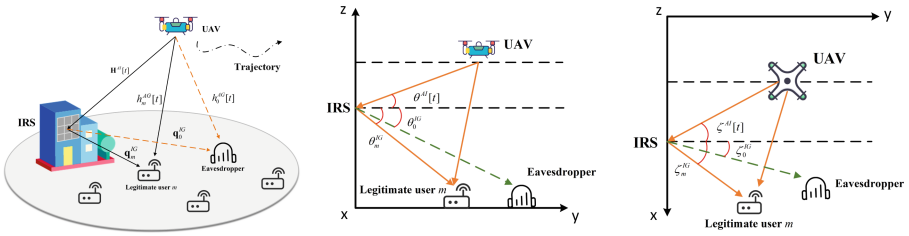
In this paper, we introduce the emerging intelligent reflective surface (IRS) technology to combat eavesdropping. The benefits of leveraging IRS in secure communication have also been presented in many studies. In [4], the author studied a secure IRS-assisted multiple-in/multiple-out (MIMO) wireless communication system where the secrecy rate was maximized by optimizing the


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 180–188, 2023.
https://doi.org/10.1007/978-981-99-2362-5_23

transmit power and phase shifts of IRS. In [5], an IRS-assisted multi-antenna secure communication is proposed to minimize the transmission power under the constraint of each legitimate user's minimum secrecy rate. Recently, many research interests are aroused to study IRS-assisted UAV secure communication. In [6,7], the UAV trajectory, power allocation, and phase shifts of IRS were jointly optimized to achieve the maximum secrecy rate for the single-antenna and multi-antenna UAV systems, respectively. These above works have mainly focused on designing the IRS-assisted UAV secure communication for one single user via two-dimensional (2D) UAV trajectory planning and IRS's phase control.

Motivated by these works, we consider an IRS-assisted secure UAV communication system where a UAV serves legitimate users with diverse data rate requirements in the presence of an eavesdropper (Eve). We aim to maximize the average secrecy rate of legitimate users by optimizing the user scheduling, phase shifts of IRS, and three-dimensional (3D) trajectory of the UAV. We propose an iterative algorithm based on alternating optimization (AO) and successive convex approximation (SCA) to obtain the suboptimal solution efficiently. The simulation results show that the proposed design can effectively enhance average secrecy rate while ensuring the minimum data rate of each legitimate user.

Notation: In this paper, scalars are italicized, vectors and matrices are bold. $\mathbb{R}^{M\times N}$ and $\mathbb{C}^{M\times N}$ denote the real-valued and complex-valued spaces of $M \times N$ dimensions, respectively. For a vector or matrix $\mathbf{a}$, $||\mathbf{a}||$ is its Euclidean norm. $\mathbf{a}^T$ is the transpose and $\mathbf{a}^H$ is the conjugate transpose of $\mathbf{a}$. diag$\{x_1, ..., x_N\}$ represents a diagonal matrix with diagonal elements $\{x_1, ..., x_N\}$.




(a) System model (b) Vertical angle (c) Horizontal angle

**Fig. 1.** IRS-assisted UAV secure network model

## 2 System Model

### 2.1 System Model

We consider a downlink communication system composed of one UAV, one IRS, $M$ legitimate users and an active Eve. We use $\mathbf{z}^A[t] = \left[q[t]^T, H^A[t]\right]$ to denote the 3D trajectory of the UAV in time slot $t \in \{1, \ldots, T\}$, where $\mathbf{q}[t] = \left[x^A[t],\right.$

$y^A[t]\big]^T \in \mathbb{R}^2$ and $H^A[t]$ denote the UAV's horizontal and vertical trajectories, respectively. For ease of exposition, the $M$ legitimate users and Eve are treated as ground users (GUs). The horizontal coordinate of GU $m$ is set as $\mathbf{z}_m^G = \left[x_m^G, y_m^G\right]^T \in \mathbb{R}^2$ $(m \in \{0\} \cup \mathcal{M})$, where $\mathcal{M} = \{1, \ldots, M\}$ is the set of the legitimate users, and Eve is indexed by $m = 0$. An IRS is fixedly deployed on the building with its horizontal location denoted by $\mathbf{z}^I = \left[x^I, y^I\right]^T \in \mathbb{R}^2$ and height $H^I$. Besides, the IRS contains $N = N_x \times N_y$ block reflection elements. As shown in Fig. 1 (a), the UAV communicates with each legitimate user over two links, i.e., the UAV-user direct link and the UAV-IRS-user cascaded link. The period of the UAV flight as $T$, and the length of each time slot is $\tau$ seconds. Assume that the maximum horizontal and vertical speeds within each time slot are $V_h$ and $V_z$, respectively. The phase shifts of IRS for legitimate user $m \in \mathcal{M}$ in time slot $t$ is expressed as $\mathbf{\Phi}_m[t] = \text{diag}\left(e^{j\theta_{m,1,1}[t]}, \ldots, e^{j\theta_{m,n_x,n_y}[t]}, \ldots, e^{j\theta_{m,N_x,N_y}[t]}\right) \in \mathbb{C}^{N\times N}$.

### 2.2 Channel Model

Due to the dominance of LOS channels in the A2G system, the channel from the UAV to GU $m$ in time slot $t$ is expressed as $h_m^{AG}[t] = \sqrt{\frac{\beta_0}{(d_m^{AG}[t])^2}}$, where $\beta_0$ represents the channel gain at the reference distance $d = 1$m and $d_m^{AG}[t]$ is the distance between the UAV and GU $m$ in time slot $t$ expressed as $d_m^{AG}[t] = \sqrt{\|\mathbf{q}[t] - \mathbf{z}_m^G\|^2 + (H^A[t])^2}$. The channel from the UAV to the IRS in time slot $t$ can be modeled as

$$\mathbf{H}^{AI}[t] = \sqrt{\frac{\beta_0}{(d^{AI}[t])^2}} \left[1, \ldots, e^{-j[(N_x-1)\gamma^{AIx}[t]+(N_y-1)\gamma^{AIy}[t]]}\right]^H \in \mathbb{C}^{N\times 1}, \quad (1)$$

where $\gamma^{AIx}[t] = \pi \sin\theta^{AI}[t] \cos\zeta^{AI}[t]$ and $\gamma^{AIy}[t] = \pi \sin\theta^{AI}[t] \sin\zeta^{AI}[t]$ are the phase shifts in x and y dimensions relative to the origin of the IRS with $\sin\theta^{AI}[t] = \frac{\|H^A[t]-H^I\|}{d^{AI}[t]}$, $\sin\zeta^{AI}[t] = \frac{\|x^A[t]-x^I\|}{\|\mathbf{q}[t]-\mathbf{z}^I\|}$ and $\cos\zeta^{AI}[t] = \frac{\|y^A[t]-y^I\|}{\|\mathbf{q}[t]-\mathbf{z}^I\|}$, $d^{AI}[t]$ is the distance between UAV and IRS in time slot $t$ expressed as $d^{AI}[t] = \sqrt{\|\mathbf{q}[t] - \mathbf{z}^I\|^2 + (H^A[t] - H^I)^2}$. As shown in Fig. 1 (b), (c), $\theta^{AI}[t]$ and $\zeta^{AI}[t]$ represent the vertical and horizontal angle-of-arrivals (AoAs) from the UAV to the IRS in time slot $t$, respectively. The channel from the IRS to GU $m \in \mathcal{M}$ in time slot $t$ can be given as

$$\left(\mathbf{q}_m^{IG}\right)^{\mathbf{H}} = \sqrt{\frac{\beta_0}{(d_m^{IG}[t])^\eta}} \left[1, \ldots, e^{-j\left[(N_x-1)\gamma_m^{IGx}+(N_y-1)\gamma_m^{IGy}\right]}\right] k_m^{IG} \in \mathbb{C}^{N\times 1}, \quad (2)$$

where $\gamma_m^{IGx} = \pi \sin\theta_m^{IG} \cos\zeta_m^{IG}$ and $\gamma_m^{IGy} = \pi \sin\theta_m^{IG} \sin\zeta_m^{IG}$. $\theta_m^{IG}$ and $\zeta_m^{IG}$ represent the vertical and horizontal angle-of-departures (AoDs) from the IRS to GU $m$ in time slot $t$, respectively, with $\sin\theta_m^{IG} = \frac{H^I}{d_m^{IG}}$, $\sin\zeta_m^{IG} = \frac{\|x^I - x_m^G\|}{\|\mathbf{z}^I - \mathbf{z}_m^G\|}$ and

$\cos\zeta_m^{IG} = \frac{\|y^I - y_m^G\|}{\|\mathbf{z}^I - \mathbf{z}_m^G\|}$. In Eq. (2), $\eta$ is the path loss exponent, $k_m^{IG}$ is the random scattering components modeled by a circularly symmetric complex Gaussian random variable with zero mean and unit variance. The distance between IRS and GU $m$ in time slot $t$ is expressed as $d_m^{IG}[t] = \sqrt{\|\mathbf{z}^I - \mathbf{z}_m^G\|^2 + (H^I)^2}$.

## 3 Problem Formulation

Substituting the channel coefficients and phase shifts of IRS, the achievable rates at legitimate user $m$ and Eve in time slot $t$ are respectively given by

$$R_m^G[t] = \alpha_m[t]\, B \log_2\left(1 + \frac{P}{\sigma^2}\left|\frac{\sqrt{\beta_0}}{d_m^{AG}[t]} + \frac{\beta_0 |k_m^{IG}| \Theta_m^L[t]}{d_m^{IG\frac{\eta}{2}} d^{AI}[t]}\right|^2\right) \tag{3}$$

$$\begin{aligned} R_{m,0}^G[t] &= \mathbb{E}\left[\alpha_m[t] B \log_2\left(1 + \tfrac{P}{\sigma^2}\left|\left(h_0^{AG}[t] + \left(\mathbf{q}_0^{IG}\right)^H \mathbf{\Phi}_m[t]\, \mathbf{H}^{AI}[t]\right)\right|^2\right)\right] \\ &\le \alpha_m[t] B \log_2\left(1 + \tfrac{P}{\sigma^2}\left(\frac{\beta_0}{\left(d_0^{AG}[t]\right)^2} + \frac{\beta_0^2\left|\Theta_{m,0}^G[t]\right|^2}{\left(d_0^{IG}\right)^{\eta}\left(d^{AI}[t]\right)^2}\right)\right) = \widetilde{R}_{m,0}^G[t], \end{aligned} \tag{4}$$

where $\alpha_m[t] \in \{0,1\}$ is the scheduling variable that indicates the association between the UAV and legitimate user $m \in \mathcal{M}$ in time slot $t$, $B$ is the system bandwidth, $P$ is the transmit power of the UAV, $\sigma^2$ is the noise power. Since the small scale fading between the IRS and Eve is difficult to obtain, we assume that Eve can achieve the upper bound of $R_{m,0}^G[t]$ using Jensen's inequality, which is the worst-case scenario. When serving user $m \in \mathcal{M}$, the phase of the UAV-IRS-user or Eve link is $\Theta_m^G[t] = e^{-j\{\pi[(N_x-1)\phi_m^{AIG1} + (N_y-1)\phi_m^{AIG2}] - \theta_{m,n_x,n_y}[t] + \arg(k_m^{IG})\}}$ or $\Theta_{m,0}^G[t] = e^{-j\{\pi[(N_x-1)\phi_0^{AIG1} + (N_y-1)\phi_0^{AIG2}] - \theta_{m,n_x,n_y}[t]\}}$, where $\phi_m^{AIG1} = \sin\theta_m^{IG} \cdot \cos\zeta_m^{IG} - \sin\theta^{AI}[t]\cos\zeta^{AI}[t]$, $\phi_m^{AIG2} = \sin\theta_m^{IG}\sin\zeta_m^{IG} - \sin\theta^{AI}[t]\sin\zeta^{AI}[t]$.

We jointly optimize the user scheduling $\mathbf{A} = \{\alpha_m[t], m \in \mathcal{M}, \forall t\}$, phase shifts of IRS $\mathbf{\Psi} = \{\mathbf{\Phi}_m[t], m \in \mathcal{M}, \forall t\}$, the horizontal trajectory of UAV $\mathbf{Q} = \{\mathbf{q}[t], \forall t\}$, and the vertical trajectory of UAV $\mathbf{H} = \{H^A[t], \forall t\}$ as follows

$$\begin{aligned} \mathrm{P}_1: \max_{\mathbf{A},\mathbf{\Psi},\mathbf{Q},\mathbf{H}} \quad & \overline{R}_s = \tfrac{1}{T}\sum_{m=1}^{M}\sum_{t=1}^{T}\left[R_m^G[t] - \widetilde{R}_{m,0}^G[t]\right]^+ \\ \text{s.t.} \quad & C1: \alpha_m[t] \in \{0,1\}, m \in \mathcal{M}, \forall t, \\ & C2: \sum_{m=1}^{M} \alpha_m[t] \le 1, \forall t, \\ & C3: \sum_{t=1}^{T} R_m^G[t]\,\tau \ge C_m, m \in \mathcal{M}, \\ & C4: \mathbf{z}^A[1] = \mathbf{z}^{begin}, \mathbf{z}^A[T] = \mathbf{z}^{end}, \\ & C5: \theta_{m,n_x,n_y}[t] \in [0, 2\pi], m \in \mathcal{M}, \forall n_x, n_y, t, \\ & C6: \left\|\mathbf{q}[t+1] - \mathbf{q}[t]\right\| \le V_h\tau, \forall t, \\ & C7: \left\|H^A[t+1] - H^A[t]\right\| \le V_z\tau, \forall t, \end{aligned} \tag{5}$$

where $C2$ ensures that within each time slot, the UAV is associated with at most one legitimate user for information transmission, $C3$ ensures the minimum QoS of each legitimate user.

## 4 Proposed Algorithm

We divide Problem $\mathrm{P}_1$ into four sub-problems and propose an iterative algorithm based on AO and SCA to find the suboptimal solution.

### 4.1 User Scheduling

By fixing $\mathbf{\Psi}$, $\mathbf{Q}$ and $\mathbf{H}$ and relaxing the binary variable $\alpha_m[t]$, Problem $\mathrm{P}_1$ is reduced to the following problem:

$$\begin{aligned} \mathrm{P}_2: \max_{\mathbf{A}} \ & \overline{R}_s \\ \text{s.t.} \ & C2, C3, \\ & \overline{C1}: 0 \leq \alpha_m[t] \leq 1, m \in \mathcal{M}, \forall t. \end{aligned} \tag{6}$$

$\mathrm{P}_2$ is a standard convex optimization problem and can be efficiently solved by existing optimization tools, such as CVX [8].

### 4.2 Optimizing the Phase Shifts of IRS

The IRS can only acquire the small-scale fading of the legitimate users and cannot know that of the Eve, so we use the phase alignment technology at the legitimate user $m$ to obtain the optimal phase shift in time slot $t$ :

$$\theta^*_{m,n_x,n_y}[t] = \pi\left[(N_x-1)\,\phi_m^{AIG} + (N_y-1)\,\phi_m^{AIG}\right] + \arg\left(k_m^{IG}\right), m \in \mathcal{M}. \tag{7}$$

Right now, we can calculate that $\Theta_m^G[t]$ is a constant $N$. Similarly, we can get $\Theta_{m,0}^G[t]^*$ by substituting $\theta^*_{m,n_x,n_y}[t]$ into $\Theta_{m,0}^G[t]$.

### 4.3 Optimizing the Horizontal Trajectory of the UAV

Equation (3) and Eq. (4) can be respectively rewritten as

$$R_m^G[t] = \alpha_m[t]\,B\log_2\left(1+\frac{P}{\sigma^2}\left|Dd_m^{AG}[t]^{-1} + Ed^{AI}[t]^{-1}\right|^2\right) \tag{8}$$

$$\widetilde{R}_{m,0}^G[t] = \alpha_m[t]\,B\log_2\left[1+\frac{P}{\sigma^2}\left(D^2 d_0^{AG}[t]^{-2} + Fd^{AI}[t]^{-2}\right)\right], \tag{9}$$

where $D=\sqrt{\beta_0}$,$E=\beta_0 N\left|k_m^{IG}\right|{d_m^{IG}}^{\frac{-\eta}{2}}$ and $F=\beta_0^2\left|\Theta_{m,0}^G[t]^*\right|^2 {d_0^{IG}}^{-\eta}$.

Given $\mathbf{\Psi}$, $\mathbf{A}$, and $\mathbf{H}$, Problem $\mathrm{P}_1$ can be reformulated as

$$\begin{aligned} \mathrm{P}_3: \max_{\mathbf{Q}} \ & \overline{R}_s \\ \text{s.t.} \ & C3, C4, C6. \end{aligned} \tag{10}$$

Owing to the objective function $\overline{R}_s$ and constraint $C3$ are neither concave nor convex respect to $\mathbf{q}[t]$, $\mathrm{P}_3$ is still a non-convex problem. Specifically, we require optimizing $R_m^G[t]$ and $\widetilde{R}_{m,0}^G[t]$ separately to tackle the non-convexity.

**Optimizing $\boldsymbol{R_m^G}[t]$** Eq. (8) is non-convex with respect to $\mathbf{q}[t]$, so we introduce two slack variables $\chi_m^h[t]$ and $\gamma^h[t]$ which satisfy the constraints: $\chi_m^h[t] \geq \left(d_m^{AU}[t]\right)^2$ and $\gamma^h[t] \geq \left(d^{AI}[t]\right)^2$. Thus, $R_m^G[t]$ ($m \in \mathcal{M}$) can be re-expressed as $R_m^{Gh}[t] = \alpha_m[t] B \log_2[1 + \frac{P}{\sigma^2}(\frac{D^2}{\chi_m^h[t]} + \frac{E^2}{\gamma^h[t]} + \frac{2DE}{(\chi_m^h[t]\gamma^h[t])^{\frac{1}{2}}})]$, where $J_m^h[t] = 1 + \frac{P}{\sigma^2}(\frac{D^2}{\chi_m^h[t]^{(l)}} + \frac{E^2}{\gamma^h[t]^{(l)}} + \frac{2DE}{(\chi_m^h[t]^{(l)}\gamma^h[t]^{(l)})^{\frac{1}{2}}})$, $S_m^h[t] = -\frac{P}{\sigma^2}(\frac{D^2}{(\chi_m^h[t]^{(l)})^2} + \frac{DE}{(\chi_m^h[t]^{(l)})^{\frac{3}{2}}(\gamma^h[t]^{(l)})^{\frac{1}{2}}})$, and $Y_m^h[t] = -\frac{P}{\sigma^2}(\frac{E^2}{(\gamma^h[t]^{(l)})^2} + \frac{DE}{(\chi_m^h[t]^{(l)})^{\frac{1}{2}}(\gamma^h[t]^{(l)})^{\frac{3}{2}}})$. Since $R_m^{Gh}[t]$ is convex with respect to $\chi_m^h[t]$ and $\gamma^h[t]$, we acquire the lower-bound of $R_m^{Gh}[t]$ as $\widetilde{R}_m^{Gh}[t]$ to make it concave:

$$\widetilde{R}_m^{Gh}[t] = \alpha_m[t] B \left( \log_2\left(J_m^h[t]\right) + \frac{S_m^h[t]\left(\chi_m^h[t] - \chi_m^h[t]^{(l)}\right) + Y_m^h[t]\left(\gamma^h[t] - \gamma^h[t]^{(l)}\right)}{J_m^h[t] \ln 2} \right). \tag{11}$$

**Optimizing $\boldsymbol{\widetilde{R}_{m,0}^G}[t]$** For optimizing the achievable rate of the Eve, we introduce two slack variables $\zeta^h[t]$ and $\varpi^h[t]$ into Eq. (9):

$$\widetilde{R}_{m,0}^{Gh}[t] = \alpha_m[t] B \log_2 \left[1 + \frac{P}{\sigma^2}\left(\frac{D^2}{\zeta^h[t]} + \frac{F}{\varpi^h[t]}\right)\right] \tag{12}$$

which is a joint convex function of $\zeta^h[t]$ and $\varpi^h[t]$. Accordingly, new constraints $\zeta^h[t] \leq \left(d_0^{AG}[t]\right)^2$ and $\varpi^h[t] \leq (d^{AI}[t])^2$are introduced. Since the right-hand terms of the constraints are convex with respect to $\mathbf{q}[t]$, we perform the first-order Taylor expansion at $l$th iteration to make them concave: $(d_0^{AG}[t])^2 \geq q^{lb1}[t] = (d_0^{AG}[t]^{(l)})^2 + 2(q^{(l)}[t] - \mathbf{z}_0^G)^T(\mathbf{q}[t] - \mathbf{q}^{(l)}[t])$ and $(d^{AI}[t])^2 \geq q^{lb2}[t] = (d^{AI}[t]^{(l)})^2 + 2(q^{(l)}[t] - \mathbf{z}^I)^T(\mathbf{q}[t] - \mathbf{q}^{(l)}[t])$.

At this point, Problem $\mathrm{P}_3$ is transformed into Problem $\mathrm{P}_4$:

$$\begin{aligned} \mathrm{P}_4: \max_{\mathbf{q}[t], \chi_m^h[t], \gamma^h[t], \zeta^h[t], \varpi^h[t]} \ & \widehat{R}_h \\ \text{s.t.} \ & C4, C6, \\ \overline{C3}: \ & \sum_{t=1}^{T} \widetilde{R}_m^{Gh}[t]\tau \geq C_m, m \in \mathcal{M}, \\ C9: \ & \chi_m^h[t] \geq \left(d_m^{AU}[t]\right)^2, m \in \mathcal{M}, \forall t, \\ C10: \ & \gamma^h[t] \geq \left(d^{AI}[t]\right)^2, \forall t, \\ C11: \ & \zeta^h[t] \leq q^{lb1}[t], \forall t, \\ C12: \ & \varpi^h[t] \leq q^{lb2}[t], \forall t, \end{aligned} \tag{13}$$

where $\widehat{R}_h = \frac{1}{T}\sum_{m=1}^{M}\sum_{t=1}^{T}\left(\widetilde{R}_m^{Gh}[t] - \widetilde{R}_{m,0}^{Gh}\right)$. Problem $\mathrm{P}_4$ is convex and can be efficiently solved.

### 4.4 Optimizing the Vertical Trajectory of the UAV

Given $\mathbf{\Psi}$, $\mathbf{A}$, and $\mathbf{Q}$, the sub-problem for finding the UAV's vertical trajectory can be equivalently reformulated as

$$\begin{aligned} \mathrm{P}_5 : \max_{\mathbf{H}} \ & \overline{R}_s \\ \text{s.t.} \ & C3, C4, C5, C8. \end{aligned} \tag{14}$$

Since $\overline{R}_s$ and constraint $C3$ are neither concave nor convex, Problem $\mathrm{P}_4$ is still a non-convex problem. Notice that $R_m^G[t]$ and $\widetilde{R}_{m,0}^G[t]$ can be optimized using the same method applied to deal with Problem $\mathrm{P}_3$. Hence, the problem for optimizing the UAV's vertical trajectory is transformed as follows:

$$\begin{aligned} \mathrm{P}_6 : \max_{H^A[t],\chi_m^v[t],\gamma^v[t],\zeta^v[t],\varpi^v[t]} \ & \widehat{R}_v \\ \text{s.t.} \ & C4, C7, \\ \widetilde{C3}: \ & \sum_{t=1}^{T}\widetilde{R}_m^{Gv}[t]\,\tau \geq C_m, m \in \mathcal{M}, \\ C13: \ & \chi_m^v[t] \geq \left(d_m^{AU}[t]\right)^2, m \in \mathcal{M}, \forall t, \\ C14: \ & \gamma^v[t] \geq \left(d^{AI}[t]\right)^2, \forall t, \\ C15: \ & \zeta^v[t] \leq H^{lb1}[t], \forall t, \\ C16: \ & \varpi^v[t] \leq H^{lb2}[t], \forall t, \end{aligned} \tag{15}$$

where $\widehat{R}_v = \frac{1}{T}\sum_{m=1}^{M}\sum_{t=1}^{T}\left(\widetilde{R}_m^{Gv}[t] - \widetilde{R}_{m,0}^{Gv}[t]\right)$. $\widetilde{R}_m^{Gv}[t]$ and $\widetilde{R}_{m,0}^{Gv}[t]$ can be derived using the similar approach as given by Eq. 11 and Eq. 12, respectively, thus being omitted for brevity.

### 4.5 Algorithm Description

Problem $\mathrm{P}_1$ is divided into four sub-problems, and an iterative algorithm based on AO and SCA methods is designed to solve them alternatively, as summarized in Algorithm 1. In each iteration $l$, the following steps are conducted sequentially. In step 1, the optimal user scheduling $\mathbf{A}^{(l+1)}$ is obtained by solving $\mathrm{P}_2$ given the phase shifts of IRS $\mathbf{\Psi}^{(l)}$ and 3D UAV trajectory $\left\{\mathbf{Q}^{(l)}, \mathbf{H}^{(l)}\right\}$. In step 2, $\mathrm{P}_4$ is solved iteratively using SCA to find the optimal horizontal trajectory $\mathbf{Q}^{(l+1)}$ given the user scheduling, phase shifts and vertical trajectory $\left\{\mathbf{A}^{(l+1)}, \mathbf{\Psi}^{(l)}, \mathbf{H}^{(l)}\right\}$. In step 3, $\mathrm{P}_6$ is solved iteratively using SCA to find the optimal vertical trajectory $\mathbf{H}^{(l+1)}$ given the user scheduling, phase shifts and horizontal trajectory $\left\{\mathbf{A}^{(l+1)}, \mathbf{\Psi}^{(l+1)}, \mathbf{Q}^{(l+1)}\right\}$. In step 4, the phase shift matrix $\mathbf{\Psi}^{(l+1)}$ is iteratively updated. The above procedure is repeated until convergence, i.e., the tolerance is less than $\varepsilon_3$.

**Algorithm 1.** An iterative algorithm based on AO and SCA for Problem P1

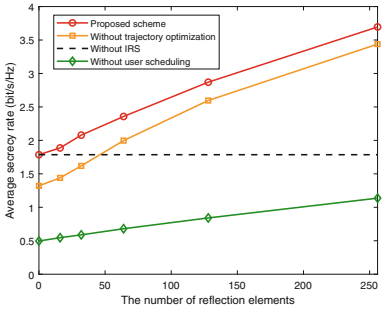
**Initialize** the $\mathbf{A}^0$, $\mathbf{\Psi}^0$, $\mathbf{Q}^0$, $\mathbf{H}^0$, $l = 0$, $\varepsilon_1$, $\varepsilon_2$, $\varepsilon_3$;
**repeat**
  Solve $\mathrm{P}_2$ (Eq. (6)) to find $\mathbf{A}^{(l+1)}$ for given $\left\{\mathbf{Q}^{(l)}, \mathbf{H}^{(l)}, \mathbf{\Psi}^{(l)}\right\}$;
  **repeat**
    Solve $\mathrm{P}_4$ (Eq. (13)) to find $\mathbf{Q}^{(l+1)}$ for given $\left\{\mathbf{A}^{(l+1)}, \mathbf{\Psi}^{(l)}, \mathbf{H}^{(l)}\right\}$;
  **until** $\left|\widehat{R}_h^{(l)} - \widehat{R}_h^{(l-1)}\right| < \varepsilon_1$;
  **repeat**
    Solve $\mathrm{P}_6$ (Eq. (15)) to find $\mathbf{H}^{(l+1)}$ for given $\left\{\mathbf{A}^{(l+1)}, \mathbf{\Psi}^{(l+1)}, \mathbf{Q}^{(l+1)}\right\}$;
  **until** $\left|\widehat{R}_v^{(l)} - \widehat{R}_v^{(l-1)}\right| < \varepsilon_2$;
  Update $\mathbf{\Psi}^{(l+1)}$with $\left\{\mathbf{A}^{(l+1)}, \mathbf{Q}^{(l+1)}, \mathbf{H}^{(l+1)}\right\}$, and $l = l + 1$;
**until** $\left|\overline{R}_s^{(l)} - \overline{R}_s^{(l-1)}\right| < \varepsilon_3$;

## 5 Simulation Results

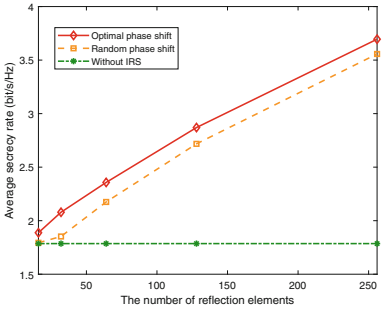
In this section, we present the simulation results to evaluate the performance of our proposed algorithm. The horizontal coordinates of the legitimate users and Eve are $[200, 100; 270, 50; 230, 60]^T$ and $[100, 40]$, respectively. The position of the IRS is $[200, 0, 10]$. The initial and final position of the UAV are set as $[0, 80, 80]$ and $[450, 70, 80]$, respectively. Other parameters are set as $V_h = 40$ m/s, $V_z = 3$ m/s, $P = 1$ W, $T = 50$, $\tau = 1$ s, $\sigma^2 = -80$ dBm, $\beta_0 = -20$ dB, $\eta = 2.3$, $\varepsilon_1 = \varepsilon_2 = 10^{-4}$ and $\varepsilon_3 = 10^{-8}$.

For comparison, we demonstrate the performances of the four schemes: (1) Our proposed scheme that jointly optimizes user scheduling, 3D UAV trajectory and IRS phase shifts; (2) The proposed scheme without UAV trajectory optimization; (3) The proposed scheme without user scheduling; (4) The transmission scheme without the aid of IRS. At first, we plot in Fig. 2 (a) the average secrecy rate curves for the four transmission policies. It is observed that for any IRS-aided transmission scheme, the average secrecy rate increases with the increasing number of reflecting elements of IRS. Our proposed scheme performs the best among the four transmission policies. It is worth mentioning that the proposed scheme with user scheduling is significantly superior to the scheme without user scheduling. In brief, it is highly necessary to jointly optimize the 3D UAV trajectory, phase shifts of IRS and user scheduling.

In Fig. 2 (b), we show the average secrecy rate for different IRS phase control policies, i.e., optimal phase control, random phase control, and without IRS. We can see that the average secrecy rate of the legitimate users increases with the number of reflection elements when IRS is adopted with optimal or random phase control. Through the optimal phase control for IRS, the proposed scheme induces a significant performance gain compared to the transmission scheme without IRS, and achieves a larger secrecy rate than the scheme with random phase control.




(a) Different transmission policies




(b) Different IRS phase control policies

**Fig. 2.** Average achievable rate versus the number of reflection elements

## 6 Summary

In this paper, we studied the IRS-assisted secure UAV communication. The average secrecy rate of legitimate users is improved by jointly optimizing the user scheduling, 3D trajectory of the UAV, and phase shifts of the IRS. To solve the non-convex, we proposed an iterative algorithm based on AO and SCA. Numerical results demonstrated the superiority of the algorithm.

## References

1. Wang, H., Wang, J., Chen, J., Gong, Y., Ding, G.: Network-connected UAV communications: potentials and challenges. China Commun. **15**(12), 111–121 (2018)
2. Zhang, G., Wu, Q., Cui, M., Zhang, R.: Securing UAV communications via joint trajectory and power control. IEEE Trans. Wireless Commun. **18**(2), 1376–1389 (2019)
3. Chen, X., et al.: Securing aerial-ground transmission for NOMA-UAV networks. IEEE Network. **34**(6), 171–177 (2020)
4. Hong, S., Pan, C., Ren, H., Wang, K., Nallanathan, A.: Artificial-noise-aided secure mimo wireless communications via intelligent reflecting surface. IEEE Trans. Wireless Commun. **68**(12), 7851–7866 (2020)
5. Chu, Z., Hao, W., Xiao, P., Shi, J.: Intelligent reflecting surface aided multi-antenna secure transmission. IEEE Wirel. Commun. Le. **9**(1), 108–112 (2020)
6. Fang, S., Chen, G., Li, Y.: Joint optimization for secure intelligent reflecting surface assisted UAV networks. IEEE Wirel. Commun. Le. **10**(2), 276–280 (2021)
7. Pang, X., Zhao, N., Tang, J., Wu, C., Niyato, D., Wong, K.K.: IRS-assisted secure UAV transmission via joint trajectory and beamforming design. IEEE Trans. Commun. **70**(2), 1140–1152 (2022)
8. Matlab Software for Disciplined Convex Programming. http://cvxr.com/cvx (2016)

# Intelligent Reflecting Surface-Assisted Fresh Data Collection in UAV Communications

Hongli Huang, Juan Liu(✉), and Lingfu Xie

Ningbo University, Ningbo 315211, China
{2011082102,liujuan1,xielingfu}@nbu.edu.cn

**Abstract.** This paper considers employing Intelligent reflecting Surface (IRS) in unmanned aerial vehicle (UAV) assisted wireless communications to ensure the freshness of the collected data in Internet of Things (IoT). We aim to minimize the average Age of Information (AoI) of the sensor nodes (SNs) by jointly optimizing the UAV flight trajectory, the SNs' scheduling and the IRS phase shift matrix. It is modeled as a Markovian Decision Process (MDP) problem. A deep reinforcement learning algorithm based on a Twin Delayed Deep Deterministic Policy Gradient (TD3) is proposed to learn and find the optimal UAV trajectory and scheduling of the SNs. For a scheduled transmission, the IRS is used based on the channel information to align the signal phase shifts. Simulation results show that IRS-assisted UAV data collection can significantly reduce the AoI of the SNs.

**Keywords:** Unmanned aerial vehicle · intelligent reflecting surface · deep reinforcement learning · age of information

## 1 Introduction

Reliable and timely sensory information by ground sensor nodes (SNs) is critical to applications in Internet of Things (IoT). It is generally challenging for the SNs with limited battery capacity to communicate reliably over long distances. In recent years, unmanned aerial vehicles (UAVs) are routinely used as mobile data collectors in IoT due to their high mobility and easy deployment. The age of information (AoI) is a measure to SNs' information freshness. In [1], AoI was defined as the time elapsed from the generation of the latest packet by a source node to its reception by a target node. For AoI-oriented UAV data collection, [2] designed an online flight trajectories of the UAV based on deep reinforcement learning (DRL) method to minimize the SNs' weighted sum of AoIs, and [3] studied the influence of SNs' sampling and the queueing process on the SNs' average AoI. The above works only account for the Line of Sight (LoS) channel between the UAV and the SN.

In the urban environment, however, the LoS link between the UAV and the SN is likely to be blocked by obstructions like tall buildings. Intelligent reflecting Surfaces (IRS) is one of the technologies that have great potential in future


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 189–197, 2023.
https://doi.org/10.1007/978-981-99-2362-5_24

wireless networks. It is is a plane composed of a large number of low-cost passive reflective elements, each of which can independently adjust the phase of an incoming signal. This allows for intelligently reconstructing the wireless propagation environment and improving the channel quality [4]. IRS is also amenable to installation. As a result, IRS can be used to overcome the channel blockage between the UAV and the SNs. Moreover, with IRS, other aspects of the communication systems, e.g., the UAV energy consumption [5] and the network throughput [6], can be improved.

In contrast to the above works, we consider the deployment of an IRS for UAV-assisted data collection in the IoT. We assume the energy of the UAV is limited and there is no charging station. For either periodic or random sampling of the SNs, we aim to minimize the SNs' average AoI by jointly optimizing the UAV flight trajectory, the SNs' scheduling and IRS phase shift matrix. This is modeled as a finite Markov decision process (MDP). To solve this problem, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [7] in DRL is proposed to learn and find the optimal policy for the flight trajectory and node scheduling. For a scheduled transmission from a SN to the UAV, the IRS, based on the channel information, is operated in such a way that the phase shifts of the signals are aligned. Simulation results demonstrate that with IRS and the learned optimal policy, both the average AOI and the transmission power of the SNs can be significantly reduced.

## 2 System Model

As shown in Fig. 1, we consider an IoT, where $I$ SNs are distributed in the rectangular area with side lengths $x_{max}^U$ and $y_{max}^U$ to sample the environment and a rotary-wing UAV acts as a mobile base station to collect status-update information. The horizontal location of the $i$-th SN is expressed as $q_i = [x_i, y_i] \in \mathbb{R}^2 \, (i \in \mathcal{I} = \{1, ..., I\})$. However, obstructions such as tall buildings and trees in congested cities cause severe path loss and high attenuation to the air-to-ground channels between the UAV and SNs. In this case, we deploy an IRS on a high-rise building with height $H_R$ to improve channel quality by reflecting signals controllably. The horizontal location of the IRS is defined as $q_R = [x_R, y_R] \in \mathbb{R}^2$. For simplicity, we assume a time-slotted system where the length of each time slot is $T_{ts}$ seconds. The UAV flies at height $H_U$ over the rectangular target area. The horizontal location of the UAV at the time slot $t$ can be defined by $q_t^U = \left[x_t^U, y_t^U\right] \in \mathbb{R}^2 \, (t \in \mathcal{T} = \{1, \ldots, T\})$. Furthermore, $T \in \mathbb{N}$ depends on the UAV's maximum carried energy $e_{max}$ and the service process.

### 2.1 Channel Model

The ground-to-air communication channel between each SN and the UAV includes two links: the direct link from the SN to the UAV, and the indirect link reflected by the IRS. The distance between the UAV and SN $i$ at slot $t$ is given by $d_t^{i,U} = \sqrt{||q_t^U - q_i||^2 + (H_U)^2}$. Similarly, the distances between the SN

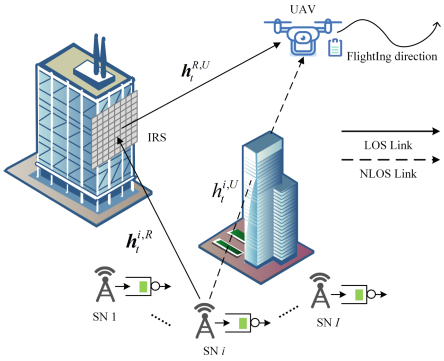



**Fig. 1.** IRS-assisted UAV data collection

$i$ and IRS and between the IRS and UAV are denoted as $d_{i,R}$ and $d_t^{R,U}$, respectively. Assume that the IRS is composed of an $M \times M$ uniform planar array (UPA) with $M^2$ reflective elements. The set of reflective elements is defined as $\mathcal{M} = \{1, ..., M^2\}$.

**1) Direct link:** According to the channel model in [8], the features of LoS and non-line-of-sight (NLoS) links are preserved and appear with a certain probability, respectively. Then, the channel gain between the UAV and SN $i$ at slot $t$ is given by

$$h_t^{i,U} = \begin{cases} \sqrt{\frac{\beta_0}{\left(d_t^{i,U}\right)^{\alpha_1}}}\widetilde{h}_{i,t}, & \text{LoS link} \\ \sqrt{\nu\frac{\beta_0}{\left(d_t^{i,U}\right)^{\alpha_1}}}\widetilde{h}_{i,t}, & \text{NLoS link} \end{cases}, \tag{1}$$

where $\beta_0$ is the pathloss at the reference distance of 1 m, $\alpha_1$ is the path loss exponent for the direct link, $\nu < 1$ denotes the attenuation factor due to NLoS, and $\widetilde{h}_{i,t}$ is the small-scale fading that follows the complex Gaussian distribution with mean 0 and variance 1.

**2) Indirect link:** The channel gain between the IRS and SN $i$ obeys the Rician fading at slot $t$, $\boldsymbol{h}_t^{i,R} \in \mathbb{C}^{M^2 \times 1}$, which can be expressed as $\boldsymbol{h}_t^{i,R} = \sqrt{\frac{\beta_0}{(d_{i,R})^{\alpha_2}}}\left(\sqrt{\frac{k}{1+k}}\boldsymbol{h}_{LoS}^{i,R} + \sqrt{\frac{1}{1+k}}\boldsymbol{h}_{t,NLoS}^{i,R}\right)$, where $k$ is the Rician factor, $\alpha_2$ is the path loss exponent between the SN and the IRS, $\boldsymbol{h}_{LoS}^{i,R}$ is the LOS component, and $\boldsymbol{h}_{t,NLoS}^{i,R}$ is the NLOS component modeled as a complex Gaussian variable with mean 0 and variance 1. Here, $\boldsymbol{h}_{LoS}^{i,R} = [1, \cdots, e^{-j\frac{2\pi d}{\lambda}(M-1)\sin\theta_{i,R}\cos\zeta_{i,R}}]^H \otimes [1, \cdots, e^{-j\frac{2\pi d}{\lambda}(M-1)\sin\theta_{i,R}\sin\zeta_{i,R}}]^H \in \mathbb{C}^{M^2\times 1}$, where $d$ is the distance between the IRS elements, $\lambda$ is the carrier length, $\theta_{i,R}$ and $\zeta_{i,R}$ represent the vertical and horizontal angle-of-departures (AoDs) from the SN $i$ to IRS at slot $t$, respectively. In addition, the geographical relationships are $\sin\theta_{i,R} = \frac{H_R}{\|q_R - q_i\|}$, $\sin\zeta_{i,R} = \frac{|x_i - x_R|}{\|q_R - q_i\|}$ and $\cos\zeta_{i,R} = \frac{|y_i - y_R|}{\|q_R - q_i\|}$.

The channel between the UAV and the IRS is dominated by the LoS link. Similarly, the channel gain at slot $t$ is expressed as $\boldsymbol{h}_t^{R,U} = \sqrt{\frac{\beta_0}{\left(d_t^{R,U}\right)^2}} \boldsymbol{h}_{t,LoS}^{R,U} \in \mathbb{C}^{M^2 \times 1}$, where $\boldsymbol{h}_{t,LoS}^{R,U}$ is the LOS component from the IRS to the UAV. The IRS phase shift matrix at slot $t$ is defined as $\Theta_t = \text{diag}\left(e^{j\theta_t^1}, \ldots, e^{j\theta_t^{M^2}}\right)$, where $\theta_t^m \in [0, 2\pi)$ is the phase shift of the $m$-th element. Therefore, the signal-to-noise ratio (SNR) can be computed as $\eta_t^{i,U} = \frac{P\left|\left(\boldsymbol{h}_t^{R,U}\right)^H \Theta_t \boldsymbol{h}_t^{i,R} + h_t^{i,U}\right|^2}{\sigma^2}$, where $P$ is the SN's transmit power and $\sigma^2$ is the noise power. If the SNR is less than a threshold $\eta_{th}$, i.e., $\eta_t^{i,U} < \eta_{th}$, the UAV cannot decode the received signal successfully.

### 2.2 Queuing Model

Each SN samples periodically or randomly the environmental information, referred to as fixed sampling and random sampling. The sensed information is packaged into an update packet of $\omega$ bits with a timestamp [3]. Then, the packet is stored in the buffer of the SN and waits for collection by the UAV. Let $g_t^i \in \{0, 1\}$ denotes the sampling action of SN $i$ at slot $t$. Specifically, $g_t^i = 1$ denotes that SN $i$ generates an update packet at slot $t$, and otherwise $g_t^i = 0$. Once an update packet arrives at SN $i$ in the beginning of each slot $t$, its lifetime is recorded and updated as

$$U_t^i = \begin{cases} 0, & g_t^i = 1 \\ U_{t-1}^i + 1, & \text{otherwise} \end{cases}. \tag{2}$$

Let $c_t^i \in \{0, 1\}$ be the binary user scheduling variable. $c_t^i = 1$ indicates that the SN $i$ is associated and ready to send one update packet to the UAV at slot $t$, and otherwise $c_t^i = 0$. To fully exploit the IRS, it is assumed that the UAV is associated with at most one SN in each time slot. If the update packet is successfully delivered to the UAV with the aid of IRS, we say that SN $i$ is served at slot $t$. Accordingly, the service state of SN $i$ is set $z_t^i = 1$. If the transmission is failed or no transmission takes place, the service state is set as $z_t^i = 0$. After a successful transmission, the AoI of this SN is updated according to the lifetime of the delivered update packet, and otherwise the AoI increases by one after a time slot. At the beginning of each slot $t$, the AoI is updated as

$$A_t^i = \begin{cases} U_{t-1}^i + 1, & z_{t-1}^i = 1 \\ A_{t-1}^i + 1, & \text{otherwise} \end{cases}. \tag{3}$$

The average AoI of all SNs at time slot $t$ is given by $\overline{A_t} = \frac{1}{I} \sum_{i=1}^{I} A_t^i$.

### 2.3 Problem Description

The UAV consumes energy on flight and hovering. The energy consumption for receiving and decoding the update packets is relatively small and can be omitted.

The UAV hovers and collects data during the transmission interval $T_s$ and then flies to the next location during the remaining time $T_{ts} - T_s$. If no data needs to be collected, the UAV flies across the entire time slot. Therefore, the energy consumption at slot $t$ can be expressed as

$$e_t^{co} = \begin{cases} P_t^f T_{ts}, & \sum_{i=1}^{I} c_t^i = 0 \\ P_{ho} T_s + P_t^f \left(T_{ts} - T_s\right), & \text{otherwise} \end{cases}, \tag{4}$$

where $P_t^f$ is the UAV flight power as a function of flight speed $v_t^U$ and $P_{ho}$ is the hovering power, which can be obtained from Eq. (11) in [5]. Then, the remaining energy of the UAV at slot $t$ can be computed as $e_t^{re} = e_{t-1}^{re} - e_t^{co}$.

The objective is to minimize the weighted sum of the average AoI of all SNs and the UAV's energy consumption by jointly optimizing the UAV flight trajectory $\boldsymbol{Q} = \left[q_t^U, \forall t \in \mathcal{T}\right]$, SN scheduling $\boldsymbol{C} = \left[c_t^i, \forall i \in I, \forall t \in \mathcal{T}\right]$, and IRS's phase shift matrix $\boldsymbol{\Phi} = [\Theta_t, \forall t \in \mathcal{T}]$. The optimization problem can be formulated as

$$\min_{\boldsymbol{Q},\boldsymbol{C},\Phi} \quad \frac{1}{T} \sum_{t=1}^{T} \left(\overline{A_t} + \delta e_t^{co}\right) \tag{5}$$

$$\begin{aligned} s.t.\ & C1: \sum_{i=1}^{I} c_t^i \le 1,\ c_t^i \in \{0,1\},\ \forall t \in \mathcal{T}, \\ & C2: 0 \le \theta_t^m < 2\pi,\ \forall m \in \mathcal{M}, \\ & C3: 0 \le x_t^U \le x_{max}^U,\ 0 \le y_t^U \le y_{max}^U,\ \forall t \in \mathcal{T}, \\ & C4: 0 < \left\| q_t^U - q_{t-1}^U \right\| < v_{max}^U T_{ts},\ \forall t \in \mathcal{T}, \end{aligned}$$

where $v_{max}^U$ is the maximum flying speed of the UAV and $\delta$ is the relative importance factor. It is quite difficult to solve the above mixed integer non-convex problem. Hence, we propose the TD3-based algorithm for UAV-enabled data collection which is able to make the best decision quickly and accurately even when the scale of the problem is very large.

## 3 TD3-Based UAV Data Collection Method

Then, a TD3 algorithm is proposed for the UAV-enabled data collection to find the optimal UAV trajectory and SN scheduling policy efficinetly. During each packet transmission, the IRS's phase shift matrix is optimized based on the perfectly estimated channel state to maximize the received SNR at the UAV.

### 3.1 Optimization of IRS Phase Shift Matrix

Given the SN scheduling and UAV's location, the received SNR is maximized by optimizing the phase shifts of IRS, which is equivalent to the following problem:

$$\min_{\Theta_t} \left| \left(\boldsymbol{h}_t^{R,U}\right)^H \Theta_t \boldsymbol{h}_t^{i,R} + h_t^{i,U} \right|^2 \tag{6}$$

$$s.t.\ 0 \leq \theta_t^m < 2\pi, \forall m \in \mathcal{M},$$

From [9], the optimal phase shifts of IRS can be obtained by aligning the phases of the direct and indirect links between the UAV and the associated SN. In particular, when scheduling SN $i$, the optimal phase shift of the $m$-th element of IRS can be expressed as $\theta_t^{m,*} = \phi_t^{i,U} - \left(\omega_{t,m}^{i,R} + \omega_{t,m}^{R,U}\right), \forall m \in \mathcal{M}$, where $\phi_t^{i,U}$, $\omega_{t,m}^{i,R}$ and $\omega_{t,m}^{R,U}$ are the phases of the direct SN-UAV link, and the SN-IRS and IRS-UAV links via element $m$, respectively.

### 3.2 TD3 Algorithm Design

**MDP Problem Formulation:** The optimization problem can be modeled as a finite MDP. In the sequel, we define the state space, action space and reward function, respectively.

**1) State space:** The system state at slot $t$ is defined as $s_t = \left[q_t^U, e_t^{co}, \boldsymbol{A}_t, \boldsymbol{U}_t\right]$, where $\boldsymbol{A}_t = \left[A_t^i, \forall i \in I\right]$ and $\boldsymbol{U}_t = \left[U_t^i, \forall i \in I\right]$.

**2) Action space:** The system action at slot $t$ is defined as $a_t = \left[\mu_t^U, d_t^U, \boldsymbol{C}_t\right]$, which includes the UAV's flight angle $\mu_t^U \in (0, 2\pi]$ and distance $d_t^U \in [0, d_{i,max}]$. Therefore, the horizontal position of the UAV at slot $t+1$ is updated as $q_{t+1}^U = \left[x_t^U + d_t^U \cos\mu_t^U, y_t^U + d_t^U \sin\mu_t^U\right]$.

**3) Reward function:** Given the state $s_t$ and action $a_t$, the reward function at time slot $t$ can be defined as $r_t(s_t, a_t) = -\left(\overline{A_t} + \delta e_t^{co}\right) + p_t$, where $p_t$ is the penalty at slot $t$ that gives punishment for an invalid action. For example, if the current action $a_t$ causes the UAV to fly out of the target area, we set $p_t < 0$ and otherwise $p_t = 0$.

The objective is to find the optimal policy $\pi^*$ to minimize the long-term return function $C_\pi = \mathbb{E}_\pi\left[\sum_{t=1}^{T}(\gamma)^{t-1} r_t(s_t, a_t)|s_1\right]$, where $\mathbb{E}_\pi$ is the expectation under policy $\pi$, $\gamma \in [0, 1]$ is the discount factor, and $s_1$ is the initial state.

**TD3 Algorithm:** The TD3 algorithm is based on an Actor-Critic framework consisting of deep neural networks (DNN) [7] to find the optimal policy, which has one Actor network that obtains the deterministic policy $\pi_\vartheta(s)$ , and two Critic networks that obtains the value function $Q_\varphi(a, s)$. In addition, there are two target Critic networks with function $Q_{\varphi'}(a, s)$ and one target Actor network with function $\pi_{\vartheta'}(s)$. The Actor network can randomly extract mini-batches of samples from the replay buffer to train the network parameters. The policy gradient is $\nabla_\vartheta J(\vartheta) = N^{-1}\sum \nabla_a Q_{\varphi_1}(s, a)|_{a=\pi_\vartheta(s)} \nabla_\vartheta \pi_\vartheta(s)$, where $N$ is the mini-batch size. The target Actor network copys the Actor network parameters periodically to stabilize the training process, and the target Critic network is the same. The smaller Q value in the two target Critic networks is selected as the target value: $y = r + \gamma \min_{l=1,2} Q_{\varphi_l'}\left(s', \pi_{\vartheta'}\left(s'\right) + \varepsilon\right)$, where $\varepsilon \sim \text{clip}\left(N(0, \sigma), -c, c\right)$ denotes the noise trimmed according to the normal distribution,which can avoid the overestimation problem. Then, the loss function is used to train the two Critic networks, which is expressed as $L(\varphi_i) = N^{-1}\sum\left(y - Q_{\varphi_l}(s, a)\right)^2$. The details of TD3-based UAV Data Collection are shown in Algorithm 1.

**Algorithm 1.** TD3-based UAV Data Collection

Initialize Critic and actor networks $Q_{\varphi_1}, Q_{\varphi_2}$, $\pi_\vartheta$ with random $\varphi_1, \varphi_2$ and $\vartheta$;
Initialize target networks $\varphi_1^{'} \leftarrow \varphi_1$ ,$\varphi_2^{'} \leftarrow \varphi_2$ and $\vartheta^{'} \leftarrow \vartheta$;
Initialize replay buffer and learning rate $\alpha$;
**for** $episode = 1 : episode_{max}$ **do**
  Set $t = 1$, $e_t^{co} = e_{max}$, observe the initial state $s_t$;
  **repeat**
    Select action with exploration noise $a_t \sim \pi_\vartheta(s_t) + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$;
    Execute the action $a_t$, calculate the optimal IRS phase shift matrix $\Theta_t^*$, update the UAV position $q_{t+1}^U$, energy $e_t^{re}$, and AoI $A_t^i\,(i = 1, 2, \ldots, I)$;
    Obtain the reward $r_t$ and the new state $s_{t+1}$, store experience $(s_t, a_t, r_t, s_{t+1})$ in replay buffer;
    Sample a mini-batch of $N$ transitions from replay buffer;
    Update the Critic networks $\varphi_l = \underset{\varphi_l}{\operatorname{argmin}} L(\varphi_l), l = 1, 2$;
    **if** $t \bmod t_{update}$ **then**
      Update the Actor network $\vartheta = \vartheta - \alpha \nabla_\vartheta J(\vartheta)$;
      Update target networks: $\varphi_l^{'} \leftarrow \tau\varphi_l + (1-\tau)\varphi_l^{'}, l = 1, 2$, $\vartheta^{'} \leftarrow \tau\vartheta + (1-\tau)\vartheta^{'}$;
    **end if**
  **until** $e_t^{re} < e_{th}$;
**end for**

## 4 Simulation Results

We consider a 300 m × 400 m rectangular target area, and set the lower left corner of the area as the coordinate origin. The coordinate of IRS is set as $[0, 150, 30]$, and the horizontal coordinate of three SNs are set as $[10, 180]$, $[85, 350]$, $[225, 50]$. The random sampling process is modeled as a Poisson process. The system simulation parameters are set as follows: $T_{ts} = 1\,s$, $T_s = 0.5\,s$, $H_U = 60\,\text{m}$, $\beta_0 = -45\,\text{dB}$ ,$\alpha_1 = 3.1$, $\alpha_2 = 2.3$, $\sigma^2 = -110\,\text{dBm}$, $\omega = 110\,\text{KB}$, $\delta = 0.001$, $\eta_{th} = 0.77$, $v_{max}^U = 40\,\text{m/s}$, $e_{max} = 1.2\,e5\text{J}$, $e_{th} = 8\,e3\text{J}$. If there is no specific explanation, the reinforcement learning parameters are shown in Table 1.

**Table 1.** Learning parameters

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Learning rate for actor | 1e-4 | Learning rate for critic | 1e-3 |
| Exploration noise | 0.1 | Policy noise | 0.15 |
| Software update factor | 0.004 | Reward discount | 0.98 |
| Total number of training episodes | 6e4 | Batch size | 128 |

Figure 2 shows the convergence curves of the proposed TD3 and PPO algorithms, when fixed sampling with rate 0.2 is applied and the RIS is $15 \times 15$. It is observed that the TD3 algorithm converges faster and more stably, and is more

suitable for our studied problem. This is because the PPO algorithm tends to have insufficient explorations and usually find a suboptimal policy.

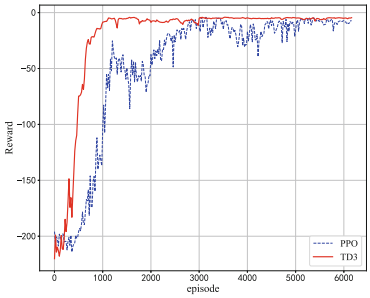
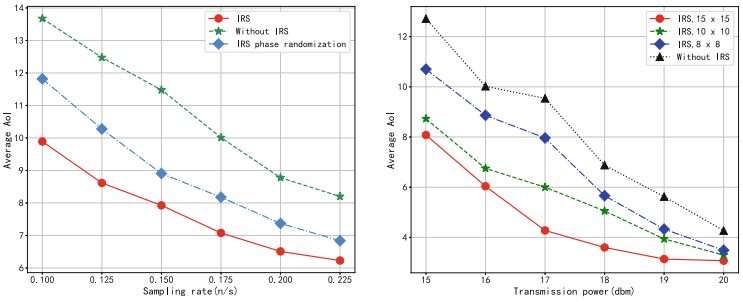



**Fig. 2.** Convergence comparison between TD3 and PPO algorithms.

In Fig. 3(a) and Fig. 3(b), we illustrate the average AoI performance of the proposed TD3-based method for different random sampling rates and transmission powers, respectively. As shown in Fig. 3(a), given any IRS phase control policy, it is observed that a higher sampling rate leads to the smaller average AoI, since the sensing data can be collected more frequently. By optimal phase alignment for IRS, our proposed scheme achieves the minimum AoI value for any sampling rate, which indicates that the IRS-aided UAV data collection scheme can significantly improve the information freshness. In Fig. 3(b), as either the number of IRS reflecting elements or the transmission power of the SN increases, the average AoI can be greatly reduced. In both subfigures, the transmission scheme without IRS leads to the highest AoI of SNs.




(a) IRS phase control (b) The number of IRS elements

**Fig. 3.** The average AoI performance of the proposed TD3 method

## 5 Conclusion

This paper investigated the efficient IRS-assisted UAV data collection problem for IoT. The IRS was deployed on a tall building to improve the channel quality of the UAV and SNs, and the optimal IRS reflection coefficient was obtained by phase alignment. The problem was modeled as a MDP problem, and we proposed the TD3 algorithm in DRL to find to find the optimal UAV trajectory and SN scheduling policy to minimizes the average AoI. Simulation results showed that integrating IRS into UAV data collection can effectively reduce the average AoI regardless of whether the SNs periodically or randomly sample environmental information, and the TD3 algorithm outperforms the PPO algorithm in terms of convergence speed and stability after convergence for the problems in this paper. The larger the IRS the lower the transmission power of the SN with guaranteed average AoI.

## References

1. Kaul, S., Yates, R., Gruteser, M.: Real-time status: how often should one update? In: IEEE INFOCOM, pp. 2731–2735 (2012)
2. Abd-Elmagid, M.A., Ferdowsi, A., Dhillon, H.S., Saad, W.: Deep reinforcement learning for minimizing age-of-information in UAV-assisted networks. In: IEEE Global Communications Conference (GLOBECOM), pp. 1–6 (2019)
3. Tong, P., Liu, J., Wang, X., Bai, B., Dai, H.: Deep reinforcement learning for efficient data collection in UAV-aided Internet of Things. In: IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–6 (2020)
4. Wu, Q., Zhang, S., Zheng, B., You, C., Zhang, R.: Intelligent reflecting surface-aided wireless communications: a tutorial. IEEE Trans. Commun. **69**(5), 3313–3351 (2021)
5. Cai, Y., Wei, Z., Hu, S., Ng, D., W. K., Yuan, J.: Resource allocation for power-efficient IRS-assisted UAV communications. In: IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–7 (2020)
6. Nguyen, K.K., Masaracchia, A., Sharma, V., Poor, H.V., Duong, T.Q.: RIS-assisted UAV communications for IoT with wireless power transfer using deep reinforcement learning. IEEE J. Sel. Top. Sig. Process. **16**(5), 1086–1096 (2022)
7. Fujimoto, S., Hoof, H.V., Meger, D.: Addressing function approximation error in actor-critic methods. arXiv:1802.09477v3 (2018)
8. Al-Hourani, A., Kandeepan, S., Jamalipour, A.: Modeling air-to-ground path loss for low altitude platforms in urban environments. In: IEEE Global Communications Conference, pp. 2898–2904 (2014)
9. Wu, Q., Zhang, R.: Intelligent reflecting surface enhanced wireless network: joint active and passive beamforming design. In: IEEE Global Communications Conference (GLOBECOM), pp. 1–6 (2014)

# UAV-Assisted Fresh Data Collection with MCS in Wireless Powered IoT

Fei Yang, Juan Liu, and Lingfu Xie(✉)

Ningbo University, Ningbo 315211, China
{2011082148,liujuan1,xielingfu}@nbu.edu.cn

**Abstract.** In this paper, we study the unmanned aerial vehicle (UAV) assisted data collection problem to improve the information freshness in wireless powered Internet of things (IoT) networks. In our system, one UAV collects the sensing information from the sensors timely. To improve the system sustainability, one unmanned vehicles acts as a mobile charging station (MCS) to recharge the sensors wirelessly and replace the battery of the UAV when necessary. We aim to design the trajectory of the UAV and the moving path of the MCS jointly to minimize the average age of information (AoI) collected from the sensors. The problem is formulated as a partially observed Markov decision process (POMDP) with large state and action spaces. To seek the optimal solution, a multi-agent deep reinforcement learning algorithm based on value-decomposition networks (VDN) is developed to make real-time decisions according to partial observations in the environment. Simulation results verify the effectiveness of the proposed algorithm.

**Keywords:** Unmanned aerial vehicle · age of information · value-decomposition networks

## 1 Introduction

In IoT networks, due to sensors' energy limitation and channel characteristics, they can not communicate with destination nodes reliably via long distances. By establishing a high-quality short-distance channel with sensors, unmanned aerial vehicle (UAV)-assisted data collection can save the energy consumption of sensors and improve the communication quality and efficiency.

However, the sustainability challenges for UAV-assisted data collection still exist as limited battery capacity of UAVs. To keep long-term work of Internet of things (IoT) system and long-distance flight of UAVs, energy replenishment for IoT system is very necessary and useful. In [1], a mobile charging station (MCS) is used to charging the UAV in a fixed position. A deep reinforcement learning (DRL) approach is used to plan the action strategies of the MCS and the UAV to maximize data collection rate. UAVs can also carry energy harvesting devices for energy supplementation. In [2], the UAV utilizes the harvested energy to transmit sensing data. The transmission opportunities of the sensors


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 198–206, 2023.
https://doi.org/10.1007/978-981-99-2362-5_25

and the UAV's flight trajectory are jointly optimized to minimize the AoI and improve energy efficiency. Applying energy-saving strategies is another way to prolong the service time of UAVs. In IoT Networks, the timeliness of data is an important indicator. In [3], the authors investigate age-optimal data collection and dissemination problems on graphs, where one UAV flies along optimal trajectories to minimize peak and average AoI, respectively.

Different from these works, we consider deploying an MCS to replenish energy for sensors and UAV in an IoT Network. By designing the trajectory of the UAV and the moving path of the MCS efficiently, our proposed UAV-assisted data collection strategy can ensure the information freshness in the long run. Then we formulate this problem as a partially observed Markov decision process (POMDP) problem, and use the value-decomposition networks (VDN) based approach to find the optimal trajectories of the UAV and the MCS. The rest of this paper is organized as follows. Section II introduces the system model and Section III presents the POMDP formulation of the UAV-aided data collection strategy with the aid of MCS. Section IV proposes a VDN-based algorithm to make real-time decisions to find the optimal solution and Section V demonstrates the simulation results. Finally, Section VI concludes the paper.

## 2 System Model

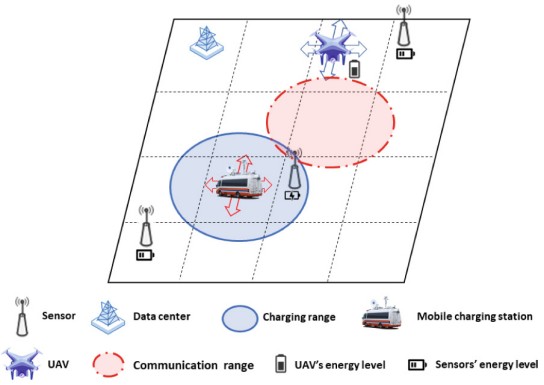



**Fig. 1.** A UAV-assisted data collection system with energy recharging by an MCS

As shown in Fig. 1, we study a UAV-assisted IoT network where $M$ sensors are randomly distributed and sense the real-time environment, and a UAV acts as a mobile relay to collect the sensors' information and forword it to the data center (DC). The coordinates of the sensors and the DC are denoted by $L_m = [x_m, y_m] \in \mathbb{R}^2$ $(m \in \mathcal{M} = \{1, \cdots, M\})$ and $L_0 = [x_0, y_0]$, respectively. The flying altitude of the UAV is $H$. To maintain the IoT network, an MCS can replenish energy for the UAV and sensors. For ease of exposition, the area is

divided into $I$ square grids with the location of each grid center presented by $\boldsymbol{\Omega}_i = [X_i, Y_i] \in \mathbb{R}^2$ $(i = 1, 2, \cdots, I)$, and the side length is denoted by $L_s$. The UAV's opeartion time is divided into $T$ time slots each of which is equal to $T_s$ seconds. In each time slot, the UAV and the MCS move from one grid center to the next ones to perform their own tasks. In each slot $n$, the flying (moving) time of the UAV (MCS) denoted by $T_{u,f}(n) \in [0, T_s]$ $(T_{c,mov}(n) \in [0, T_s])$, and the rest time is used for data collection or energy replenishment.

### 2.1 Ground-to-Air Transmission Channel

In time slot $n$, the trajectory of the UAV is denoted as $L_u(n) = [x_u(n), y_u(n)]$. The distance between the UAV and sensor (or DC) is expressed as $d_{u,m}(n) = \sqrt{||L_u(n) - L_m||^2 + H^2}$ $(m \in \{0\} \cup \mathcal{M})$. When having sufficient energy, the UAV collects the sensing data from the sensors in its coverage range with radius $r_1$, and forwards to the DC. We adopt the widely used hybrid channel model of line-of-sight (LoS) and non-line-of-sight (NLoS) propagation [4]. The channel gain between the UAV and sensor $m$ can be obtained by:

$$h_m(n) = \begin{cases} \beta_0 \left[d_{u,m}(n)\right]^{-k}, & \text{LoS link} \\ \zeta\beta_0 \left[d_{u,m}(n)\right]^{-k}, & \text{NLoS link} \end{cases}, \tag{1}$$

where $\beta_0$ is the channel gain at the reference distance (e.g., 1 m), $\zeta \in (0, 1]$ denotes the additional attenuation coefficient due to NLoS propagation, and $k$ is the path loss exponent. The probability of LoS propagation is expressed as:

$$P_{LoS,m}(n) = \frac{1}{1 + a \exp\left\{-b\left[\frac{180}{\pi} \arcsin(\frac{H}{d_{u,m}(n)}) - a\right]\right\}}, \tag{2}$$

where $a$, $b$ are constants related to the environment. The average channel gain between the UAV and sensor $m$ is

$$h_m^{\text{avg}}(n) = P_{LoS,m}(n) h_{LoS,m}(n) + P_{NLoS,m}(n) h_{NLoS,m}(n), \tag{3}$$

where $P_{NLoS,m}(n) = 1 - P_{LoS,m}(n)$ is the probability of NLoS propagation.

The process of UAV-assisted transmission can be divided into two stages: (1) The UAV collects the sensing data from the sensors when hovering over a grid; (2) The UAV forwards the received information to the DC. Let $T_{u,tx}(n)$ denote the transmission time in slot $n$. When hovering over a grid center $\boldsymbol{\Omega}_i$, the UAV receives data from the sensors using FDMA. With the transmit power $P_m$ at sensor $m$, the achievable data rate is expressed as $R_m(n) = \frac{B}{|\mathcal{M}_i|} \log_2\left(1 + \frac{h_m^{\text{avg}}(n) P_m}{N_0 \frac{B}{|\mathcal{M}_i|}}\right)$, where $B$ and $N_0$ are the system bandwidth and noise power spectral density, respectively, and $\mathcal{M}_i$ is the set of sensors located within the grid $i$. Hence, the transmission time of sensor $m$ in the first stage is $T_{m,tx}(n) = W_m / R_m(n)$, where $W_m$ is the packet size. The energy consumed by sensor $m$ is

$$E_{m,tx}(n) = P_m \cdot T_{m,tx}(n). \tag{4}$$

When sensor $m$ finished its transmission within $T_{tx,1}$ seconds, the information can be successfully received by the UAV. Thus, at the end of the first stage, the total amount of information received by the UAV is expressed as $\sum_{\{m\in\mathcal{M}_i, T_{m,tr}(n)\le T_{tx,1}\}} W_m$. In the second stage, the UAV forwards the received information to the DC. The data rate can be given by:

$$R_2^{tx}(n) = B\log_2\left(1+\frac{h_0^{\text{avg}}(n)P_u(n)}{N_0 B}\right), \tag{5}$$

where $P_u(n)$ is the transmit power of the UAV in slot $n$. To complete the information delivery within $T_{tx,2}(n) = T_{u,tx}(n) - T_{tx,1}$, the transmit power required by the UAV should satisfy

$$P_u(n) = N_0 B(h_0^{\text{avg}}(n))^{-1}\left[2^{\frac{\sum_{\{m\in\mathcal{M}_i, T_{m,tx}(n)\le T_{tx,1}\}} W_m}{BT_{tx,2}(n)}} - 1\right]. \tag{6}$$

The energy consumption of the UAV for transmission can be calculated as:

$$E_{u,tx}(n) = P_u(n)\cdot T_{tx,2}(n). \tag{7}$$

### 2.2 Flying and Hovering Energy Consumption

In time slot $n$, the flight speed can be expressed as $v_{u,f}(n) = L_s/T_{u,f}(n)$. According to the UAV flight energy consumption model [5], the minimum flight power of the UAV is writen as $p_{u,f}^{min}(n) = (v_{u,f}(n)\sin\theta + \hat{v}(n))F$, where $\theta$ is pitch angle and$\hat{v}(t)$ is the required inducted velocity for giving a thrust $F$. $F = m_u g + f_d$ is related to the gravity $m_u g$ and drag coefficient $f_d$. And $\hat{v}(t)$ can be obtained by solving the implicit equation [5]:

$$\hat{v}(t) = \frac{2F}{n_r\pi R^2\rho\sqrt{\left[v_{u,f}(t)^2\cos\theta\right]^2 + \left[v_{u,f}(t)\sin\theta + \hat{v}(t)\right]^2}}, \tag{8}$$

where $n_r$, $R$ and $\rho$ represent the number of rotors, rotor radius and air density, respectively. When the UAV hovers, the minimum hovering power of the UAV is expressed as $p_{u,h}^{min}(n) = \frac{F\sqrt{F}}{\sqrt{\frac{1}{2}\pi n_r R^2\rho}}$. The energy consumption of the UAV for flying and hovering in slot $n$ can be calculated as:

$$E_u^{f,h}(n) = \eta^{-1}\left[p_{u,f}^{min}(n)T_{u,f}(n) + p_{u,h}^{min}(n)T_{u,tx}(n)\right], \tag{9}$$

where $\eta \in (0,1]$ represents the energy conversion efficiency of the UAV.

### 2.3 Energy Supplement

In the system, the MCS moves around to recharge the UAV and sensors. The moving path of MCS is denoted by $L_c(n) = [x_c(n), y_c(n)] \in \boldsymbol{\Omega}_i$ in time slot $n$. When the UAV is hovering over the grid in which the MCS is located, i.e.,

$L_u(n) = L_c(n)$, and is in the recharging state with $C_u(n) = 1$, the UAV lands on the MCS to replace its battery quickly, and takes off to continue data collection. From [5], the energy consumption for ascending and descending motions denoted as $E_u^{as}$ and $E_u^{des}$ are given by:

$$\begin{aligned} E_u^{as} &= \frac{m_u g}{2} v_{as} + \frac{m_u g}{2} \sqrt{v_{as}^2 + \frac{2 m_u g}{\rho \pi R^2}} \\ E_u^{des} &= \frac{m_u g}{2} v_{des} - \frac{m_u g}{2} \sqrt{v_{des}^2 - \frac{2 m_u g}{\rho \pi R^2}} \end{aligned}, \tag{10}$$

where $v_{as}$and $v_{des}$ are ascending and descending flight speeds, respectively.

At the same time, the MCS charges the sensors in grid $i$ via wireless power transfer. Following the power charging model in [6], the charging power can be given by:

$$P_{c,m}(n) = \begin{cases} \frac{\alpha}{(d_{c,m}(n)+\beta)^2} & d_{c,m}(n) \le D \\ 0 & d_{c,m}(n) > D \end{cases}, \tag{11}$$

where $d_{c,m}(n) = \sqrt{||L_c(n) - L_m||^2}$ is the distance between the sensor $m$ and the MCS, $D$ is the charging radius, $\alpha$ and $\beta$ are constants. The energy transferred to the sensor in time slot $n$ is given by $E_{m,c}(n) = P_{c,m}(n)(T_s - T_{c,mov}(n))$.

### 2.4 AoI

AoI is defined as the time since the latest update packet was generated at the source node until received at the destination node [7]. If the data of sensor $m$ is successfully collected ($T_m(n) = 1$), the AoI of sensor $m$ is reset to one, and otherwise increases by one at the beginning of the next slot, i.e.,

$$A_m(n+1) = \begin{cases} 1 & T_m(n) = 1 \\ \min\{A_m(n) + 1, A_{m,max}\} & \text{otherwise} \end{cases}, \tag{12}$$

where $A_{m,max}$ is a relatively large integer. During energy replenishment for the UAV, the AoI values of all sensors increase linearly. The weighted average AoI is evaluated as $\text{AoI}(n) = \sum_{m=1}^{M} p_m A_m(n)$, where $p_m$ represents the importance of sensor node $m$.

## 3 Problem Formulation and Algorithm Design

In this work, we adopt the multi-agent VDN approach to find the trajectory of the UAV and moving path of the MCS jointly. In particular, the UAV-aided data collection and energy recharging problem is formulated in a POMDP framework, where the UAV and MCS act as two intelligent agents and select their moving directions independently after being well trained.

### 3.1 POMDP Formulation

**State Space:** The system state includes the location and energy state of the UAV $\boldsymbol{s}_u(n) = (L_u(n), E_u(n), C_u(n)) \in \boldsymbol{\Omega} \times [0, E_{u,max}] \times \{0, 1\}$, the position of the MCS $L_c(n) \in \boldsymbol{\Omega}$, the energy state and AoI value of each sensor $m$ $\boldsymbol{s}_m(n) = (A_m(n), E_m(n)) \in \{1, \cdots, A_{m,max}\} \times [0, E_{m,max}]$. In the partially observed environment, the observations of the UAV $\boldsymbol{z}_u(n)$ include its own state $\boldsymbol{s}_u(n)$ and the states of the sensors in its coverage range denoted by $\widetilde{\boldsymbol{s}}_u(n)$. Similarly, the observations of the MCS $\boldsymbol{z}_c(n)$ include its location state $L_c(n)$ and the states of the sensors within its coverage denoted by $\widetilde{\boldsymbol{s}}_c(n)$.

**Action Space:** The UAV (MCS) chooses its moving direction or stays still in each slot $n$. Hence, the action of the UAV or MCS can be denoted by $a_{i,f}(n) = [a_{i,h}(n), a_{i,v}(n)] \in \{(-L, 0), (L, 0), (0, -L), (0, L), (0, 0)\} (i \in \{u, c\})$, where $a_{i,h}(n), i \in \{u, c\}$ ( $a_{i,v}(n) i \in \{u, c\}$) represents the moving distance along the x-axis and y-axis, respectively.

**State Update:** The location of the UAV or MCS is updated as $L_i(n+1) = L_i(n) + [a_{i,h}(n), a_{i,v}(n)] (i \in \{u, c\})$. The UAV's energy is updated in three cases. When the UAV is in its recharging state and in the same grid with the MCS, i.e., $C_u(n) = 1$ and $L_u(n) = L_c(n)$, it lands on the MCS to replace its battery quickly, and then takes off to continue its data collection task. During data collection, the UAV collects data for some sensors, i.e., $T_m(n) = 1 (m \in \mathcal{M})$, and its energy is consumed for flying, hovering, and transmission. In other cases, the UAV flies or hovers for a whole time slot with energy consumption on flight and hovering. The energy update for the UAV is given by

$$E_u(n+1) = \begin{cases} E_{u,max} - E_u^{as}, & \text{if } C_u(n) = 1 \text{ and } L_u(n) = L_c(n) \\ E_u(n) - E_u^{f,h}(n) - E_{u,tx}(n), & \text{elsif } T_m(n) = 1, m \in \mathcal{M} \\ E_u(n) - E_u^{f,h}(n), & \text{otherwise} \end{cases} . \tag{13}$$

Similarly, we express the energy update for each sensor $m$ as

$$E_m(n+1) = \begin{cases} \min\{E_m(n) - e_m(n) + E_{m,c}(n), E_{m,max}\} & \text{if } d_{c,m}(n) \le D \\ E_m(n) - e_m(n) - E_{m,tx}(n) & \text{elsif } T_m(n) = 1 \\ E_m(n) - e_m(n) & \text{otherwise} \end{cases} , \tag{14}$$

where $e_m(n)$ is the energy consumption on circuits, $E_{m,c}(n)$ is the wirelessly transferred energy, $E_{m,tr}(n)$ is the transmission energy, and $E_{m,max}$ is the capacity of the battery at sensor $m$. The UAV transits to the recharging state, $C_u(n+1) = 1$, if its energy is below the threshold, i.e., $E_u(n) < E_{u,th}$. Then, the UAV goes back to the data collection state with $C_u(n+1) = 0$ after being quickly equipped with a new battery full of energy.

**Action-observation History:** In the partially observed environment, the action-observation history is employed with the current observation to find the optimal policy. Let $\tau(n) = (\tau_u(n), \tau_c(n))$ denote the joint action-observation history of the UAV and MCS, where $\tau_i(n) = (a_{i,f}(0), \boldsymbol{z}_i(1), \cdots, a_{i,f}(n-1), \boldsymbol{z}_i(n))$ is the action-observation history of agent $i \in \{u, c\}$ till slot $n$.

**Cost Function:** We define the system cost function as the sum of the AoI, energy consumption of the UAV and the charging capacity of the sensors

$$C(s(n), a(n)) = \text{AoI}(n) + \varsigma_1 E_u^{f,h}(n) + \varsigma_2 \sum_{m=1}^{M} p_m (E_{m,max} - E_m(n)) + pf(n), \quad (15)$$

where $\varsigma_1$, $\varsigma_2$ are the weight coefficients, and $pf(n)$ is the non-negative penalty for disciplining inappropriate actions. We aim to find the optimal policy $\pi$ to minimize the long-term cost $C_\pi = E_\pi[\sum_{n=1}^{N} \gamma^{n-1} C(s(n), a(n)|s(1))]$, where $E_\pi$ denotes the expected value under policy $\pi$, and $\gamma \in [0, 1]$ is the discount factor.

**Algorithm 1.** Training of the VDN-based algorithm for joint UAV-aided data collection and energy recharging via MCS

Initialize the Q value functions $Q(\tau(n), a(n); \theta(n)) = Q(\tau(n), a(n); \hat{\theta}(n)) = 0$ with random network parameters $\theta = \hat{\theta}$;
Initialize the replay memory with capacity $G$, global shared counter $episode = 1$;
**for** $episode = 1 : episode_{max}$ **do**
  Get the initial observations $\boldsymbol{z}_i(n)$;
  **for** $n = 1 : N - 1$ **do**
    **for** $i \in \{u, c\}$ in parallel **do**
      Select an action $a_{i,f}(n)$ using the $\varepsilon$-greedy policy;
    **end for**
    **if** $E_u(n) \leq E_u^{des}$ **then**
      The UAV lands on the ground waiting for energy supplement;
    **end if**
    The MCS charges the sensors in its coverage and supply energy for the UAV if they are located in the same grid. Otherwise, the UAV performs data collection and MCS transfers energy wirelessly at the same time. Then, the UAV and MCS execute the joint moving action;
    Update $E_u(n)$, $E_m(n)$, AoI$(n)$, $L_u(n)$, $L_c(n)$, and $C_u(n)$ respectively;
    Calculate the cost $r(n)$ by (15) and receive the next observation $\tau(n+1) = (\tau_u(n), \tau_c(n))$;
    Store $(\tau(n), a(n), r(n), \tau(n+1))$ in the replay buffer;
    Sample a mini-batch of sequential samples from the replay buffer to update the current network parameters as $\theta = \theta + \alpha \nabla_\theta Floss(\theta)$
    Update the target network every $O$ timesteps by set $\hat{\theta} = \theta$ and probability $\varepsilon$;
  **end for**
**end for**

### 3.2 Algorithm Design

VDN is a very popular centralized training decentralized execution algorithm which shares information between the agents during training. the two agents learn to act on their own observations to approach the joint action-value

function that is additively decomposed into local action-value functions, i.e., $Q_{tot}(\tau, a) \approx \sum_{i\in\{u,c\}} Q_i(\tau_i(n), a_{i,f}(n); \theta_i)$, where $\tau$ is a joint action-historical trajectory of observations, $a$ is a joint action. Each agent $i$ maintains an agent network $\theta_i$ to approximiate the local action-value function $Q_i(\tau_i(n), a_{i,f}(n); \theta_i)$. The agent network is composed of an input multi-layer perceptron (MLP) layer, a gated recurrent unit (GRU) layer, and an output MLP layer. The GRU layer is used to capture the long-term and short-term dependencies on the observation-action history so as to relieve severe partial observability. The details of VDN-based UAV-aided data collection and energy recharging via MCS are shown in Algorithm 1. In exception of energy shortage, $E_u(n) \leq E_u^{des}$, the UAV can land to the ground to avoid crashing. In other cases, the UAV and MCS operate to finish the fresh data collection task efficiently.

## 4 Simulation Results

Consider the IoT network consists of $M = 6$ sensors, one UAV and one MCS in a square area of $200m \times 200m$. The area is divided into $10 \times 10$ grids, and each grid is $20m \times 20m$. Major system parameters are set as: $T_s = 1s$, $T_{u,f}(n) = 0.4s$, $T_{c,move}(n) = 0.6s$, $r_1 = 30\sqrt{3}m$, $P_m = 1\,Watt$, $W_m = 3Mbit$, $\beta_0 = -60dB$, $\sigma_2 = -90dB$, $H = 50m$, $k = 2.3$, $\zeta = 0.2$, $n_r = 4$, $R = 0254$, $\rho = 1.225kg/m^3$, $\eta = 0.7$, $f_d = 9.6998N$, $E_{m,max} = 20Joul$, $E_{u,max} = 1.2e5Joul$, $E_{u,th} = 8e3Joul$, $\alpha = 10000$, $\beta = 40$, $D = 10\sqrt{2}$. And we trained $100,000$ episodes in our algorithm, each episode has $2,00$ steps, and batch size $B_s = 2000$.

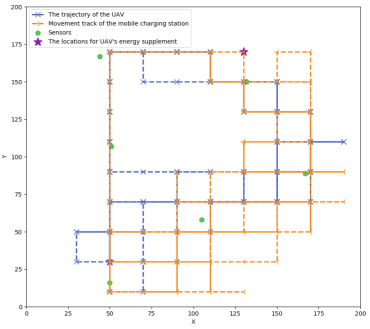



**Fig. 2.** Trajectories of the UAV and MCS

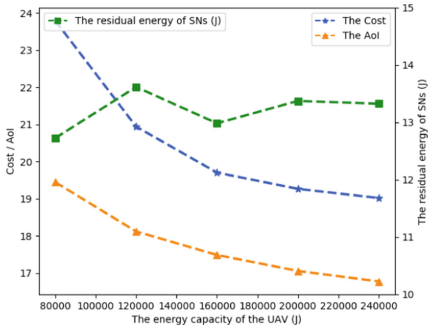



**Fig. 3.** The AoI and energy curves

In Fig. 2, we show the trajectory of the UAV and moving path of the MCS during a training episode. The UAV and the MCS move around the sensors constantly to collect data and recharge sensors independently. During the data collection task, the UAV can find some places to meet the MCS and get energy supply. The locations of replenishing energy for the UAV are marked by purple stars.

In Fig. 3, we study the impact of the energy capacity of the UAV on the system performance, i.e., the cost $r(n)$ , the AoI $\mathrm{AoI}(n)$ and the sensors' average residual energy $\overline{E_m}(n) = \sum_m p_m E_m(n)$. It can be seen that as the UAV's energy capacity increases, the cost $r(n)$ and AoI $\mathrm{AoI}(n)$ significantly decrease, while in sensors' average residual energy nearly keeps constant. This is because when the energy capacity of UAV is low, the UAV needs to spend more time on replenishing energy such that it provide less data service for the sensors. However, the sensors can always get energy replenishment when the MCS moves around.

## 5 Conclusions

In this paper, a problem of UAV-aided data collection and energy recharging via MCS in wireless powered IoT networks was studied. The data collection strategy of the UAV and the scheduling of the MCS were jointly optimized to minimizing the average AoI and maximizing the charging benefit. In the POMDP problem formulation, a multi-agent VDN-based algorithm was designed to find the trajectory of the UAV and the moving path of the MCS in the way of centralized training and decentralized execution. Finally, simulation results showed that the UAV and the MCS are well trained to improve the system performance in terms of low AoI and high durability.

## References

1. Zhang, B., Liu, C.H., Tang, J., Xu, Z., Ma, J., Wang, W.: Learning-based energy-efficient data collection by unmanned vehicles in smart cities. IEEE Trans. Indust. Inform. **14**(4), 1666–1676 (2018)
2. Zhang, N., Liu, J., Xie, L., Tong, P.: A deep reinforcement learning approach to energy–harvesting uav-aided data collection. In: 2020 International Conference on Wireless Communications and Signal Processing (WCSP), pp. 93–98 (2020)
3. Tripathi, V., Talak, R., Modiano, E.: Age optimal information gathering and dissemination on graphs. In: IEEE INFOCOM 2019 – IEEE Conference on Computer Communications, pp. 2422–2430 (2019)
4. Stolaroff, J.K., Samaras, C., ONeill, E.R., Lubers, A., Mitchell, A.S., Ceperley, D.: Modeling air-to-ground path loss for low altitude platforms in urban environments. In: 2014 IEEE Global Communications Conference, pp. 2898–2904 (2014)
5. Stolaroff, J., Samaras, C., O'Neill, E., Lubers, A., Mitchell, A., Ceperley, D.: Energy use and life cycle greenhouse gas emissions of drones for commercial package delivery. Nat. Commun. **9**, 02 (2018)
6. Dai, H., Wang, X., Liu, A.X., Ma, H., Chen, G.: Optimizing wireless charger placement for directional charging. In: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, pp. 1–9 (2017)
7. Kaul, S., Yates, R., Gruteser, M.: Real-time status: How often should one update? In: 2012 Proceedings IEEE INFOCOM, pp. 2731–2735 (2012)

# Resource Allocation for Full-Duplex Vehicular Communications in Overlay and Underlay Modes

Fang Qu[1], Liang Han[1,2(✉)], and Xiaocheng Wang[1,2]

[1] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
hanliang@tjnu.edu.cn

[2] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

**Abstract.** In this paper, we perform power control and spectrum allocation in full-duplex (FD) overlay and FD underlay modes to minimize the total transmit power of vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) links while ensuring the quality-of-service of each link, measured by outage probability. Power control in underlay mode is a non-convex optimization, thus we utilize the monotonicity and convexity of constraints to propose a double-nested binary search algorithm to solve the non-convex problem. Since the spectrum allocation problem of V2I and V2V links in overlay mode is also non-convex, we first discretize the bandwidth ratio coefficient and obtain the optimal value through traversal. The simulation results show the sum of the minimum transmit power of vehicles satisfying both V2I and V2V links outage probability constraints and the total ergodic capacity under this optimization both in overlay and underlay modes based on FD technique.

**Keywords:** Resource allocation · overlay · underlay · full-duplex · vehicular network

## 1 Introduction

A variety of works have been conducted in vehicular networks to support various applications for traffic safety, traffic efficiency, and passenger infotainment [1–3]. For different applications, the requirements for the transmission links that carry messages are also different. Generally speaking, the applications related to traffic safety require high reliability and low latency, and that related to traffic efficiency and infotainment frequently require to exchange of large amounts of data.

There are two modes of spectrum allocation for vehicle-to-vehicle (V2V) links, one is to occupy orthogonal spectrum, i.e., overlay mode, and the other is to reuse the spectrum of vehicle-to-infrastructure (V2I) link, i.e., underlay mode [4]. The overlay mode will not cause mutual interference, but its spectrum efficiency is relatively lower;


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 207–215, 2023.
https://doi.org/10.1007/978-981-99-2362-5_26

the underlay mode can greatly improve the spectrum utilization, but it will cause mutual interference between links.

Wireless terminals enabled by full-duplex (FD) are characterized by transmitting and receiving simultaneously over the same frequency band and thus FD offers the potential to double their spectral efficiency (SE) [5]. The large self-interference (SI) from a node's own transmission can be effectively attenuated by self-interference cancellation (SIC) [5, 5]. The introduction of FD in V2V links is beneficial to reduce latency and improve SE to alleviate the current situation of tight spectrum resources.

There are many existing related works concerned with resource management to optimize communication networks. In [7], a three-step scheme was proposed to maximize the overall network throughput while guaranteeing the quality of service (QoS) for both device-to-device (D2D) users and cellular users. Inspired by [7], the authors in [8] performed resource allocation (RA) based on slowly varying large-scale fading information of wireless channels and approached the optimization problem in two steps, and jointly considered the performance of V2I and V2V links. In [9], the authors developed a novel decentralized resource allocation mechanism for V2V communications based on deep reinforcement learning. The authors in [10] investigate the mode selection for FD-enabled D2D communications by maximizing the system SE while fulfilling the minimum rate requirements and maximum transmit power constraints for both cellular and D2D users.

Most of the existing papers on vehicular network RA are mostly based on half-duplex mode, and there are fewer optimizations in FD mode. In this paper, we consider both large-scale fading caused by path loss and small-scale fading caused by the Doppler effect and implement power control and spectrum allocation in overlay and underlay modes. By minimizing the total transmit power in a cluster while satisfying the QoS of V2I and V2V links, we reduce power consumption in a fast-fading vehicular environment, thereby effectively saving resources.

## 2 System Model

A V2V pair is formed by a pair of adjacent vehicle users denoted as $\mathrm{VUE}_1$ and $\mathrm{VUE}_2$, which form a cluster with a CUE. In this paper, we study resource allocation in a cluster, assuming that the total spectrum bandwidth is $W$. The V2I link composed of CUE and BS occupies $\mu W$ bandwidth, and the V2V links composed of the V2V pair occupy the remaining $(1-\mu)W$ bandwidth in overlay mode.

Due to the fast channel variations caused by high mobility in a vehicular environment, fast fading and slow fading caused by path loss both need to be taken into account. The channel power gain between CUE and BS can be expressed as

$$h_{C,B} = g_{C,B} \cdot \alpha_{C,B}, \tag{1}$$

where $g_{C,B}$ denotes fast fading channel power gain that obeys Rayleigh distribution, assuming to be exponentially distributed with unit mean, and $\alpha_{C,B}$ denotes path loss channel power gain. Similarly, we denote the channel power gain of the V2V pair as $h_{1,2}$ and $h_{2,1}$, the channel power gain from CUE to $\mathrm{VUE}_1$ and $\mathrm{VUE}_2$ as $h_{C,1}$ and $h_{C,2}$,

respectively, and the channel power gain from $\text{VUE}_1$ and $\text{VUE}_2$ to BS as $h_{1,B}$ and $h_{2,B}$, respectively.

Figure 1 shows the FD overlay mode of vehicular networks, in which V2V links of $\text{VUE}_1$ transmitting $\text{VUE}_2$ receiving and $\text{VUE}_2$ transmitting $\text{VUE}_1$ receiving share $(1-\mu)W$ bandwidth. There occurs SI at each VUE side, thus the received signal-to-interference-plus-noise-ratio (SINR) at BS for CUE can be written as

$$\gamma_c = \frac{P_C \alpha_{C,B} g_{C,B}}{\mu W N_0}, \tag{2}$$

and the received SINR at the $\text{VUE}_1$ and $\text{VUE}_2$ can be written as

$$\gamma_{2,1} = \frac{P_2 \alpha_{2,1} g_{2,1}}{P_1 \beta + (1-\mu) W N_0}, \gamma_{1,2} = \frac{P_1 \alpha_{1,2} g_{1,2}}{P_2 \beta + (1-\mu) W N_0}, \tag{3}$$

respectively, where $P_C$, $P_1$ and $P_2$ represent the transmit powers at CUE, $\text{VUE}_1$, and $\text{VUE}_2$, $N_0$ denotes the noise power spectral density, and $\beta$ denotes the SIC coefficient. The smaller $\beta (0 \le \beta \le 1)$ denotes the greater SI cancellation ability.

This paper focuses on minimizing the total transmit power of CUE and VUE while satisfying the outage probability constraints of V2V and V2I links. In FD overlay mode, the optimization problem can be expressed as

$$\min_{\{P_C, P_1, P_2\}} (P_C + P_1 + P_2) \tag{4}$$

$$\text{s. t. } \Pr\left\{\mu W \log_2(1+\gamma_c) \le R_0^c\right\} \le p_0^c \tag{4a}$$

$$\Pr\left\{(1-\mu) W \log_2\left(1+\gamma_{1,2}\right) \le R_0^v\right\} \le p_0^v, \Pr\left\{(1-\mu) W \log_2\left(1+\gamma_{2,1}\right) \le R_0^v\right\} \le p_0^v, \tag{4b}$$

$$0 \le P_C \le P_{\max}^C,\ 0 \le P_1 \le P_{\max}^V,\ 0 \le P_2 \le P_{\max}^V, \tag{4c}$$

where $\Pr\{\cdot\}$ represents probability operations, $R_0^c$ and $R_0^v$ denote the minimum capacity requirements of each V2I link and V2V link, $p_0^c$ and $p_0^v$ denote the tolerable outage probabilities of V2I link and V2V link, respectively, $P_{\max}^C$ and $P_{\max}^V$ represent the maximum transmit powers at CUE and $\text{VUE}_1/\text{VUE}_2$, respectively.

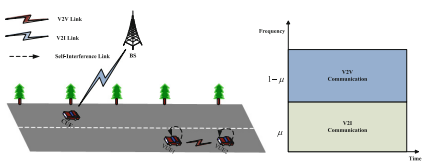



**Fig. 1.** FD + Overlay

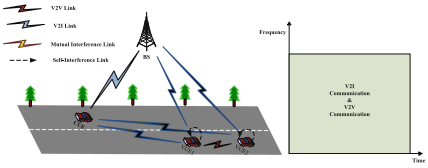



**Fig. 2.** FD + Underlay

Figure 2 shows FD underlay mode of vehicular communication, in which a V2I link and two V2V links share $W$ bandwidth. In this case, mutual interference occurs between V2I and V2V links, and SI occurs at each VUE side, thus the received SINR at the BS for the CUE can be written as

$$\gamma_c = \frac{P_C \alpha_{C,B} g_{C,B}}{P_1 \alpha_{1,B} g_{1,B} + P_2 \alpha_{2,B} g_{2,B} + WN_0}, \tag{5}$$

and the received SINR at the $\text{VUE}_1$ and $\text{VUE}_2$ can be written as

$$\gamma_{2,1} = \frac{P_2 \alpha_{2,1} g_{2,1}}{P_C \alpha_{C,1} g_{C,1} + P_1 \beta + WN_0}, \gamma_{1,2} = \frac{P_1 \alpha_{1,2} g_{1,2}}{P_C \alpha_{C,2} g_{C,2} + P_2 \beta + WN_0}. \tag{6}$$

In FD underlay mode, the optimization problem can be expressed as

$$\min_{\{P_C, P_1, P_2\}} (P_C + P_1 + P_2) \tag{7}$$

$$\text{s. t. } \Pr\left\{W \log_2(1 + \gamma_c) \le R_0^c\right\} \le p_0^c \tag{7a}$$

$$\Pr\left\{W \log_2\left(1 + \gamma_{1,2}\right) \le R_0^v\right\} \le p_0^v, \Pr\left\{W \log_2\left(1 + \gamma_{2,1}\right) \le R_0^v\right\} \le p_0^v, \tag{7b}$$

$$0 \le P_C \le P_{\max}^C,\ 0 \le P_1 \le P_{\max}^V,\ 0 \le P_2 \le P_{\max}^V. \tag{7c}$$

## 3 Power Control Scheme

In Sect. 2, we analyze the system models and express the optimization objectives and constraints of each mode. In this section, we will analyze the optimization problems according specifically to the formulas above.

To simplify the optimization, let $P_1 = P_2 = P_V$. Since $\alpha_{1,2}$ and $\alpha_{2,1}$ represent the path loss of the same path, thus $\alpha_{1,2} = \alpha_{2,1}$. Since $\text{VUE}_1$ and $\text{VUE}_2$ are adjacent, it can be considered as $\alpha_{1,B} \approx \alpha_{2,B}$, $\alpha_{C,1} \approx \alpha_{C,2}$.

### 3.1 Full-Duplex Overlay Mode

Putting formulas (2) (3) into constraints (4a) (4b), the outage probability constraint of the V2I link can be expressed as

$$1 - \exp\left(-\frac{\mu W N_0}{P_C \alpha_{C,B}}\left(2^{\frac{R_0^c}{\mu W}} - 1\right)\right) \le p_0^c, \tag{8}$$

and the outage probability constraints of the two V2V links formed by $\text{VUE}_1$ and $\text{VUE}_2$ are equal, which can be expressed as

$$1 - \exp\left(-\frac{(P_V \beta + (1-\mu) W N_0)}{P_V \alpha_{1,2}}\left(2^{\frac{R_0^v}{(1-\mu) W}} - 1\right)\right) \le p_0^v. \tag{9}$$

Inequalities (8) and (9) are further transformed into

$$\begin{aligned} P_C &\geq -\frac{\mu W N_0}{\alpha_{C,B}\ln\left(1-p_0^c\right)}\left(2^{\frac{R_0^c}{\mu W}}-1\right)=P_C^*, \\ P_V &\geq -\frac{(1-\mu)WN_0}{\alpha_{1,2}\ln\left(1-p_0^v\right)+\beta\left(2^{\frac{R_0^v}{(1-\mu)W}}-1\right)}\left(2^{\frac{R_0^v}{(1-\mu)W}}-1\right)=P_V^*. \end{aligned} \tag{10}$$

Combining constraints (4c), if exits $0 \leq P_C^* \leq P_{\max}^C$, $0 \leq P_V^* \leq P_{\max}^V$, then we can get

$$\min_{\{P_C,P_V\}}(P_C+P_1+P_2)=\min_{\{P_C,P_V\}}(P_C+2P_V)=P_C^*+2P_V^*. \tag{11}$$

In this case, the sum ergodic capacity of V2I and V2V links can be expressed as

$$\begin{aligned} &\mu W\mathbb{E}\left\{\log_2\left(1+\gamma_c\right)\right\}+(1-\mu)W\left(\mathbb{E}\left\{\log_2\left(1+\gamma_{1,2}\right)\right\}+\mathbb{E}\left\{\log_2\left(1+\gamma_{2,1}\right)\right\}\right) \\ &=\mu W\frac{e^{\frac{1}{a}}}{\ln 2}E_1\left(\frac{1}{a}\right)+2(1-\mu)W\frac{e^{\frac{1}{b}}}{\ln 2}E_1\left(\frac{1}{b}\right), \end{aligned} \tag{12}$$

where $\mathbb{E}\{\cdot\}$represent the expectation operation, $a=\frac{P_C^*\alpha_{C,B}}{\mu WN_0}$, $b=\frac{P_V^*\alpha_{1,2}}{P_V\beta+(1-\mu)WN_0}$, and $E_1(x)=\int_x^\infty\frac{e^{-t}}{t}dt$, the derivation process can refer to [8] [11, Eq. (3.352.4)].

### 3.2 Full-Duplex Underlay Mode

Putting formulas (5) (6) into constraints (7a) (7b), the outage probability constraint of the V2I link can be expressed as

$$\left(\frac{P_C\alpha_{C,B}}{P_C\alpha_{C,B}+\left(2^{\frac{R_0^c}{W}}-1\right)P_V\alpha_{1,B}}\right)^2\exp\left(-\frac{WN_0}{P_C\alpha_{C,B}}\left(2^{\frac{R_0^c}{W}}-1\right)\right)\geq 1-p_0^c, \tag{13}$$

and the outage probability constraints of the two V2V links formed by $\mathrm{VUE}_1$ and $\mathrm{VUE}_2$ are equal, which can be expressed as

$$1-\frac{P_V\alpha_{1,2}}{P_V\alpha_{1,2}+\left(2^{\frac{R_0^v}{W}}-1\right)P_C\alpha_{C,2}}\exp\left(-\frac{(P_V\beta+WN_0)}{P_V\alpha_{1,2}}\left(2^{\frac{R_0^v}{W}}-1\right)\right)\leq p_0^v. \tag{14}$$

Inequalities (13) and (14) are further transformed into

$$\begin{cases} P_V \le \dfrac{P_C\alpha_{C,B}}{\left(2^{\frac{R_0^c}{W}}-1\right)\alpha_{1,B}}\left(\sqrt{\dfrac{1}{1-p_0^c}\exp\left(-\dfrac{WN_0}{P_C\alpha_{C,B}}\left(2^{\frac{R_0^c}{W}}-1\right)\right)}-1\right) = G(P_C) \\ P_C \le \dfrac{P_V\alpha_{1,2}}{\left(2^{\frac{R_0^v}{W}}-1\right)\alpha_{C,2}}\left(\dfrac{1}{1-p_0^v}-1\right)\exp\left(-\dfrac{1}{\alpha_{1,2}}\left(2^{\frac{R_0^v}{W}}-1\right)\left(\beta+\dfrac{WN_0}{P_V}\right)\right) = F(P_V) \end{cases}. \tag{15}$$

Since the first and second derivatives of $G(P_C)$ and $F(P_V)$ are as follows

$$\begin{aligned} &\frac{\partial G}{\partial P_C} = \left(\frac{\alpha_{C,B}}{\left(2^{\frac{R_0^c}{W}}-1\right)\alpha_{1,B}}+\frac{WN_0}{2P_C\alpha_{1,B}}\right)\sqrt{\frac{1}{1-p_0^c}\exp\left(-\frac{WN_0}{P_C\alpha_{C,B}}\left(2^{\frac{R_0^c}{W}}-1\right)\right)} \\ &-\frac{\alpha_{C,B}}{\left(2^{\frac{R_0^c}{W}}-1\right)\alpha_{1,B}} > 0 \frac{\partial^2 G}{\partial P_C^2} = \frac{(WN_0)^2}{4(P_C)^3\alpha_{1,B}\alpha_{C,B}}\left(2^{\frac{R_0^c}{W}}-1\right) \\ &\sqrt{\frac{1}{1-p_0^c}\exp\left(-\frac{WN_0}{P_C\alpha_{C,B}}\left(2^{\frac{R_0^c}{W}}-1\right)\right)} > 0 \end{aligned} \tag{16}$$

$$\begin{aligned} &\frac{\partial F}{\partial P_V} = \left(\frac{\alpha_{1,2}}{\left(2^{\frac{R_0^v}{W}}-1\right)\alpha_{C,2}}+\frac{WN_0}{P_V\alpha_{C,2}}\right)\left(\frac{1}{1-p_0^v}-1\right) \\ &\exp\left(-\frac{1}{\alpha_{1,2}}\left(2^{\frac{R_0^v}{W}}-1\right)\left(\beta+\frac{WN_0}{P_V}\right)\right) > 0 \\ &\frac{\partial^2 F}{\partial P_V^2} = \frac{(WN_0)^2}{(P_V)^3\alpha_{C,2}\alpha_{1,2}}\left(2^{\frac{R_0^v}{W}}-1\right)\left(\frac{1}{1-p_0^v}-1\right) \\ &\exp\left(-\frac{1}{\alpha_{1,2}}\left(2^{\frac{R_0^v}{W}}-1\right)\left(\beta+\frac{WN_0}{P_V}\right)\right) > 0, \end{aligned} \tag{17}$$

which indicates that $G(P_C)$ and $F(P_V)$ are monotonically increasing convex functions with respect to $P_C$ and $P_V$, respectively. According to formula (7c) (16) (17), we can get the feasible region of FD underlay mode as shown in Fig. 3.

Let $G(P_C) = 0$, we can get $P_{\min}^c = -\frac{WN_0}{\alpha_{C,B}\ln(1-p_0^c)}\left(2^{\frac{R_0^c}{W}}-1\right)$, since $\lim_{P_V\to 0} F(P_V) = 0$, thus set $P_{\min}^v = 0$. We can find that if the feasible region is shown in (a) of Fig. 3, the minimum transmit powers of $P_C$ and $P_V$ satisfying the constraints are obtained at the intersection of curves $P_V = g(P_C)$ and $P_C = f(P_V)$. Conversely, if the feasible region is shown in (b) of Fig. 3, there does not exist optimal transmit power.

Aiming at the monotonicity and convexity of the curves $P_V = g(P_C)$, $P_C = f(P_V)$, we propose a double-nested binary search (DNBS) algorithm to find the optimal transmit

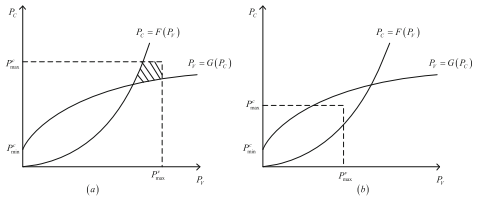


**Fig. 3.** The feasible region for (13a) (13b) (13c)

power. The algorithm is specifically expressed as follows: within the value range of $P_C$ and $P_V$, a two-layer binary search method nested inside and outside is used to traverse until the intersection of the two curves is found.

According to the algorithm, we can get the optimal transmit power in this mode. In this case, the sum ergodic capacity of V2I and V2V links can be expressed as

$$\begin{aligned}
&W\mathbb{E}\left\{\log_2\left(1+\gamma_C\right)\right\}+W\mathbb{E}\left\{\log_2\left(1+\gamma_{1,2}\right)\right\}+W\mathbb{E}\left\{\log_2\left(1+\gamma_{2,1}\right)\right\}\\
&=\frac{Wt_1}{\ln 2}\left(\frac{t_1}{(t_1-t_2)(t_1-t_3)}e^{\frac{1}{t_1}}E_1(\frac{1}{t_1})+\frac{t_2}{(t_2-t_1)(t_2-t_3)}e^{\frac{1}{t_2}}E_1(\frac{1}{t_2})+\frac{t_3}{(t_3-t_1)(t_3-t_2)}e^{\frac{1}{t_3}}E_1(\frac{1}{t_3})\right)\\
&+\frac{2W}{\ln 2}\frac{t_4}{t_4-t_5}\left(e^{\frac{1}{t_4}}E_1\left(\frac{1}{t_4}\right)-e^{\frac{1}{t_5}}E_1\left(\frac{1}{t_5}\right)\right),
\end{aligned} \tag{18}$$

where $t_1=\frac{P_C\alpha_{C,B}}{WN_0}$, $t_2=\frac{P_V\alpha_{1,B}}{WN_0}$, $t_3=\frac{P_V\alpha_{2,B}}{WN_0}$, $t_4=\frac{P_V\alpha_{1,2}}{WN_0+P_V\beta}$, $t_5=\frac{P_C\alpha_{C,2}}{WN_0+P_V\beta}$.

## 4 Spectrum Allocation in Overlay Mode

In overlay mode, the V2I link composed of CUE and BS occupies $\mu W$ bandwidth. To minimize the total transmit powers of vehicles in each cluster, the selection of parameter $\mu$ is also crucial. The total transmit power in FD overlay mode is a non-convex function with respect to $\mu$, so to simplify the optimization process, we discretize the value of $\mu$. The optimal $\mu$ with the minimum total transmit power is obtained by traversing, specifically expressed as

$$\mu_1^*=\min_{\mu\in\{0.1,0.2,\ldots,0.9\}}\left\{-\frac{\mu WN_0\left(2^{\frac{R_0^c}{\mu W}}-1\right)}{\alpha_{C,B}\ln\left(1-p_0^c\right)}-2\frac{(1-\mu)WN_0\left(2^{\frac{R_0^v}{(1-\mu)W}}-1\right)}{\alpha_{1,2}\ln\left(1-p_0^v\right)+\beta\left(2^{\frac{R_0^v}{(1-\mu)W}}-1\right)}\right\} \tag{19}$$

## 5 Simulation Results

We set carrier frequency $f=2$ GHz, cell radius $r=500$ m, BS antenna height $H_{BS}=25$ m, BS antenna gain $h_{BS}=8$ dBi, BS receiver noise figure $N_{BS}=5$ dB, the distance

between BS and highway $D = 35$ m, vehicle antenna height $H_v = 1.5$ m, vehicle antenna gain $h_v = 3$ dBi, vehicle receiver noise figure $N_v = 9$ dB, $W = 1$ Hz, $R_0^v = 3$ bps, $p_0^c = 0.01$, $P_{\max}^c = P_{\max}^v = 23$ dBm, $N_0 = -114$ dBm. The slow fading parameters can refer to [12]. Let the positions of BS, CUE, $\text{VUE}_1$, and $\text{VUE}_2$ be (0, 0), (10, 37), (300, 41) and (303, 41) respectively.

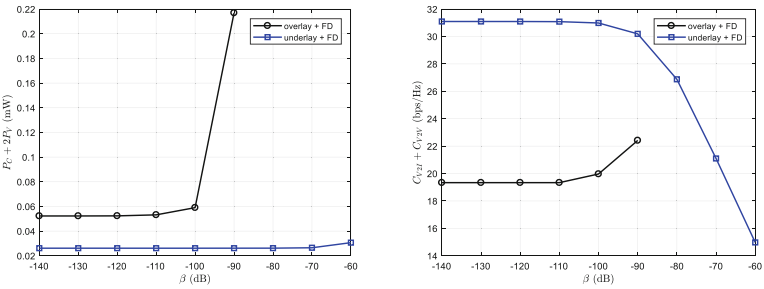



**Fig. 4.** The sum of the minimum transmit power (a) and the sum of ergodic capacity (b) under different SIC coefficients $\beta$, $R_0^c = 2$ bps, $p_0^v = 0.001$.

Figure 4 shows the greater the SIC capability, the total transmitted power in a cluster decreases, which can be interpreted that the smaller SI is, the smaller $P_V$ required for V2V link outage probability is. At the same time, FD underlay mode outperforms FD overlay in terms of power consumption and ergodic capacity when the SI cancellation capability is larger. Figure 5 performs when the V2V link outage probability is reduced, the required VUE transmit power is decreased, thus the total minimum transmission power is reduced. On the other hand, FD underlay is better than FD overlay when the outage probability is relatively small.

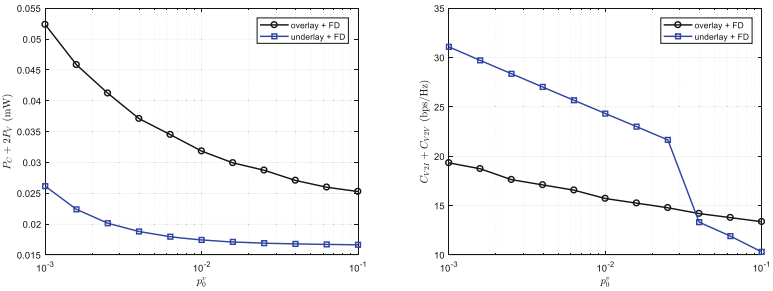



**Fig. 5.** The sum of the minimum transmit power (a) and the sum of ergodic capacity (b) under different $p_0^v$, $\beta = -120$ dB, $R_0^c = 2$ bps.

## 6 Conclusion

In this paper, we consider both the large-scale fading caused by path loss and the small-scale fading caused by the Doppler effect and implement power control and spectrum allocation in the two different modes based only on the CSI of path loss, including FD overlay mode and FD underlay mode. And we attempt to minimize the total transmit power of V2I and V2V links while ensuring the QoS of each link. The limitation of this paper is that we only consider the resource allocation among vehicles in one cluster, and do not simulate the problem of multi-cluster resource allocation in which multiple vehicles are randomly distributed, which is also the direction of our future efforts.

**Acknowledgment.** This work was supported by the National Natural Science Foundation of China (62001327, 61701345), and the Natural Science Foundation of Tianjin (18JCZDJC31900).

## References

1. Karagiannis, G., Altintas, O., Ekici, E., et al.: Vehicular networking: a survey and tutorial on requirements, architectures, challenges, standards and solutions. IEEE Commun. Surv. Tutorials **13**(4), 584–616 (2011)
2. Gyawali, S., Xu, S., Qian, Y., et al.: Challenges and solutions for cellular based V2X communications. IEEE Commun. Surv. Tutorials **23**(1), 222–255 (2020)
3. MacHardy, Z., Khan, A., Obana, K., et al.: V2X access technologies: regulation, research, and remaining challenges. IEEE Commun. Surv. Tutorials **20**(3), 1858–1877 (2018)
4. Lin, X., Andrews, J.G., Ghosh, A.: Spectrum sharing for device-to-device communication in cellular networks. IEEE Trans. Wireless Commun. **13**(12), 6727–6740 (2014)
5. Sabharwal, A., Schniter, P., Guo, D., et al.: In-band full-duplex wireless: challenges and opportunities. IEEE J. Sel. Areas Commun. **32**(9), 1637–1652 (2014)
6. Duarte, M., Dick, C., Sabharwal, A.: Experiment-driven characterization of full-duplex wireless systems. IEEE Trans. Wireless Commun. **11**(12), 4296–4307 (2012)
7. Feng, D., Lu, L., Yuan-Wu, Y., et al.: Device-to-device communications underlaying cellular networks. IEEE Trans. Commun. **61**(8), 3541–3551 (2013)
8. Liang, L., Li, G.Y., Xu, W.: Resource allocation for D2D-enabled vehicular communications. IEEE Trans. Commun. **65**(7), 3186–3197 (2017)
9. Ye, H., Li, G.Y., Juang, B.H.F.: Deep reinforcement learning based resource allocation for V2V communications. IEEE Trans. Veh. Technol. **68**(4), 3163–3173 (2019)
10. Han, L., Zhang, Y., Li, Y., et al.: Spectrum-efficient transmission mode selection for full-duplex-enabled two-way D2D communications. IEEE Access **8**, 115982–115991 (2020)
11. Gradshteyn, I.S., Ryzhik, I.M.: Table of Integrals, Series, and Products. Elsevier (2007)
12. Kyosti, P.: WINNER II channel models. IST Tech. Rep. IST-4-027756 WINNER II D1. 1.2 V1. 2 (2007)

# Reinforcement Learning Based Resource Allocation for Dedicated Full-Duplex V2V Communications

Keshan Zheng[1], Liang Han[1,2(✉)], and Yupeng Li[1,2]

[1] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
hanliang@tjnu.edu.cn

[2] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

**Abstract.** In this paper, we investigate resource allocation for full-duplex vehicle-to-vehicle (V2V) communication based on deep reinforcement learning in the case of dedicated resources. According to the deep reinforcement learning mechanism, a pair of V2V links is regarded as an agent, and the other parts are regarded as the environment. Each agent can interact with the environment independently and tends to meet the requirements of V2V information transmission according to the set rewards while ensuring that the multiplexing channel works properly. From the simulation results, under the self-interference cancellation of 120 dB, the overall capacity of full-duplex V2V is higher than that of half-duplex.

**Keywords:** Full-duplex V2V · Resource Allocation · Deep Reinforcement Learning

## 1 Introduction

In recent years, with the continuous advancement of urbanization, the number of vehicles continues to increase, and traffic accidents occur frequently. Efficient and orderly traffic control has become an urgent problem to be solved. Currently, the V2V [1–3] is an effective way to solve this problem.

With the increase of the number of mobile communication terminals, a large number of devices access to the spectrum, and the spectrum resources become extremely tight. In order to improve the spectrum efficiency and reliability of V2V communication, we use reinforcement learning (RL) to solve the problem of vehicle resource allocation [4–7] in the case of V2V dedicated resources.

In previous work, most of the resource allocation on V2V communication is based on half-duplex mode, and the spectrum utilization rate is low. For this reason, we take full-duplex [8–10] as the starting point and study the resource allocation in full-duplex V2V communication, and use reinforcement learning to complete resource allocation. Therefore, the innovation of this paper is to use deep reinforcement learning to resolve resource allocation in full-duplex V2V communication and improve spectrum utilization.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 216–222, 2023.
https://doi.org/10.1007/978-981-99-2362-5_27

## 2 System Model

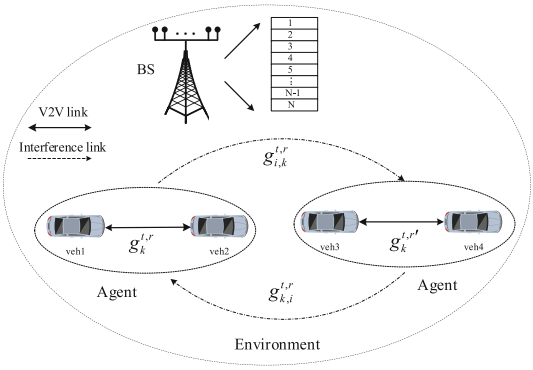



**Fig. 1.** System model

As shown in Fig. 1, in the coverage area of the base station, there are K active V2V (Vehicle-to-vehicle) pairs, expressed as $\mathcal{K} = \ominus\{1, ..., K\}$, N resource blocks have been allocated in the base station, expressed as $\mathcal{N} = \ominus\{1, ..., N\}$, $(N \leq K)$, , each pair Both V2V select a frequency band for communication from N RBs (Resource Blocks). The $k^{th}$ V2V pair, denoted as $D_k, k \in K$. $D_k^t$ and $D_k^r$ represent the transmitter and receiver of V2V pair $D_k$, respectively. The respective communication goals between vehicles have been determined, and full-duplex communication is performed between V2V through RB.

We assume that the transmit power of full-duplex V2V is $P^d$. On resource block $n$ in the BS, the instantaneous SINR of the signal received by the receiver $D_k^r$ of the $k^{th}$ pair of V2V (receiving the signal $D_k^t$ from the V2V transmitter) is:

$$\xi_k^{d,r} = \frac{P^d g_k^{t,r}}{\sum\limits_{i \in D_n, i \neq k} \left( P^{d'} g_{i,k}^t + P^{d'} g_{i,k}^{t\ '} \right) + P^d \theta + \sigma^2} \tag{1}$$

where: $\sum\limits_{i \in D_n, i \neq k} \left( P^{d'} g_{i,k}^t + P^{d'} g_{i,k}^{t\ '} \right)$ denote the interference power introduced by sharing the spectrum with other V2V and $D_n$ denotes the set of V2V pairs assigned to resource block $n$. $g_{i,k}^t$ and $g_{i,k}^{t\ '}$ are the channel gain of other V2Vs that reuse the same frequency band interferes with the current V2V. $P^d \theta$ denote the interference power after self-interference suppression, and $\sigma^2$ is the noise power. Hence the full-duplex capacity of the $k^{th}$ VUE can be expressed as:

$$\mathcal{R}_k^c = W \left[ \log\left(1 + \xi_k^{d,t}\right) + \log\left(1 + \xi_k^{d,r}\right) \right] \tag{2}$$

where $W$ is the bandwidth of each RB.

Through the above theoretical analysis, our goal is to reduce the mutual interference between V2V links as much as possible by controlling the power while reusing the same frequency band, while ensuring the communication delay of V2V. Here, it is noted that the power allocation problem formulated in (2) is an NP-hard combinatorial optimization problem with nonlinear constraints, which corresponds a forbidden complexity. Thus, to resolve this issue, we investigate DRL-based resource allocation in full-duplex V2V communication.

## 3 Deep Reinforcement Learning for Resource Allocation of Full-Duplex V2V Communication

In our system model, a V2V link corresponds to an agent, and everything beyond the particular V2V link is regarded as the environment. When V2V link communication is to be carried out, the selection of spectrum and power during V2V link communication is made based on their local observations. We make the environment and the agent interact with each other continuously to obtain rewards and maximize the expected cumulative discounted rewards.

As shown in Fig. 2, at each time $t$, the agent observes a state $s_t$ from the state space $S$, and accordingly implements an action $a_t$ from the action space $A$ to choose sub-band and transmission power. Based on the actions taken by the agents, the environment transits to a new state, $s_{t+1}$, and the agent receives a reward, $r_t$, from the environment.

Assuming that after the end of the interaction, the rewards received in each round in this process are as follows: $r_t, r_{t+1}, r_{t+2}, ..., r_n$. We expect that the sum of all rewards from time $t$ to the end is:

$$R_t = r_t + \mu r_{t+1} + \mu^2 r_{t+2} + ... + \mu^{n-t} r_n = r_t + \mu R_{t+1} \tag{3}$$

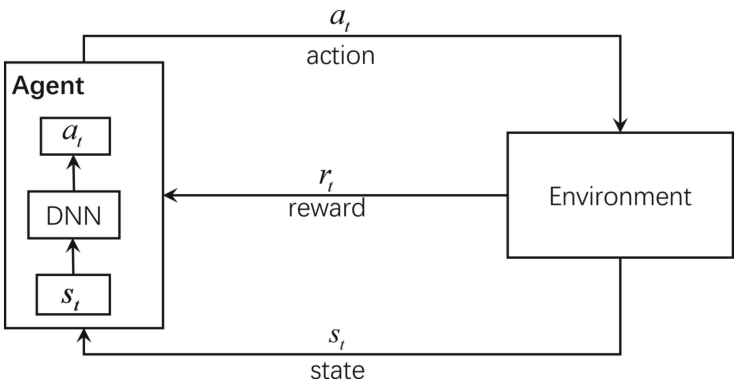



**Fig. 2.** Reinforcement learning process diagram

where $\mu$ is called the discount factor, and we expect that the value $R_t$ can be maximized. In our model, when the agent interacts with the environment, we express the reward from the environment to the agent in the following form:

$$r_t = \lambda_d \sum_{k \in K} C_k^v - \lambda_p (T_0 - U_t) \tag{4}$$

where $T_0$ is the maximum tolerable latency, $\lambda_d$ and $\lambda_p$ are weights. The reward is determined by the capacities of V2V links and the latency constraints of the corresponding V2V link. The largest sum of discount rewards is:

$$R_t = \sum_{n=0}^{\infty} \mu^n r_{t+n} \tag{5}$$

where $\mu \in (0, 1)$ is the discount factor. Therefore, we need to train the agent to make it act based on a certain strategy $\pi$, so that the final $R_t$ is the largest. In reinforcement learning, the action-value function is defined as follows:

$$Q_\pi(s_t, a_t) = E[R_t | s_t, a_t] \tag{6}$$

whose calculated value is denoted as $q$. Since the result also depends on the strategy $\pi$, further maximize both sides of Eq. (7):

$$Q^*(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t) \tag{7}$$

Thus, the final $Q$ value will be independent of the influence of strategy $\pi$. Similarly, the above transformation is also performed on the Eq. (8) to obtain:

$$Q(s_t, a_t) = r_t + \mu \cdot Q(s_{t+1}, a_{t+1}) \tag{8}$$

The value of $q$ can be calculated by using the equation and the real-time rewards obtained in each round. We use a neural network, to simulate a function to implicitly calculate $Q$ each round. The neural network can be trained using the DQN algorithm.

## 4 Simulation Results

The simulation environment is as follows: a BS is located in the center of Manhattan streets simulated by 3GPP TR 3.855 and covers the entire street area. Vehicles are randomly generated in lanes around the street according to the spatial Poisson distribution, and each vehicle communicates with the four closest vehicles. The detailed parameters can be found in Table 1.

In Fig. 3, we plot the relationship between the number of vehicles, self-interference suppression and the sum of V2V link capacity. As can be seen from Fig. 3, when the communication time used for full-duplex is half of half-duplex, from a horizontal perspective, whether it is full-duplex or half-duplex, the sum capacity increases with the increase of the number of vehicles. From vertical, it can be seen that with a fixed number of vehicles, the full-duplex V2V and capacity will go from lower than half-duplex

**Table 1.** Simulation Parameters

| Parameter | Value |
|---|---|
| Blocks of resource $N$ | 20 |
| Number of V2V links $K$ | [20, 40, 60, 80, 100] |
| Carrier frequency | 2 GHz |
| Bandwidth per channel | 1.5 MHz |
| BS antenna height | 25 m |
| BS antenna gain | 8 dBi |
| BS receiver noise figure | 5 dB |
| Vehicle antenna height | 1.5 m |
| Vehicle antenna gain | 3 dBi |
| Vehicle speed | 36 km/h |
| Neighbor distance threshold | 150 m |
| Number of lanes | 3 in each direction (12 in total) |
| Latency constraint for V2V link. | 100 ms |
| V2V transmit power $P^d$ | [23,10,5] dBm |
| Noise power $\sigma^2$ | $-114$ dBm |
| self-interference cancellation coefficient $\beta$ | $[-90, -100, -110, -120]$ dB |

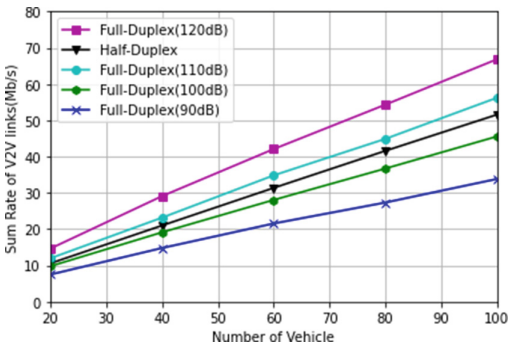



**Fig. 3.** Sum capacity versus the number of vehicles and self-interference suppression

V2V and capacity to higher than half-duplex V2V and capacity as self-interference suppression increases.

Figure 4 is drawn on the basis of Fig. 3, depicting the relationship between the ratio of full-duplex V2V and capacity to half-duplex V2V and the sum capacity, the number of vehicles and self-interference cancellation. According to Fig. 4, it can be seen from the horizontal that with the increase in the number of vehicles, the ratio of full-duplex V2V and capacity to half-duplex V2V and capacity fluctuates in a small range, and tends

to be relatively stable as a whole. It can be seen from the longitudinal direction that when the number of vehicles is fixed, the ratio increases with the increase of self-interference suppression, and when the self-interference suppression is less than 110 dB, the ratio is less than 1, which means that the performance of full-duplex is not as good as that of half-duplex at this time. When the self-interference suppression is greater than 110 dB, the ratio is between 1 and 1.4, which means that the performance of full-duplex V2V is better.

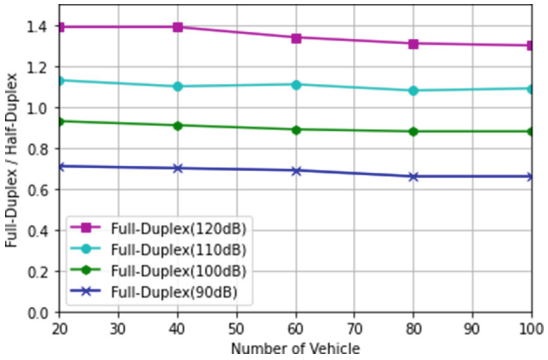



**Fig. 4.** Performance comparison between full-duplex V2V and half-duplex V2V

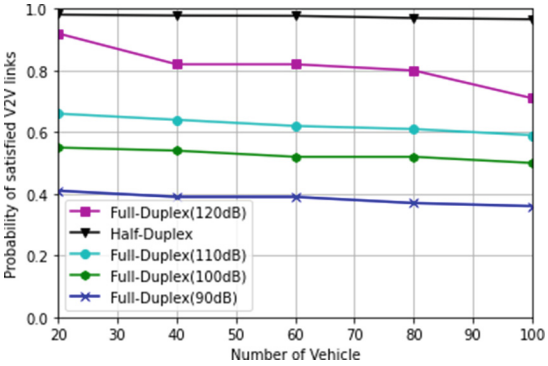



**Fig. 5.** Probability of satisfied V2V links versus the number vehicles

Figure 5 depicts the influence of different self-interference suppression and the number of vehicles on the probability of successful V2V communication. On the whole, the probability of successful V2V communication in full-duplex is less than that in half-duplex. Specifically, according to Fig. 5, from a horizontal point of view, the probability of successful V2V communication decreases gradually with the increase in the number of vehicles. Longitudinally, when the number of vehicles is fixed, with the increase of self-interference suppression, the probability of successful communication of full-duplex V2V gradually increases, and when the self-interference suppression is too small, communication is almost impossible.

## 5 Conclusion

The main content of this paper is to use deep reinforcement learning to realize the selection of full-duplex V2V communication resources, so that it can meet the transmission requirements of periodic security information between V2V. In addition, we focus on exploring the impact of the number of vehicles and different self-interference suppression on full-duplex V2V communication. Therefore, in general, in the case of satisfying certain performance requirements, the selection of self-interference suppression should be comprehensively considered when conducting full-duplex V2V communication.

**Acknowledgment.** This work was supported by the National Natural Science Foundation of China (61901301, 61701345), and the Natural Science Foundation of Tianjin (18JCZDJC31900).

## References

1. Mei, J., Zheng, K., Zhao, L., Teng, Y., Wang, X.: A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems. IEEE Trans. Wireless Commun. **17**(6), 3850–3860 (2018)
2. Abbas, F., Fan, P., Khan, Z.: A novel low-latency V2V resource allocation scheme based on cellular V2X communications. IEEE Trans. Intell. Transp. Syst. **20**(6), 2185–2197 (2018)
3. Liang, L., Li, G.Y., Xu, W.: Resource allocation for D2D-enabled vehicular communications. IEEE Trans. Commun. **65**(7), 3186–3197 (2017)
4. Ye, M., Tianqing, C., Wenhui, F.: A single-task and multi-decision evolutionary game model based on multi-agent reinforcement learning. J. Syst. Eng. Electron. **32**(3), 642–657 (2021)
5. Toan, N.D., Gon-Woo, K.: Environment exploration for mapless navigation based on deep reinforcement learning. In: 2021 21st International Conference on Control, Automation and Systems (ICCAS), pp. 17–20 (2021)
6. Ye, H., Li, G.Y., Juang, B.F.: Deep reinforcement learning based resource allocation for V2V communications. IEEE Trans. Veh. Technol. **68**(4), 3163–3173 (2019)
7. Luong, N.C., et al.: Applications of deep reinforcement learning in communications and networking: a survey. IEEE Commun. Surv. Tutorials **21**(4), 3133–3174 (2019)
8. Roychowdhury, D., Moallemi, S., Ozev, S., Kitchen, J.: Self-interference signal path characterization in full-duplex transceivers using built-in self-test. In: IEEE Radio and Wireless Symposium (RWS), pp. 16–19 (2020)
9. Sepanek, R., Hickle, M., Stuenkel, M.: In-band full-duplex self-interference canceller augmented with Bandstop-Configured Resonators. In: IEEE/MTT-S International Microwave Symposium (IMS), pp. 1199–1202 (2020)
10. Wu, D., Sun, Y.-X., Wang, B., Lian, R.: A high isolation circularly polarized antenna for full-duplex applications. In: Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC), pp. 360–361 (2021)

# Deep Learning Based Resource Allocation for Full-Duplex-Enabled Two-Way Device-To-Device Communications

Xiaolei Tian[1], Yi Gao[1,2], and Liang Han[1,2](✉)

[1] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
hanliang@tjnu.edu.cn

[2] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

**Abstract.** In recent years, device-to-device (D2D) communications and full-duplex (FD) communications, which can improve the spectrum efficiency (SE) of mobile communications, have been widely studied. FD-enabled two-way D2D communications, which integrate FD into D2D communications, can further improve the SE. Traditionally, iterative algorithms that converge to a local optimum are employed to solve the SE maximization problem of FD-enabled two-way D2D communication systems. However, its high computational complexity has stimulated the need for lower-complexity approximations at the expense of performance. Inspired by the success of deep learning in the balance between complexity and performance, we propose a novel application of the deep neural network (DNN) model for FD-enabled two-way D2D communication systems. The main idea consists of mapping a limited number of iterations of the concave-convex procedure (CCCP) algorithm into trainable neural network layers. As opposed to traditional iterative power allocation, we train the DNN to learn the nonlinear relation between the channel realizations and the corresponding power allocation schemes based on the CCCP algorithm. Extensive simulation results demonstrate that the DNN can provide a good approximation of the iterative CCCP algorithm while reducing the computational overhead significantly.

**Keywords:** Full-duplex · Deep learning · Two-way D2D communications · Spectrum efficiency · Concave-convex procedure algorithm

## 1 Introduction

In the face of the rapidly growing volume of data services, the problem of insufficient wireless resources in traditional cellular networks is becoming increasingly prominent. To improve the spectrum efficiency (SE) of cellular networks, many novel techniques are being studied, among which device-to-device (D2D) communications and full-duplex (FD) communications have been widely studied [1, 2]. In recent years, some researchers have integrated FD into D2D communications which can further improve the SE and


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 223–230, 2023.
https://doi.org/10.1007/978-981-99-2362-5_28

reduce end-to-end latency. Existing studies have shown that an FD-enabled D2D communications have many advantages, such as increased SE [3–5] and improved network throughput [6, 7].

To improve the data transmission rate, it is important to optimize the power allocation for FD-enabled two-way D2D communication systems. In [8] the authors transformed the optimization problem into a difference of convex functions (D.C.) programming, and then proposed a concave-convex procedure (CCCP) algorithm to maximize SE. However, most of the works are mainly based on traditional methods to obtain optimal power allocation. The higher complexity of the traditional algorithms often means consuming a large amount of computational time.

Deep learning has been successfully applied to many fields, including speech recognition, automatic machine translation, and autonomous driving. Due to its characteristics, many researchers have tried to use DNN to solve wireless communication problems. Numerous studies have found that deep learning algorithms can not only achieve excellent performance in wireless communications but also significantly reduce computational complexity. Real-time power allocation using deep learning algorithms in real systems becomes a possibility. In [9], the authors use deep learning to solve critical signal processing problems, which shows that deep learning algorithms outperform traditional algorithms in terms of time reduction and computational complexity. To the best of our knowledge, there are few studies on power allocation schemes based on DNN methods. In this paper, we study deep learning algorithms for FD-enabled two-way D2D communication systems.

The rest of this paper is organized as follows. In Sect. 2, we introduce the system model including the system configuration, and formulate the SE maximization problem for the FD-enabled two-way D2D communication systems. In Sect. 3, we present the architecture of the system, the network structure, the process of data generation, and the training phase of DNN. In Sect. 4, we give the parameter selection and simulation results of the deep neural network to prove our hypotheses. Section 5 concludes the paper.

## 2 System Model

In this section, we consider FD-enabled two-way D2D communications with a pair of DUs ($DU_1$ and $DU_2$) sharing the uplink spectrum resource with one CU, as showed in Fig. 1. We denote the channels of the CU-BS, CU-$DU_1$, CU-$DU_2$, $DU_1$-BS, $DU_2$-BS, $DU_1$-$DU_2$ and $DU_2$-$DU_1$ links as $h_{cb}$, $h_{c1}$, $h_{c2}$, $h_{1b}$, $h_{2b}$, $h_{12}$ and $h_{21}$, respectively.

We assume all the channels are frequency-flat and quasi-static, and all the channel state information is perfectly known at the base station (BS). There has been a lot of work on self-interference cancellation techniques, but in practice self-interference cannot be eliminated. We assume the residual self-interference is subject to the complex Gaussian distribution, which can be considered the worst-case assumption about the interference. The problem of maximizing SE for the uplink FD-enabled two-way D2D communication system can be modeled as [8]:

$$
\begin{aligned}
&\max SE = \frac{1}{W}(R_c + R_1 + R_2) \\
&s.t. R_c \geq W \log_2(1 + \frac{P_c|h_{cb}|^2}{P_1|h_{1b}|^2 + P_2|h_{2b}|^2 + WN_0}) \geq R_c^{\min} \\
&R_1 \geq W \log_2(1 + \frac{P_2|h_{21}|^2}{P_c|h_{c1}|^2 + \beta P_1 + WN_0}) \geq R_1^{\min} \\
&R_2 \geq W \log_2(1 + \frac{P_1|h_{12}|^2}{P_c|h_{c1}|^2 + \beta P_2 + WN_0}) \geq R_2^{\min} \\
&0 < P_c \leq P_c^{\max}, 0 < P_1, P_2 \leq P_d^{\max}
\end{aligned} \tag{1}
$$

where $P_c$, $P_1$ and $P_2$ denote the transmit power of the CU, $DU_1$, and $DU_2$, respectively, $R_c^{\min}$, $R_1^{\min}$ and $R_2^{\min}$ denote the minimum rate requirement for CU, $DU_1$ and $DU_2$, respectively, $\beta$ is a constant that reflects the self-interference cancellation ability [8], and $N_0$ denotes the one-sided power spectral density of the additive white Gaussian noise (AWGN). We assume the maximum transmit power of the CU and DUs are $P_c^{\max}$ and $P_d^{\max}$, respectively, and the channels consist of small-scale and large-scale fading [4].

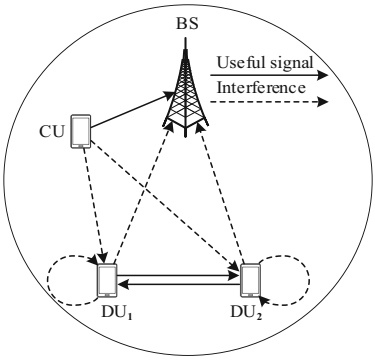



**Fig. 1.** System model

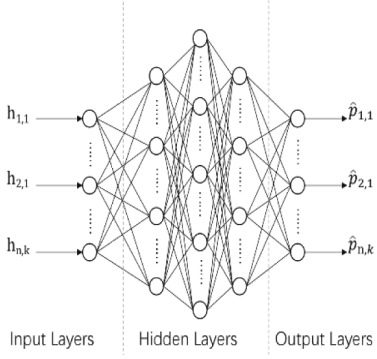



**Fig. 2.** The architecture of DNN

## 3 Proposed Deep Neural Network Approach

In this section, we present deep learning algorithms for the SE maximization problem. From the traditional algorithmic point of view, the CCCP algorithm can be treated as an unknown relationship between the channel realization and the corresponding power allocation. Deep neural networks can learn this unknown relationship very well. Although the ability of DNNs to learn this unknown relationship cannot be given a concrete explanation at the moment, people like to use DNNs, which are considered 'black boxes', to solve practical problems.

### 3.1 Making the Dataset

This paper aims to demonstrate that the DNN algorithm can be applied to FD-enabled two-way D2D communication systems with different maximum transmit power. Later we will verify the results by comparing the performance of the DNN algorithm with that of the conventional CCCP algorithm for different maximum transmit powers. The different maximum transmission powers can be expressed as $P_{\max 1}, P_{\max 2}, \ldots, P_{\max i}$. To ensure the rigor of the comparison, we pass in the traditional CCCP algorithm and the DNN algorithm in which the channel implementation is the same.

Because deep learning requires a large dataset of different instances to train a DNN, we need to prepare a large amount of training data and the corresponding labels. In the first step, a bulk channel matrix $H^{(t)}$ is generated, where $t$ is the index of the training samples. The second step is to use the conventional CCCP algorithm to obtain the optimal power allocation matrices $P^{(t)}_{\max 1}, P^{(t)}_{\max 2}, \ldots, P^{(t)}_{\max i}$, which are the labels of the corresponding $H^{(t)}$, respectively. For convenience, we use $P^{(t)}$ to denote the label at different maximum transmitting powers. Thus, $[H^{(t)}, P^{(t)}]$ denotes the $t$-th sample.

### 3.2 The Architecture of Neural Network

Once the dataset is produced, from an artificial intelligence perspective, we need to design a supervised learning method to approximate the CCCP algorithm to maximize the SE in the FD-enabled two-way D2D communication system. As shown in Fig. 2, we propose a fully connected neural network that consists of an input layer, multiple hidden layers and an output layer. The channel gains in the model diagram are the input to the DNN and the optimal power allocation schemes are the output of the DNN. It is worth noting that since this is a regression problem in machine learning, the output of the DNN should be a continuous value. To increase the learning capability of the DNN, we apply rectified linear units (ReLU) as the activation function of the hidden layer.

### 3.3 Training and Testing the DNN

We train the DNN using a normalized training dataset and normalized labels. The process mainly consists of forwarding propagation and back-propagation. The purpose of forwarding propagation is to calculate the error value of the DNN. In addition, back-propagation updates the weights of the DNN by reducing the error values so that the target power allocation results obtained from the CCCP algorithm and the current DNN output are similar. The training process can be briefly described as shown in Fig. 3. We call the function consisting of all error values a loss function. The hidden loss function is represented as

$$
\begin{aligned}
&loss = E[loss_{mse} + loss_{const}] \\
&= E[\lambda_1(\sum (\hat{P_c} - P_c)^2 + \sum (\hat{P_1} - P_1)^2 + \sum (\hat{P_2} - P_2)^2) + \lambda_2 \sum \mathrm{ReLU}(\hat{P_c} - P_c^{\max}) \\
&+\lambda_3 \sum \mathrm{ReLU}(\hat{P_1} - P_d^{\max}) + \lambda_3 \sum \mathrm{ReLU}(\hat{P_2} - P_d^{\max})]
\end{aligned}
\tag{2}
$$

where $\hat{P}_c$, $\hat{P}_1$, and $\hat{P}_2$ represents the output of the DNN.

The loss function consists of two parts, the first part is the mean square error $loss_{mse}$ between the DNN and the labeled output, and the second part is the constraint error $loss_{const}$.

The binding factors $\lambda_1$, $\lambda_2$ and $\lambda_3$ are used to balance the $loss_{mse}$ and $loss_{const}$ to ensure that the DNN can be trained well enough. Unlike other papers where only DNN and label errors are estimated, we think about the effect of constraints on the neural network in the loss function.

We experimented extensively to select suitable training parameters for the DNN. Furthermore, the training process was different when different learning rates were used. By testing different learning rates and batches, appropriate learning rates and batch sizes were selected based on the validation error of the previous 300 times DNN training sessions, as shown in Figs. 4 and 5 respectively.

Once we have optimized the weights of the DNN, we will obtain a neural network with good performance, and next we need to verify the performance of the DNN algorithm using a test dataset. To reduce the impact of chance on the performance evaluation, we average over multiple optimal power allocation schemes.

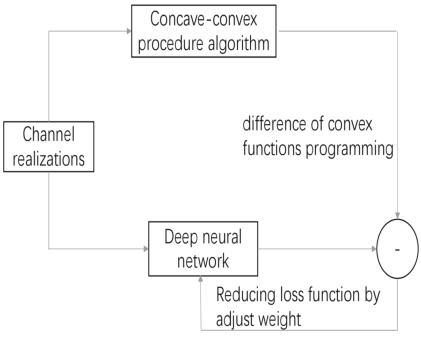



**Fig. 3.** The training flowchart

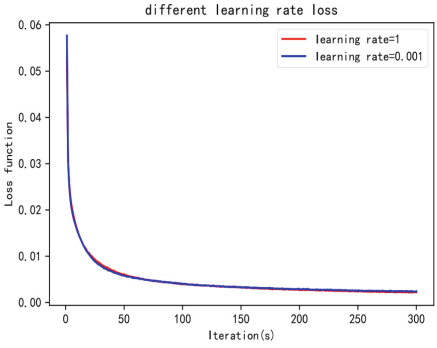



**Fig. 4.** Learning rate selection

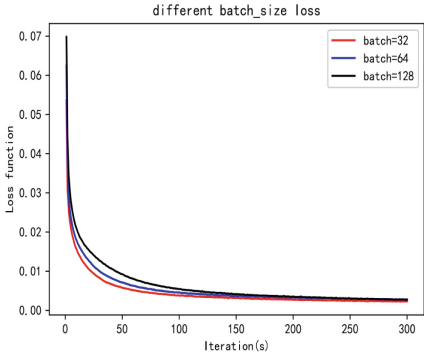



**Fig. 5.** Learning rate selection

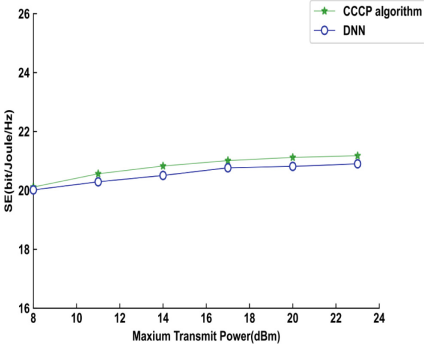



**Fig. 6.** Performance comparison

## 4 Numerical Results

In this section, we use the above channel modality to generate a 60000 samples dataset, with 50000 samples as the training dataset and 10000 as the test dataset. As the neural network is trained by reading the datasets sequentially, we randomly disrupt the order of the training datasets during training to reduce the impact of the fixed position information of the datasets on the neural network.

We consider the scenario of a single base station, a cellular user and a pair of D2D users. A new fully connected deep neural network algorithm has been developed. This network structure contains the following parts, an input layer with 7 nodes, three hidden layers with 64, 128 and 64 nodes, and an output layer with 3 nodes.

Through extensive experiments, we have come up with the most suitable batch size and learning rate of 32 and 0.001 respectively. We also found that by setting the balance factors $\lambda_1$, $\lambda_2$ and $\lambda_3$ to 1, 0.001 and 0.001 respectively, the DNN could achieve better performance. We give some numerical parameters for simulation, as shown in Table 1.

We measured the performance of the power allocation scheme of the newly proposed DNN algorithm and the power allocation scheme of the conventional CCCP algorithm as shown in Fig. 6. It is well known that the dataset for the DNN method is derived from the traditional CCCP algorithm, so the DNN scheme results cannot exceed the CCCP scheme. It is easy to observe from Fig. 6 that the performance of the proposed new DNN algorithm scheme is very close to that of the traditional CCCP algorithm.

In order to show the performance of the two algorithm allocation schemes more intuitively, we have drawn a comparison table of the accuracy of the two schemes, as shown in Table 2. From Table 2 we can see that when the target problem is SE maximization, the DNN algorithm can achieve more than 98% of the performance of the traditional CCCP algorithm and can be applied in practical scenarios.

Although the graphics processing unit (GPU) has a speed-up effect when processing neural networks, for the sake of comparative rigor, we use the central processing unit (CPU) to calculate the complexity of the DNN algorithm and the traditional CCCP algorithm separately. Their computation times are shown in Table 3.

**Table 1.** Simulation parameters

| Parameters | Value |
|---|---|
| The cellular radius | 1000 m |
| System bandwidth | 1.8 MHz |
| Noise spectrum density | −174 dBm/Hz |
| The minimum rate requirement of CU | 5 Mbps |
| The minimum rate requirement of D2Ds | 10 Mbps |
| The maximum communication range of D2Ds | 50 m |
| Path loss exponent | 4 |

**Table 2.** The accuracy of DNN

| The maximum transmit power | Accuracy |
|---|---|
| 8 dBm | 99.51% |
| 11 dBm | 98.67% |
| 14 dBm | 98.45% |
| 17 dBm | 98.83% |
| 20 dBm | 98.55% |
| 23 dBm | 98.7% |

**Table 3.** Computational times for the two methods

| The maximum transmit power | Total CPU Time(s) | |
|---|---|---|
| | DNN | CCCP (MATLAB) |
| 8 dBm | 0.000127 | 16.501179 |
| 11 dBm | 0.000100 | 14.093196 |
| 14 dBm | 0.000103 | 13.159920 |
| 17 dBm | 0.000123 | 17.190953 |
| 20 dBm | 0.000125 | 8.748877 |
| 23 dBm | 0.000157 | 8.039851 |

## 5 Conclusion

Our research indicated that it is possible to learn a well-defined algorithm very well by using finite-sized deep neural networks. Our empirical results show that, for maximizing the SE problem of FD-enabled two-way D2D communications, deep neural networks can be trained to well-approximate the behavior of the state-of-the-art algorithm CCCP. However, this paper still exists some limitations. There are several interesting issues to investigate in the future, e.g., we will consider developing an unsupervised learning based power allocation scheme for FD-enabled two-way D2D communications, without using an optimization algorithm to provide training labels.

**Acknowledgment.** This work was supported by the National Natural Science Foundation of China (61701345), and the Natural Science Foundation of Tianjin (18JCZDJC31900).

## References

1. Gupta, A., Jha, R.K.: A survey of 5G network: Architecture and emerging technologies. IEEE Access **3**, 1206–1232 (2015)

2. Ansari, R.I., et al.: 5G D2D networks: techniques, challenges, and future prospects. IEEE Syst. J. **12**(4), 3970–3984 (2018)
3. Wang, L., Tian, F., Svensson, T., Feng, D., Song, M., Li, S.: Exploiting full duplex for device-to-device communications in heterogeneous networks. IEEE Commun. Mag. **53**(5), 146–152 (2015)
4. Li, S., Ni, Q., Sun, Y., Min, G.: Resource allocation for weighted sum rate maximization in multi-user full-duplex device-to-device communications: approaches for perfect and statistical CSIs. IEEE Access **5**, 27229–27241 (2017)
5. Vu, H.V., Tran, N.H., Le-Ngoc, T.: Full-duplex device-to-device cellular networks: power control and performance analysis. IEEE Trans. Veh. Technol. **68**(4), 3952–3966 (2019)
6. Chai, X., Liu, T., Xing, C., Xiao, H., Zhang, Z.: Throughput improvement in cellular networks via full-duplex based device-to-device communications. IEEE Access **4**, 7645–7657 (2016)
7. Liu, F., Hou, X., Liu, Y.: Capacity improvement for full duplex device-to-device communications underlaying cellular networks. IEEE Access **6**, 68373–68383 (2018)
8. Han, L., Zhang, Y., Li, Y., Zhang, X.: Spectrum-efficient transmission mode selection for full-duplex-enabled two-way D2D communications. IEEE Access **8**, 115982–115991 (2020). https://doi.org/10.1109/ACCESS.2020.3004487
9. Sun, H., Chen, X., Shi, Q., Hong, M., Fu, X., Sidiropoulos, N.D.: Learning to optimize: training deep neural networks for interference management. IEEE Trans. Signal Process. **66**(20), 5438–5453 (2018)

# Analysis and Research on Dynamic Measurements Technology

Yingfang Fu[1], Haoxiang Huang[2,3], and Jianbiao Zhang[2,3](✉)

[1] Trusted Native Technology Department of Ant Group Beijing, Beijing, China
[2] Faculty of Information Technology, Beijing University of Technology, Beijing, China
zjb@bjut.edu.cn
[3] Beijing Key Laboratory of Trusted Computing, Beijing, China

**Abstract.** In cloud computing environments characterized by "dynamics", the traditional static measurement to ensure the trust of the platform system is difficult to guarantee the trust of four dimensions in the cloud computing environment, such as the platform system, computation, transmission, and storage. Dynamic measurement, as an important key technology to secure reliable system operation, has received much attention from academia and industry. This paper introduces the principles and characteristics of dynamic measurement and the common attack risks in the running system. Meanwhile, this paper surveys the state of the art of dynamic measurement schemes.

**Keyword:** Cloud computing · trusted computing · dynamic metrics · TEE

## 1 Introduction

With the rapid development of communication technology, the speed of centralization and processing of resources and data in the information industry, especially in the Internet industry, is now accelerating. At present, almost all giant information enterprises are building a complete distributed system platform, that is cloud, to provide services to users. The heterogeneous, dynamic, and distributed have become the basic characteristics of the platform network. In such an environment, security becomes more and more serious, especially malicious code attacks that seriously threaten users' privacy and property security. For example, the cryptographic ransomware virus WannaCry, caused a large number of users' computers to be unusable.

To protect user privacy and data security in the cloud-native environment, the credibility of the operation of the business system with "dynamic" as the main feature is currently the biggest difficulty and a problem that needs to be solved urgently.This paper introduces the principles and characteristics of dynamic measurement and the common attack risks in the running system. Meanwhile, this paper surveys the state of the art of dynamic measurement schemes. Finally, the research challenges and the corresponding solution ideas of dynamic measurement in a cloud computing environment are suggested.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 231–238, 2023.
https://doi.org/10.1007/978-981-99-2362-5_29

## 2 Dynamic Measurement Definition and Key Elements

Trusted computing is a comprehensive information security technology to enhance the credibility of computer systems. Dynamic measurement is the key technology of trusted computing technology. Its goal is to ensure the system is trusted. To make sure the system is trusted by measuring the integrity of system-related components or data, and certifying the system behavior of the running system. The China National Standard for Trusted Software Base [1] defines dynamic measurement as a trusted measurement method, that is, to ensure that the system is trusted by measuring the integrity of the system and evaluating its behavior.

Alibaba defines the terms related to trusted measurement as follows:

a) Trusted measurement is defined as the measurement process of the measured object, which includes four steps: measurement request, data acquisition, hash measurement calculation, and measurement result comparison.
b) Static measurement is defined as the measurement of the device system startup process and application program loading.
c) Dynamic measurement is defined as the measurement of running equipment systems and running applications.
d) Pure soft dynamic measurement is defined as the dynamic measurement process, which is done by the dynamic measurement component.

## 3 Research Status of Dynamic Measurement Technology Solutions

### 3.1 Pure Soft Dynamic Measurement Scheme

#### 3.1.1 Dynamic Measurements Based on Integrity Check

IBM proposed the integrity measurement architecture IMA [2] in 2004, which performs integrity checks on static data to determine whether the system is trusted. However, static measurements cannot accurately reflect the dynamic behavior of the runtime system, and it also lacks flexibility in the triggering method. Some scholars [3, 4] exploit periodic metrics or combined them with information flow to extend access control to provide coarse-grained trusted assurance for the runtime system by setting dynamic trigger measurements. In addition, literature [5] divided the codes in cyber systems into two types, that is, the operating system kernel and various application codes, then the Integrity measurement is triggered by monitoring the process scheduler and system calls at the kernel. This scheme is not only a significant improvement over the previous measurement triggering methods but also controls the frequency of measurements by setting a threshold value, which reduces the performance overhead introduced by measurements. Literature [6] stated that the distribution of measurement points is one of the main factors affecting system performance, and proposes a dynamic integrity measurement scheme that takes the attack time point as the benchmark and dynamically adjusts the correlation coefficient of the measurement interval function according to the security status of the system to achieve the purpose that dynamically modifying the distribution of integrity measurement points, and performs integrity measurement in real-time.

In addition, literature [7] proposed a dynamic trusted measurement based on the page table, this scheme delays the measurement point to the moment when the code page enters the memory. Literature [8] and literature [9] from the perspective of "pagination" are all in-depth studies of measurement trigger conditions. The instruction-level staking technology mentioned in [10] is to insert probes into program instructions and obtain the control flow and data flow information of the program through the execution of the probes. The literature [11, 12] edits the source code of the monitored object based on the instruction-level stubbing technique and jumps to the measurement code when the program execution triggers the probe, i.e., triggers the measurement behavior.

The above schemes adopt some innovation measures in metrics, metric objects, and metric point insertion to improve the trustworthiness of dynamic metrics. However, there also have some obvious disadvantages:

a) From theoretical analysis, it is difficult to balance security and performance overhead. That is, the deployment of measurement points is too dense, which will bring greater system overhead, and the sparse deployment of measurement points will reduce the security of the system.
b) Random time seeds are generated by software, and there is a risk of being intercepted and detected and suffering from TOC-TOU attacks; event triggering can weaken TOC-TOU attacks, but frequent event triggers may bring greater system overhead.

### 3.1.2 Dynamic Measurements Based on Entity Behavior

The literature [13] pointed out that the fundamental problem of dynamic measurements is the accurate, timely, and comprehensive acquisition of measured objects. The runtime system has the characteristics of high dynamics, flexibility, and the uncertainty of dynamic data changes, the trust of static data cannot effectively characterize the dynamic trust of the runtime system, and it is also not in line with the "behavior as expected". Integrity checks cannot prevent code-reuse attacks based on control flow hijacking [14, 15]. Literature [16] pointed out that many attacks are implemented through system calls, and thus the trusted expected behavior at runtime can be portrayed by modeling that system calls during program operation.

**System Call Short Sequence Model.** The N-gram detection method based on system call sequences proposed in the reference [17, 18] has the advantage of fast training by presetting a constant N and using enumeration of all adjacent and unique sequences of length N in the training dataset to obtain the normal behavior of the application and then using a sliding window of length N to sequentially match the detected system call sequences. However, the above scheme can only identify the sequence of system calls less than or equal to N encountered during training, which has the disadvantage of an inefficient detection rate.

**Static Analysis.** The static analysis mainly exploits the executable binary file format to extract instruction sequences utilizing decompiling, or extracting system calls and function calls to construct control flow analysis graphs, and use them as features to build malicious code, and detection models. Literature [19] proposed a method to obtain a

behavior model through static analysis of the source code, which is an early attempt to establish a normal behavior model for system calls through static analysis.

These schemes have the following disadvantages: a) Lack of systematic methods to verify the results of static analysis on a large scale, and the accuracy and completeness of the measurement results can not be guaranteed; b) The method of static analysis is efficient, but it cannot truly simulate the dynamic running process of the program and there is a problem of a high false alarm rate; c) It is susceptible to the continuous improvement of obfuscation techniques by attackers, and malicious behavior is difficult to identify.

**Dynamic Analysis.** The dynamic analysis method is an automatic model construction technology that constructs a normal behavior model from the observed software execution. Reviewing the research on dynamic analysis, this research started from Stow-away, a dynamic analysis solution proposed in the reference [20]. Literature [21] proposed and designed a dynamic analysis framework for Android applications. Reference [22] proposed a dynamic-length chain trust transfer model in the context of industrial control systems, which can adapt to different real-time requirements and integrity verification based on meeting the real-time requirements of industrial control systems. The technology of dynamic behavior analysis makes up for the disadvantages of static analysis in terms of detection effectiveness, but the models proposed by the above schemes underestimate the expected behavior and the models consider only the behavior observed during training, and not all valid program executions have been observed.

## 3.2 Dynamic Measurement Based on the Combination of Software and Hardware

### 3.2.1 SGX-Based Dynamic Measurements

SGX [23] introduced by Intel in 2013 is a hardware mechanism that is based on memory encryption and attestation, which provides a trusted execution environment in user space for applications and a new solution to the problem of runtime dynamic protection.

Literature [24] proposed a Spons & Shields framework that separates applications in a trusted execution environment through response and shielding. The reference [25] presented the Ryoan prototype, which provides a distributed sandbox [26] that uses SGX to protect sandbox instances from potentially malicious computing platforms. The reference [27] pointed out that measurement components located in TEE [28] share the context of trusted OS with other target applications (TA). Any TA attack may affect the security of measurement components. Therefore, a solution for migrating the measurement program to an environment similar to SGX called Scanclave is proposed. TEE allows applications in user space to create a protected area to prevent software from running at a higher level of authority.

Compared with purely soft dynamic measurement schemes, SGX has some technical advantages as follows: a) Using memory encryption technology to protect the security of the program's running state makes it more difficult to obtain critical information through memory leak attacks. b) By reducing the trusted computing base of the system to the CPU, more system attacks can be avoided compared with the entire operating system or privileged software (such as a hypervisor, etc.) as the trusted computing base in the past.

### 3.2.2 Multicore Architecture-Based Dynamic Measurements

With the advancement and development of processor-related hardware technologies, multi-core CPU computing platforms are being applied in practice. From the perspective of CPU multi-core characteristics, the multi-core collaborative architecture with computational and security cores in parallel becomes a new idea to solve the dynamic measurement performance overhead problem.

The reference [29] proposed a dynamic measurement scheme for split-core asynchronous systems based on the ARM platform. The scheme utilizes a special computing core to actively monitor and actively measure the system kernel resources in real-time, which can effectively mitigate the impact of the large computational overhead on the host system due to dynamic measurements and monitoring. The reference [30] proposed an improved trusted architecture for immune monitoring of aircraft hardware attacks. The feature of this scheme is that the trusted monitoring system hardware is isolated from the main processor system, and the measurement program running on the telecommunication management system is used to measure and call the lightweight measurement agent running on the mobile management system. This scheme guarantees the security of the trusted monitoring system.

Compared with the existing pure soft dynamic measurement solution, the advantage of dynamic measurement based on the multi-core architecture is that the security of the dynamic measurement component itself is improved, and the system performance overhead caused by the dynamic measurement is reduced. Its disadvantage is that it needs to change the traditional CPU architecture for trusted adaptation, and consumes computing resources on the user side.

## 4 Technology Challenges and Solutions for Dynamic Measurements

In the cloud computing environment characterized by “dynamic”, traditional measurement methods that mainly rely on static measurement methods to ensure the credibility of the platform system are difficult to ensure the credibility of the platform system, computing, transmission, and storage in the cloud computing environment. Compared with the pure soft dynamic measurement schemes, the combined software and hardware dynamic measurement schemes improve the security of the dynamic measurement component itself and reduce the system performance overhead caused by the dynamic measurement. However, from the perspective that the dynamic measurement would be applied in practice, its security and performance need to be improved. Based on the descriptions in the previous sections, the future challenges of dynamic measurement are as follows: a) How to secure the dynamic measurement component itself; b) How to ensure that the sensitive data in the dynamic area is not leaked and the integrity is not tampered with; c) How to solve the problem of performance overhead caused by dynamic measurement; d) How to improve the accuracy of dynamic measurement in identifying malicious attacks.

Based on the analysis of the difficult problems that need to be solved in dynamic measurement, the author believes that the following corresponding measures can be taken to solve them.

a) Ensure the security of dynamic measurement components.
   Independent hardware security service components are integrated with TEE technology to ensure the security of the dynamic measurement component, and to provide dynamic measurement-related secure computing services for business computing nodes to achieve the trade-off between security and performance.
b) Sensitive data located in the dynamic area of security protection.

   i. Measured objects have a transformation from statics areas (e.g. static code segments, configuration files, linked library files, etc.) to dynamic areas.
   ii. The subject's malicious behavior identification technology relies on the artificial intelligence technology and monitoring technology based on the system call behavior and control behavior to prevent the leakage of sensitive data in the dynamic area, and the integrity of the dynamic measurement component and the sensitive data in the dynamic area are not tampered with.
   iii. Measured object changes from platform system startup integrity detection to runtime multi-level, multi-dimensional integrity dynamic measurements and behavioral measurements, which improves the accuracy of the system runtime dynamic measurement to identify malicious attacks.

c) The trade-off between performance and security of dynamic measurement.

   i. Fixed interval, random interval, and event-triggered measurements to ensure dynamic measurements in terms of security and performance overhead trade-offs.
   ii. The integrated form with a multi-core CPU supports dynamic measurement schemes.

## 5 Conclusion

This paper summarizes and analyzes the mainstream risk attacks faced by the system operation, the technical principles of dynamic measurements, the key factors affecting the dynamic measurement evaluation results, and the technical status of dynamic measurement schemes. On this basis, it points out the advantages and disadvantages of the existing dynamic measurement technology solutions, and further introduces the challenges and corresponding measures that dynamic measurement technology faces in the field of cloud computing.

## References

1. GB/T 37935-2019, Information security technology—Trusted computing specification—Trusted software base (2019)
2. Sailer, R., Zhang, X., Jaeger, T., et al.: Design and implementation of a TCG-based integrity measurement architecture. In: Proceedings of the 13th USENIX Security Symposium, 9–13 August 2004, San Diego, CA, USA (2004)

3. Petroni, N.L., Hicks, M.: Automated detection of persistent kernel control-flow attacks. In: Proceedings of the 14th ACM conference on Computer and communications security (CCS 2007), pp. 103–115. Association for Computing Machinery, New York, NY, USA (2007)
4. Jaeger, T., Sailer, R., Shankar, U.: Prima: policy-reduced integrity measurement architecture (2007)
5. Liu, C., Fan, M., Feng, Y., et al.: Dynamic integrity measurement model based on trusted computing. In: International Conference on Computational Intelligence and Security, CIS 2008, vol. 1, pp. 281–284 (2008)
6. Hu, M., Yang, Y., Lv, G., Liu, X.: Research on dynamic integrity measurement strategy based on measurement point distribution. In: 2018 Chinese Control and Decision Conference (CCDC), Shenyang (2018)
7. Wu, Y.: Research and Implementation of Trustworthiness Measurement Technology for Windows Applications. Nanjing University of Science and Technology (2012)
8. Wu, T., Yang, Q.S., He, Y.P.: Method of dynamic integrity measurement for VMM based on adjacency data. J. Commun. (2015)
9. Cai, M.J., Chen, X.S., Jin, X., Zhao, C., Yin, M.Y.: Paging-measurement method for virtual machine process code based on hardware virtualization. J. Comput. Appl. (2018)
10. D'Elia, D.C., Coppa, E., Nicchi, S., et al.: SoK: using dynamic binary instrumentation for security (and how you may get caught red handed). In: Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, pp. 15–27 (2019)
11. Wang, Z.W., Zhuang, Y., Yan, Z.J.: Probe-based dynamic integrity measurement scheme for mobile devices using ARM TrustZone. J. Chin. Comput. Syst. (2021)
12. Vaduva, J.A., Dascalu, S., Florea, I.M., et al.: Observations over SPROBES mechanism on the TrustZone architecture. In: 2019 22nd International Conference on Control Systems and Computer Science (CSCS), pp. 317–322. IEEE (2019)
13. Chen, Z.F., Li, Q.B., Zhang, P., Wang, W.: Kernel integrity measurement method based on memory forensic. J. Softw. **27**(9), 2443–2458 (2016). (in Chinese). http://www.jos.org.cn/1000-9825/4875.htm
14. Peng, G.J., Liang, Y., Zhang, H.G., Fu, J.M.: Survey on software binary code reuse technologies. J. Softw. **28**(8), 2026−2045 (2017). (in Chinese). http://www.jos.org.cn/1000-9825/5270.htm
15. Dileesh, E.D., Shanthi, A.P.: An application specific dynamic behaviour model using function-call sequence and memory access-graph for execution integrity verification. Comput. Secur. **107**, 102299 (2021)
16. Maggi, F., Matteucci, M., Zanero, S.: Detecting intrusions through system call sequence and argument analysis. IEEE Trans. Dependable Secure Comput. **7**(4), 381–395 (2010)
17. Forrest, S., Hofmeyr, S.A., Somayaji, A., et al.: A sense of self for Unix processes. In: Proceedings of the IEEE Symp. on Security and Privacy, pp. 120–128 (1996)
18. Hofmeyr, S., Forrest, S., Somayaji, A.: Intrusion detection using sequences of system calls. J. Comput. Secur. **6**, 151–180 (1998)
19. Wagner, D., Dean, D.: Intrusion detection via static analysis. In: Proceedings of the IEEE Symposium on Security and Privacy, pp. 156–169 (2001)
20. Felt, A.P., Chin, E., Hanna, S., Song, D., Wagner, D.A.: Android permissions demystified. In: 18th ACM Conference on Computer and Communication Security (CCS 2011). ACM (2011)
21. Dawoud, A., Bugiel, S.: Bringing balance to the force: dynamic analysis of the android application framework. In: Network and Distributed System Security Symposium (2021)
22. Shang, W., Xing, X.: ICS software trust measurement method based on dynamic length trust chain. Sci. Program. **2021**(5), 1–11 (2021)
23. Wang, J., et al.: Analysis and research on SGX technology. Ruan Jian Xue Bao/J. Softw. **29**(9), 2778−2798 (2018). (in Chinese). http://www.jos.org.cn/1000-9825/5594.htm

24. Sartakov, V.A., O'Keeffe, D., Eyers, D., et al.: Spons & Shields: practical isolation for trusted execution. In: VEE 2021: 17th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments. ACM (2021)
25. Hunt, T., Zhu, Z., Xu, Y., et al.: Ryoan: a distributed sandbox for untrusted computation on secret data. ACM Trans. Comput. Syst. **35**(4), 1–32 (2018)
26. Indradevi, K.A.R., Sukarno, P., Jadied, E.M.: Analisis Performansi Aplikasi Sandbox pada Sistem Operasi Windows. eProceedings Eng. **5**(3) (2018)
27. Morbitzer, M.: Scanclave: verifying application runtime integrity in untrusted environments. In: 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 198–203. IEEE (2019)
28. Feng, D.G., Liu, J.B., Qin, Y., et al.: Trusted computing theory and technology in innovation-driven development (in Chinese). Sci. Sin. Inform. **50**(8), 1127–1147 (2020). https://doi.org/10.1360/SSI-2020-0096
29. Wang, W.H., Chen, G.X., Pan, X.R., et al.: Leaky cauldron on the dark land: understanding memory side-channel hazards in SGX. In: Proceedings of the 2017 ACM SIGSAC Conf. on Computer and Communications Security (CCS 2017), pp. 2421–2434 (2017)
30. Cheng, D., Zhang, C., Liu, J., et al.: An attack-immune trusted architecture for supervisory aircraft hardware. Chinese J. Aeronaut. (2021)

# Lightning Protection Design for 20 kV Insulated Crossarm of Different Altitudes

Jin Li[1(✉)], Li Wang[1], and Shuang Zhang[2]

[1] College of Engineering and Technology, Tianjin Agricultural University, Tianjin 300384, China
Davidjin2013@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** At present, there are few researches on lightning protection design of 20 kV line in China, and the insulation configuration level of 20 kV line is usually low. Once the line is caused by lightning failure, it will cause great losses to the society, so the research on 20 kV line is very necessary. Combined with the actual situation, this paper uses the insulation cross arm for lightning protection. By introducing the dry arc distance correction coefficient and the ground lightning density correction parameters, the lightning protection model of 20 kV lines with different elevations is established. And the calculated data is processed and analyzed by MATLAB, and the lightning protection design of 20 kV insulation cross arm with different elevations is finally completed. In order to avoid the line suffered from lightning, lightning protection design should be considered from the following three aspects; First of all, the insulation level of the line and cable should be improved, and the line should be inspected regularly. Secondly, according to the altitude of the line, a reasonable distance between the dry arc of the insulation cross arm is set. Finally, the lightning arrester is installed on the tension rod. According to the lightning disaster grade corresponding to the line, the optimal range of the number of lightning arrester installed on the pole tower of 20 kV line is determined.

**Keyword:** 20 kV · Insulated crossarm · Ground flash density · Lightning protection design · Altitude

## 1 The Introduction

With the rapid development of China's economy, the scale of power system in the constantly expanding, 20 kV voltage class in strengthen the power grid transmission capacity, meet user dense area of the power supply, power supply range and so on compared with 10 kV voltage level has obvious advantage.

The insulation configuration level of 20 kV voltage grade overhead lines is usually low, and lightning protection design becomes one of the most important problems [2]. Because the lightning strike is affected by the altitude factor, it is necessary to consider the altitude factor to design the lightning protection of 20 kV insulation cross arm.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 239–249, 2023.
https://doi.org/10.1007/978-981-99-2362-5_30

Various accidents caused by insulation breakdown caused by lightning strike on overhead lines are the focus of lightning protection on lines [5]. As an effective lightning protection method, the insulation cross arm can reduce the number of lightning flashover of the line and increase the reliability and continuity of the line operation [6]. In this paper, the lightning protection model of 20 kV insulation cross arm at different altitudes was established by referencing the correction coefficient of ground lightning density affected by altitude factors and installing zinc oxide lightning arrester on the tensile rod. The calculated data were processed and analyzed by using Matlab, and the lightning protection design of 20 kV insulation cross arm at different altitudes was finally completed.

## 2 Expressions of Lightning and Lightning Protection Means

The forms of lightning are direct lightning and inductive lightning. According to the observation data, the proportion of inductive lightning and direct lightning on 20 kV overhead lines is about 9:1 [4]. Therefore, in order to design the economy of lightning protection, the lightning protection design of 20 kV insulation cross arm can only consider the protection of inductive lightning. For the economic and applicability of installation, 20 kV overhead lines mainly adopt the installation of insulation cross arm, lightning arrester and other devices for lightning protection.

The lightning discharge process is in the pilot stage, and the wire is tightly surrounded by the electric field formed by the lightning cloud, the pilot channel and the ground, as shown in Fig. 1 below.

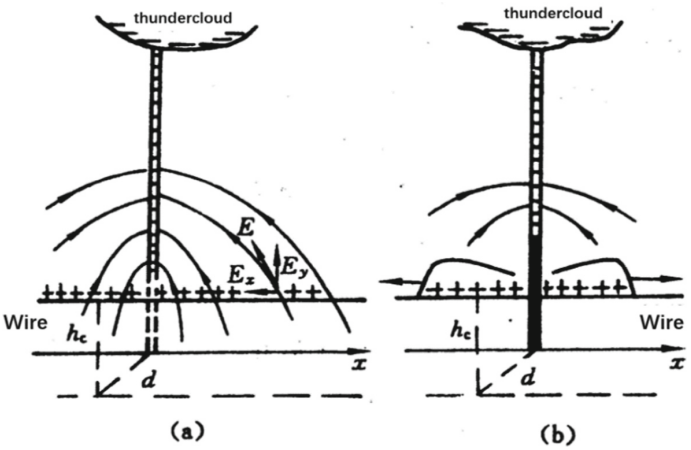



**Fig. 1.** Schematic diagram of induced lightning overvoltage mechanism

When the distance between the lightning strike location and the cable is greater than 65 m, the induced overvoltage caused by lightning on the cable is [1]:

$$U_i = 25 \times \frac{I \cdot h_c}{S} \tag{1}$$

In Eq. (1), the distance between the lightning strike location and the line is represented by S, the lightning current amplitude is usually represented by I, and the height of the wire from the ground can be represented by $h_c$.

The induced overvoltage on the 20 kV voltage class line is shown in Table 1.

**Table 1.** Induced overvoltage at different distances of 20 kV overhead lines

| distance/m | I/kA | | | | |
|---|---|---|---|---|---|
| | 20 | 40 | 60 | 80 | 100 |
| 65 | 96.2 | 192.3 | 288.5 | 384.6 | 480.8 |
| 70 | 89.3 | 178.6 | 267.9 | 357.1 | 446.4 |
| 80 | 78.1 | 156.3 | 234.4 | 312.5 | 390.6 |
| $\rho$% | 59.3 | 35.1 | 20.8 | 12.3 | 7.30 |

When overhead lines are under overvoltage, flashover often occurs, and the arc generated may burn the wires.

## 3 Design of Dry Arc Distance of Insulation Cross Arm

Cross-arm is an important part of transmission line under external load, so insulation performance is related to the safety and reliability of transmission line [10]. The composite insulation cross arm can reduce the lightning strike of the insulation cross arm.

Dry arc distance is also called arc distance. The shortest external distance between two metal parts of the insulation cross arm is h [7] under normal working voltage. This is shown in Fig. 2.

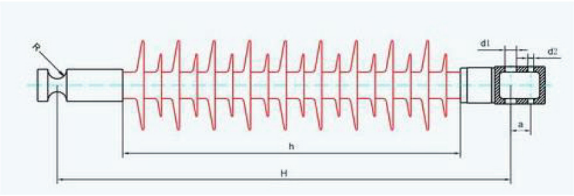



**Fig. 2.** Diagram of dry arc distance

According to the initial value, the induction lightning overvoltage of 20 kV overhead line is 480.8 kV. The 20kV lightning voltage amplitude is generally less than 500 kV. Referring to relevant literature, it can be seen that the dry arc distance of the insulation cross arm is set to 750 mm at an altitude of 1000 m and below.

The calculation formula of voltage incident wave is as follows:

$$U_0 = I_0 \cdot Z_0 \tag{2}$$

The voltage incident wave can be approximated as an induced overvoltage. According to the provisions of the relevant manual in China, the impedance of the lightning wave $Z_0$ is usually 300. The intensity of lightning in a certain area can be measured by the amplitude of lightning current. $I \approx 2I_0$. The lightning current amplitude can be calculated according to the following formula:

$$\mathrm{Ig}P = -\frac{I}{88} \tag{3}$$

In Eq. (3), P is the probability that the amplitude is not less than I. It can be obtained that the 20 kV insulation cross arm can meet the requirements of lightning protection design when the altitude is below 1000 m and the dry arc distance is 750 mm. At higher altitude, the insulation performance of the 20 kV line will decrease. According to the relevant provisions of Technical Specifications for High Altitude External Insulation Configuration, the calculation formula for altitude correction is as follows:

$$k = e^{m\frac{H_2-H_1}{8150}} \tag{4}$$

wherein, $H_2$ is the altitude of the position of the insulation cross arm, $H_1$ is the altitude of the experiment of the insulation cross arm, and m is the elevation correction factor. In the case of lightning overvoltage, 1 is usually taken. After calculation, the dry arc distance of 20 kV insulation cross arm at various altitudes can be obtained, as shown in Table 2.

**Table 2.** Dry arc distance values of 20 kV insulation cross arms at different altitudes

| Altitude H/m | Elevation correction factor k | Dry arc distance/mm |
|---|---|---|
| $H \leq 1000$ | 1 | ≥750 |
| $1000 < H \leq 2000$ | 1.131 | ≥849 |
| $2000 < H \leq 3000$ | 1.278 | ≥959 |
| $3000 < H \leq 4000$ | 1.445 | ≥1084 |
| $4000 < H \leq 5000$ | 1.634 | ≥1226 |

## 4 Arrester Design

The arrester is a kind of electrical equipment which can reduce overvoltage effectively. It is essentially a discharger [3]. The tensioning rod of 20 kV overhead line mainly uses zinc oxide lightning arrester for lightning protection [12]. It has the advantages of rapid response time, flat volt-ampere curve, relatively stable performance and simple structure.

The rated voltage Ur is 25 kV. The maximum continuous operating voltage is Uc, and the maximum continuous operating voltage is generally 0.8 times of the rated voltage, so the maximum operating voltage is generally 20 kV [9]. The main parameters of 20 kV zinc oxide arrester are shown in Table 3.

**Table 3.** Parameters of 20 kV zinc oxide arrester

| Technical parameters | Recommended values | 3EG5 | MWD22/WMK22 |
|---|---|---|---|
| The rated voltage Ur/kV | 25 | 24 | 27.5 |
| Continuous running voltage Uc/kV | 20 | 19.5 | 22 |
| DC 1 mA reference voltage/kV | 35.4 | | 36.7 |
| 30/60 $\mu s$ 500 A Operating impact residual pressure/kV | 55 | | 54.1 |
| 8/20 $\mu s$ 5 kA Lightning shock full wave residual pressure/kV | 66 | 66 | 63.8 |
| 2 ms Square wave flow capacity/A | 500 | 200 | 550 |

## 5 Lightning Protection Design Considering Altitude Factor

### 5.1 Analysis of Lightning Failure Rate of 20 kV Line

Elevation factor has become a non-negligible factor for failure of 20 kV voltage grade lines caused by lightning strike [8]. Lightning tripping rate can reflect the lightning protection level of 20 kV overhead line. It can be defined as the number of trips caused by lightning strikes per 100 km line in 40 lightning days, which can be expressed by the following formula:

$$n = N \times P_1 \times \eta \tag{5}$$

In Eq. (5): N is the total number of lightning strikes suffered by 20 kV line; $P_1$ is the probability that the lightning current is not less than the lightning resistance degree; $\eta$ is the probability that the flashover of the 20 kV line becomes a stable power frequency arc when it is subjected to the shock caused by inductive lightning. Lightning density can reflect the frequency of lightning strike in a certain local line. Its calculation formula is as follows:

$$N_g = 0.024 N_d^{1.3} \tag{6}$$

In Eq. (6), $N_\mathrm{d}$ is the annual average number of thunderstorm days. This paper uses the lightning density coefficient to represent the lightning frequency [11].

By referring to relevant literature, the lightning density, average altitude and lightning tripping rate of each county in a city from 2009 to 2011 were collected and analyzed, and the influencing factors of lightning tripping of 20 kV voltage grade lines were studied.

Table 4 shows the lightning density data of each county from 2009 to 2011. Figure 3 shows the lightning density curve of each county from 2009 to 2011 drawn by Matlab (Tables 5 and 6).

It can be concluded from the above analysis that the higher the altitude, the more likely the 20 kV line will suffer from lightning failure (Fig. 4).

**Table 4.** Ground lightning density in each county from 2009 to 2011 $n \cdot \left(km^2 \cdot a\right)^{-1}$.

| year | A county | B county | C county | D county | E county | F county |
|---|---|---|---|---|---|---|
| 2009 | 5.1 | 7.1 | 12.5 | 10.9 | 5.6 | 5.3 |
| 2010 | 6.9 | 7.7 | 9.2 | 8.3 | 6.5 | 7.7 |
| 2011 | 5.9 | 5.5 | 8.4 | 7.9 | 5 | 4.7 |

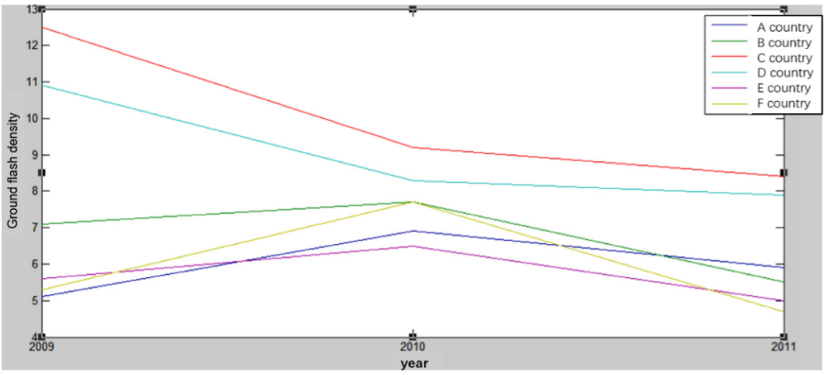



**Fig. 3.** Curse of lightning density in each county

**Table 5.** Trip rate of 20 kV line in each county from 2009 to 2011 (%)

| year | A county | B county | C county | D county | E county | F county |
|---|---|---|---|---|---|---|
| 2009 | 0.24 | 0.63 | 0.54 | 0.47 | 1.46 | 3.27 |
| 2010 | 0.53 | 1.03 | 0.88 | 0.80 | 1.79 | 2.99 |
| 2011 | 0.57 | 0.99 | 1.02 | 1.00 | 1.75 | 4.25 |

**Table 6.** Average altitudes by county

| County name | A county | B county | C county | D county | E county | F county |
|---|---|---|---|---|---|---|
| Average altitudes/m | 100 | 200 | 250 | 300 | 400 | 500 |

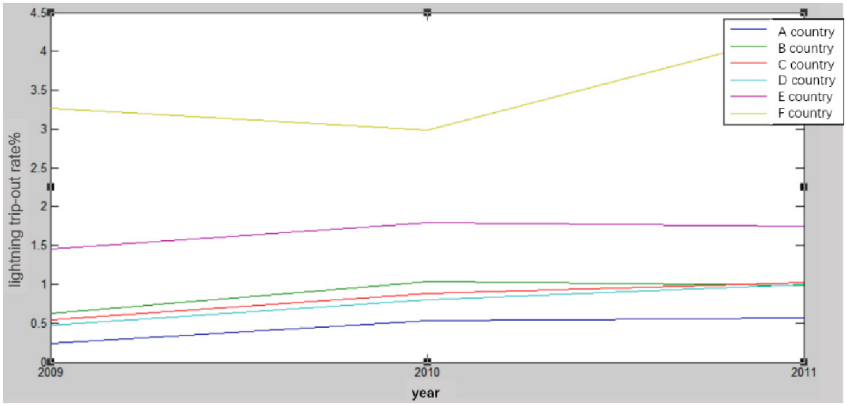



**Fig. 4.** Curve of lightning trip rate in each county

### 5.2 Design of Correction Coefficient of Ground Flash Density

In order to find out whether the elevation factor has an impact on the lightning failure rate of the 20 kV voltage grade line, the relevant data of 2011 from various regions are now collected as shown in Table 7 below.

**Table 7.** Relevant data by county in 2011

| Place | ground flash density/$n \cdot \left(km^2 \cdot a\right)^{-1}$ | average elevation/m | Lightning tripping rate (%) |
|---|---|---|---|
| A county | 5.9 | 100 | 0.57 |
| B county | 5.5 | 200 | 0.99 |
| C county | 8.4 | 250 | 1.02 |
| D county | 7.9 | 300 | 1.00 |
| E county | 5.0 | 400 | 1.75 |
| F county | 4.7 | 500 | 4.25 |

Define elevation unit values $\mu$, Its expression is as follows:

$$\mu = \frac{h_x}{h_{baisc}} \tag{7}$$

In Eq. (7), $h_x$ is the altitude of the location of the line, and $h_{basic}$ is the base altitude of 100 m. It is assumed that the failure probability of 20 kV voltage grade line caused by lightning strikes is linearly correlated with the lightning density, that is:

$$\frac{\lambda_x}{\lambda_{basic}} = \frac{\gamma_x}{\gamma_{basic}} \tag{8}$$

In Eq. (8), $\lambda_x$, $\gamma_x$ Are lightning tripping rate and ground lightning density in a certain area respectively; $\lambda_{basic}$, $\gamma_{basic}$ is the lightning tripping rate and ground lightning density in the reference area. It can be seen that the lightning tripping rate of the 20 kV voltage grade line is not positively correlated with the ground lightning density. Now, the correction coefficient $\alpha$ of ground lightning density is introduced, and its specific relationship is as follows:

$$\alpha = \frac{\lambda_x \times \gamma_{basic}}{\lambda_{basic} \times \gamma_x} \tag{9}$$

According to the data in the table and Eqs. 7 and 9, the relationship between the correction coefficient of ground lightning density and the altitude unit value can be obtained as shown in Table 8 below.

**Table 8.** Ground lightning density correction coefficient and elevation unit value data

| parameter | A county | B county | C county | D county | E county | F county |
|---|---|---|---|---|---|---|
| $\alpha$ | 0.797 | 1.484 | 1 | 1.04 | 2.86 | 7.46 |
| $\mu$ | 1 | 2 | 2.5 | 3 | 4 | 5 |

In order to study the relationship between $\alpha$ and $\mu$, MATLAB was used to carry out polynomial fitting processing on the experimental data calculated and collected above, and the graph as shown in Fig. 5 below was obtained.

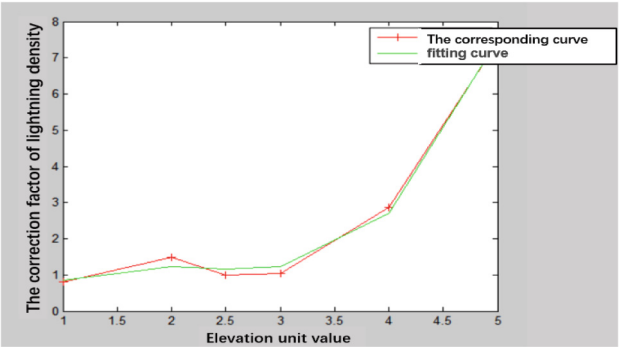



**Fig. 5.** Correction coefficient of ground lightning density and elevation curve

Using polynomial fitting method, the relationship between the correction coefficient of ground lightning density and altitude is obtained as follows:

$$\alpha = 0.3093\mu^3 - 2.0442\mu^2 + 4.3397\mu - 1.7544 \tag{10}$$

### 5.3 Lightning Protection Scheme Design

By multiplying the calculated correction coefficient of ground lightning density $\alpha$ by the value of ground lightning density $\gamma$, the modified value of ground lightning density after the influence of altitude factors can be obtained as follows:

$$\gamma_{revise} = \alpha \times \gamma \tag{11}$$

In Eq. (11), $\gamma_{revise}$ is the modified value of the ground-flash density, and $\gamma$ is the actual measured value of the ground-flash density. According to the modified value of the calculated lightning density, it can be divided into three levels of lightning disaster. The modified value of lightning density in each county is shown in Table 9 below.

**Table 9.** Modified values of lightning density in each county

| Place | A county | B county | C county | D county | E county | F county |
|---|---|---|---|---|---|---|
| $\gamma_{revise}$ | 4.7 | 8.2 | 8.4 | 8.2 | 14.3 | 35 |

In a certain area, the probability $\sigma$ of lightning strike per kilometer at general altitude is calculated as follows:

$$\sigma = \gamma_{revise} \frac{w + 4h}{10^4} \tag{12}$$

In Eq. (12), $\gamma_{revise}$ is the modified value of ground lightning density, and h is the vertical suspended height of 20 kV overhead line. W is the distance between two lightning arrester lines. Since lightning arrester lines are not installed on 20 kV overhead lines for lightning protection, the value of W is 0. The classification of lightning hazard and the probability of lightning strike per kilometer are shown in Table 10 below.

**Table 10.** Classification of lightning disaster and the probability of lightning strike

| Class of lightning disaster | Modification of ground flash density | The probability of being struck by lightning per kilometer |
|---|---|---|
| Level 1 | $\gamma_{revise} > 25$ | $\sigma > 0.12$ |
| Level 2 | $10 < \gamma_{revise} \leq 25$ | $0.6 < \sigma \leq 0.12$ |
| Level 3 | $\gamma_{revise} \leq 10$ | $\sigma \leq 0.6$ |

It can be concluded that the higher the altitude, the greater the chance of being struck by lightning. Referring to relevant literature, it can be found that the base N of the poles and towers of 20 kV overhead lines is between 13–20 per kilometer. According to the probability of being struck by lightning per kilometer, the number of poles and towers separated by two adjacent arrester poles and towers N can be calculated, that is:

$$n = \frac{N}{N \times \sigma} = \frac{1}{\sigma} \tag{13}$$

According to the calculated lightning strike probability per kilometer and Eq. (13), the best value of installing lightning arrester on 20 kV overhead line for lightning protection can be calculated. This is shown in Table 11 below.

**Table 11.** Lightning disaster grade and lightning arrester installation range

| Class of lightning disaster | Optimal range for installing arrestors |
|---|---|
| Level 1 ($\gamma_{revise} > 25$) | A group of arrestors shall be installed at an interval of 8–9 poles and towers |
| Level 2 ($10 < \gamma_{revise} \leq 25$) | A group of lightning arresters shall be installed at intervals of 10–16 poles and towers |
| Level 3 ($\gamma_{revise} \leq 10$) | A group of arrestors shall be installed on the poles and towers at intervals of 17–20 |

## 6 Summary and Prospect

In this paper, the lightning protection scheme of 20 kV voltage grade line is designed by using the insulation cross arm. This paper establishes lightning protection models of 20 kV lines at different altitudes by designing the ground lightning density correction parameters, and processes and analyzes the calculated data with MATLAB, finally completes the lightning protection design of 20 kV insulation cross supports at different altitudes. In order to avoid making the line suffer lightning strikes, should consider from the following three aspects.

(1) first of all, the insulation level of the line and cable should be improved, and the line should be inspected regularly.
(2) Secondly, according to the altitude where the line is, a reasonable value of the distance between the dry arc of the insulation cross arm is set.
(3) Finally, the lightning arrester is installed on the tension rod. According to the lightning disaster grade corresponding to the line, the optimal range of the number of lightning arrester installed on the pole tower of 20 kV line is determined.

## References

1. Chen, Y.: Study on lightning protection of distribution network voltage level from 10 kV to 20 kV. Sci. Technol. Commun. **6**(22), 137+121 (2014)
2. Ding, Y., Wang, D., Ke, G.: Application research of 20 kV medium voltage distribution voltage level. Mech. Electr. Inf. **2015**(27), 5–7 (2015)
3. Feng, X.: Building integrated lightning protection technology and system design. In: Proceedings of 2016 4th International Conference on Machinery, Materials and Information Technology Applications (ICMMITA 2016) (2016)

4. Ge, J.: Application of comprehensive lightning protection measures for overhead insulated distribution lines. East China Electric Power **12**, 79–81 (2005)
5. Han, J., et al.: Study on lightning overvoltage and comprehensive measures of lightning protection of 10 kV overhead insulated conductor. High Voltage Technol. **2008**(11), 2395–2399 (2008)
6. Hu, Y., Wang, L., Zheng, C., Shao, F., Liu, K.: Insulation tower head and crossarm of 10 kV line against lightning and pollution flash. High Voltage Technol. **2007**(12), 108–110+161 (2007)
7. Jiang, A., Nie, P., Zhu, J., Dai, H., Wan, J.: Influence of insulation cross support measures on lightning intrusion wave in 10 kV substation. Electric Porcelane Lightning Arrester **2017**(05), 165–170 (2017)
8. Jiang, X.: Research on lightning protection measures of 10 kV overhead insulated conductors. Mech. Electr. Inf. **09**, 20–21 (2014)
9. Pang, M.: Lightning protection design of 10 kV insulation crossarm. In: China Electric Power Research Institute. Proceedings of 2018 Smart Grid Informatization Construction Seminar. China Electric Power Research Institute: The Sixth Research Institute of China Electronics Information Industry Group Co., Ltd., vol. 2 (2018)
10. Wen, Z., Wang, J., Su, H.: Design of a porcelain crossarm type series clearance arrester. Mech. Electr. Inf. **2018**(30), 118–119 (2018)
11. Li, X., Li, H., Liu, Y., et al.: Insulation coordination of arcing horns on HVDC electrode lines: protection performance evaluation, influence factors and improvement method. Energies **11**(2) (2018)
12. Zhang, C.: Selection of scheme for preventing lightning breakage of overhead insulated lines. Hubei Electric Power **01**, 57–59 (2007)

# Modeling and Inspection Design of High Voltage Line Deicing Device

Jin Li[1(✉)], Li Wang[1], and Yuming Du[2]

[1] College of Engineering and Technology, Tianjin Agricultural University, Tianjin 300384, China
Davidjin2013@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** In order to ensure the quality and stability of power and remove the ice on the high voltage line, this study designed a patrol check and de-icing device that walks on the high voltage line. The device is a multi-functional device that integrates patrol check, de-icing, weather information collection, early warning and feedback. With STM32 microcontroller as the control core, the device can detect the data of some weather air near the high voltage line through a variety of different sensors, and can also distinguish some early trouble hidden dangers (such as broken stock, rust, deformation, etc.) of the high voltage line. This device adopts physical deicing method, deicing knife. This research adopts 3D printing method to make, through the ice cutting separation, to achieve the purpose of deicing. This device provides reference for the intellectualization of power system.

**Keyword:** high voltage line · Deicing device · 3D modeling · Inspection · STM32 · microcontroller

## 1 Introduction

In recent years, with the growing impact of global warming, the number of ice disasters on the power grid has increased. Icing disaster of high voltage transmission lines has become one of the problems troubling China's electric power. The ice and snow on the high voltage line will make its electrical and mechanical properties decline sharply, and a variety of serious accidents will occur, resulting in inestimable losses, and even casualties [5], so the inspection is essential. However, the high-voltage tower is not limited to the edge of the city, and most of them are located in the remote mountains and forests where people are extremely rare, so it is very difficult to patrol. Because the workload of inspection is large, the span between poles and towers is large, the time required is long, the work is very cumbersome and the tools carried are many, at present, the operation is basically by manpower, and the difficulty of inspection of poles and towers at different locations will also be different. This kind of inspection is not only cumbersome but also inefficient. Therefore, how to clear the ice of high voltage power lines is more important [6] -- it is very necessary to develop a patrol deicing


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 250–258, 2023.
https://doi.org/10.1007/978-981-99-2362-5_31

device. Machines can not only replace people to complete some high-risk work, but also it can have a variety of functions, uninterrupted work. Therefore, based on the above requirements, the content of this study includes: The first part is to determine the overall framework of the equipment, including the design of the structure and function of the device, what method to deice, etc.; The second step is to complete the design of the de-icing part and the modeling of the de-icing knife. In this study, 3Done software was used for modeling. After printing, the de-icing knife should be detected and adjusted. The third step is to select the components and sensors to be used according to their functions, connect them to the SCM, write programs, and debug the sensors to ensure their working efficiency and accuracy. The final step is to assemble the device from components and sensors, and to debug and modify its overall structure and function.

## 2 Inspection of Deicing Device

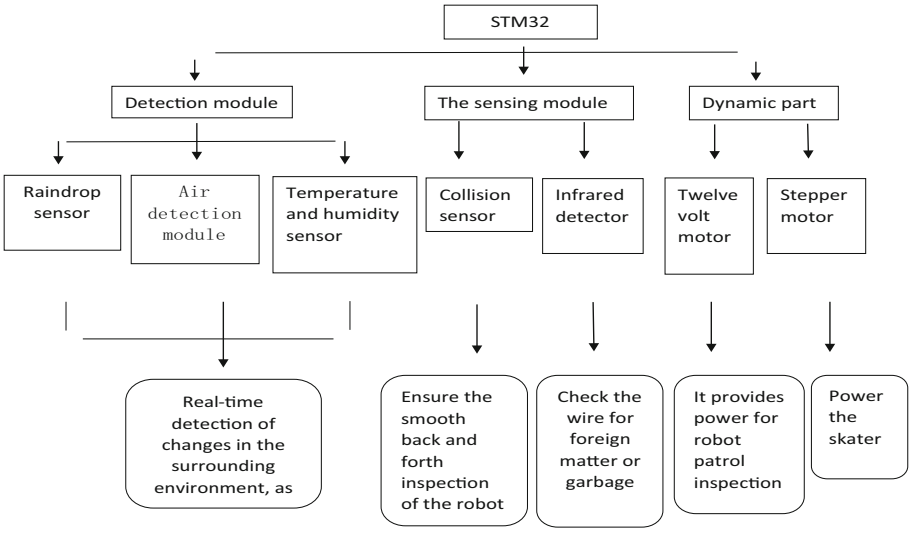



**Fig. 1.** Structure diagram of deicing device

As shown in Fig. 1, the device is a multi-functional device integrating patrol inspection, de-icing, weather information collection, early warning and feedback [7]. In this study, the development board of STM32 is used to control the device. When the device works, it walks back and forth on the high voltage wire, detects the operation condition of the wire, and detects the weather condition near the power line tower in real time, and gives feedback to the control center.

In order to quickly and effectively inspect the electric wires and clear the ice, the following methods are adopted in the design of the device:

(1) The roller is used to realize the device walking. There is a set of roller before and after the device. The roller is made of rubber material to increase friction and avoid

skidding. Each roller adopts an independent stepper motor, which can solve the problem of walking and braking by using the characteristics of large torque and high precision control force of stepper motor. At the same time, the front and rear two wheels can ensure the flexibility of the device. When an obstacle is encountered, a single roller can be used as a support, and another roller can cross the obstacle, so that the inspection range of the device is expanded.

(2) Visual identification technology and infrared detection technology are proposed to detect the external and internal problems of the device, collect the air and weather quality near the line, give early warning to some abnormal and special environments, and respond to problems in a timely manner.

(3) The de-icing part is an important part of the device. Closing the icing on the high voltage line can reduce many accidents along the high voltage line and prolong the service life. In the design part of skates, 3Done software is used to get the desired shape of skates through modeling.

(4) Since the device does not need to be inspected on the high-voltage line all the time, and the de-icing part is only used in winter, the de-icing part and the inspection part should be detachable. Convenient, easy to disassemble, and ensure firm and safe.

In this way, the device can walk smoothly on the high voltage line, collect and transmit the weather parameters along the line, quickly and effectively remove the ice on the high voltage line, and is easy to install and disassemble.

## 3 The De-icing Part

In order to adapt to the structure of the inspection device, this study chose the rolling and cutting de-icing. This approach is simple and effective, and does not require too much complex structure. The structure of deicing in this study mainly includes deicing knife, DC motor, gear, rotating shaft, infrared rangefinder and trap. In addition to ice skate design inspiration comes from soybean milk machine blade, break through the high-speed rotating blade attached to the wire on the ice, make its fall, this study aimed at this to design a special ice skates: the ice skates have four pieces of blade, blade [3] serrate, tip slightly bent upward, middle hollow, a circular ring (Fig. 2), can make the high tension line through; Because the ice cutter needs to be disassembled, it is divided into two equal parts, and a raised semicircle and a concave semicircle are added on both sides of the separate place, so that it can fit with each other and fix the position when it is installed. The “ring” of the ice-breaker extends back, and a groove is designed in the middle to fix the ice-breaker on the ice-breaker equipment. The first half is outside the equipment, and the second half is inside. And a design is made to change one part of the curved surface to a plane, so that the two halves of the ice-breaker can be stuck, and play a fixed role. The two halves of the ice-breaker can be fixed with a gear (as shown in Fig. 3), which is specially designed to fit with the structure of the ice-breaker. The gear is connected with the motor inside the equipment to provide power support for the ice-breaker.

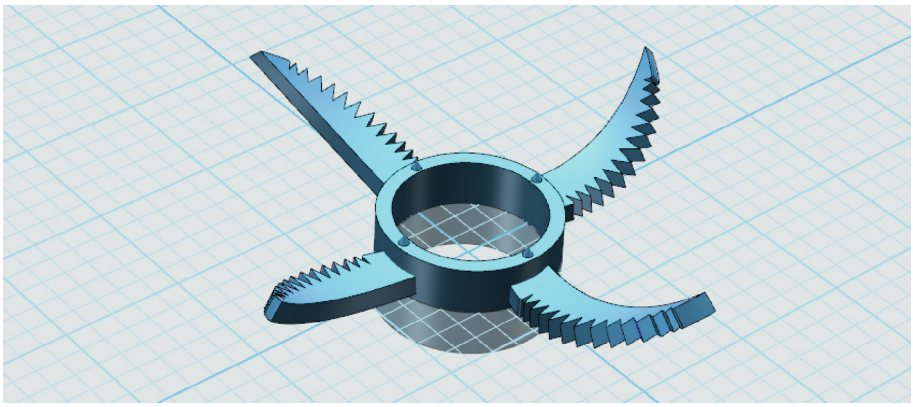


**Fig. 2.** Overall preliminary design of the skater

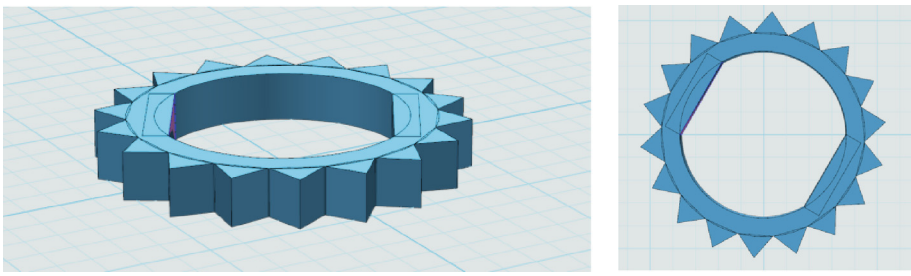


**Fig. 3.** Gear design modeling

In order to ensure the efficient operation of the ice-breaker, it is necessary to calculate the cutting force of the ice-breaker, so as to select the suitable motor power. We assume that the external temperature of deicing is $T = -10°C$, and then the tensile strength of the icing is $\sigma_b = 1.5$ MPa, and the shear strength of icing is $\tau_b = \sigma_b/1.64 = 0.92$ MPa [11]. The speed of the proposed device at work is 1 m/min, and the thickness of the icing is about 30 mm. According to the modeling, the length from the rotating shaft to the tool tip is about $D = 20$ mm, the thickness of the tool tip is $h = 2$ mm, the number of gears is $Z = 18$, and the speed at work is 120 rpm.

When the blade is fully in contact with the ice, the contact area is about 15 $mm^2$, the shear force of a single blade on the ice is:

$$F_\tau = \tau_b A_\tau = 0.92 \times 10^6 \times 15 \times 10^{-6} = 13.8\,N$$

The total shear force is:

$$F_\tau = F_\tau \times 4 = 13.8 \times 4 = 55.2\,N$$

According to the modeling data, the extrusion area received by a single blade is 66 mm, the extrusion pressure received by each blade is:

$$F_\sigma = \sigma_b A_\sigma = 1.5 \times 10^6 \times 66 \times 10^{-6} = 99\,N$$

The total extrusion pressure is:

$$F_\sigma = F_\sigma * 4 = 99 * 4 = 396\,N$$

The total cutting force acting on the four blades is:

$$F = \sqrt{F_\tau^2 + F_\sigma^2} = \sqrt{55.2^2 \times 396^2} = 399.8\,N$$

Required torque:

$$F = \sqrt{F_\tau^2 + F_\sigma^2} = \sqrt{55.2^2 \times 396^2} = 399.8\,N$$

Power of de-icing stepper motor:

$$P = M\omega = 4 \times 2\pi \times 120/60 = 50.24\,W$$

If the safety factor is 2.5, then the power of the motor of the ice-breaker is:

$$P = 50.24 \times 2.5 = 125.6\,W$$

## 4 Patrol Inspection

This study intends to adopt the method of visual identity to find some hidden trouble to the naked eye can see, its principle is to use computer to simulate human eyes recognise certain objects, and then through the computer information processing in a certain way, extract useful information from the image, intelligent identifying possible failure after processing, eventually used in actual measurement, detection and control; For parts or circuit inside some of the hidden trouble, this study intends to use the infrared detection technology to scan analysis (as shown in Fig. 4), the technology USES infrared detection of high pressure along the temperature of the device, failure occurs when the parts, in the vicinity of a fault can appear local warming, change the thermal distribution of the image, it can be found by sensor failure [10]. On the sensor, the device uses different kinds of sensors for different purposes: air detection sensor - monitoring and prevention of various media in the atmosphere [1] caused by chemical action of material damage and equipment damage, to ensure the normal supply of electric energy; Weather detection sensor -- real-time detection of nearby weather conditions, improve the economy and practicability of the device; Collision sensor -- in the process of inspection, when encountered obstacles, timely response to avoid obstacles.

In order to be able to can device in the process of checking the collected data and problems timely feedback to the engineers and monitoring center, this device to design communication module is divided into two parts, namely cable communication module and wireless communication module, cable communication device is mainly used for internal communication and data exchange between sensors, wireless communication is mainly used for devices with engineers, information interaction between test center, real-time data transmission and instructions.

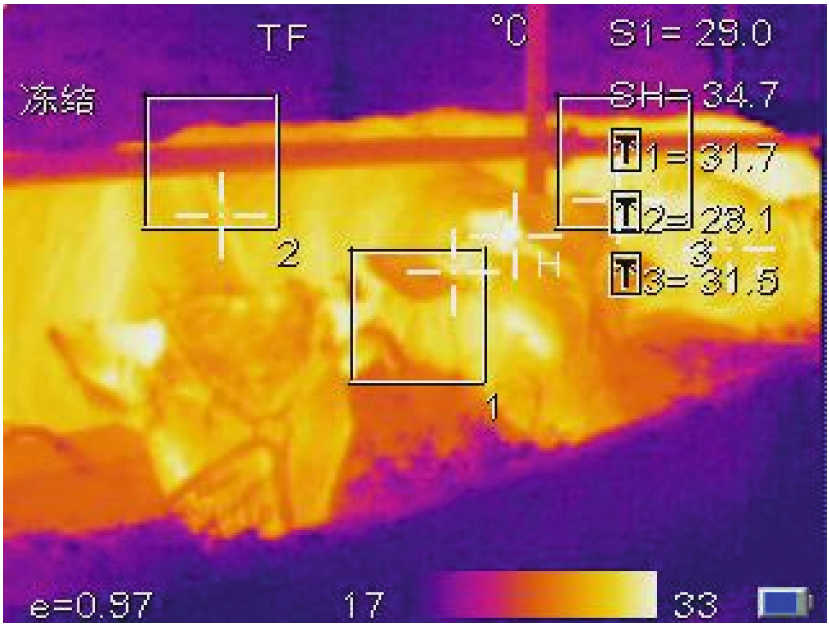



**Fig. 4.** Infrared detection

For the problem of high-voltage wire shaking, which affects the safety and stability of the inspection, the corresponding measures are also made. Some anti-shock hammers (as shown in Fig. 5 [9]) are installed near the hanging wire clip. When the wire is shaking, the anti-shock hammers will follow the shaking, producing a shaking out of sync or even opposite force, which can reduce the shaking of the wire [4].

Because of the functional design, the inspection device automatically completes the inspection and de-icing work, so the power supply should be matched with it. The conditions of the device determine that it is impossible to carry a large-capacity battery, so this study plans to install a solar charger to ensure sufficient power supply [8] (as shown in Fig. 6).

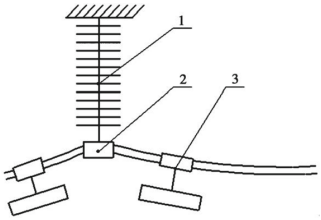



**Fig. 5.** Common obstacles on high voltage lines **1** suspension insulator **2** wire clip **3** shock hammer

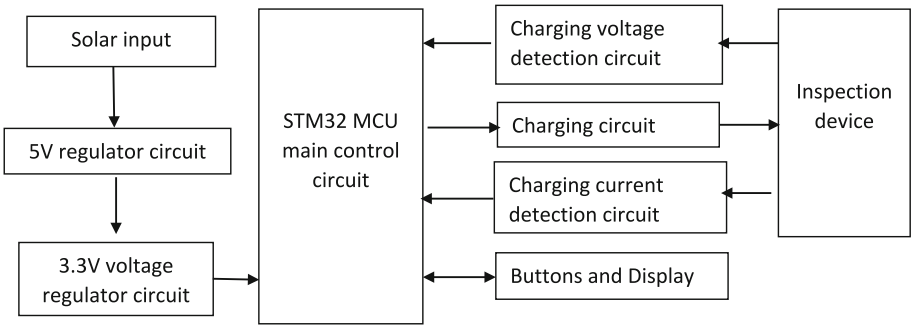



**Fig. 6.** Hardware connection diagram

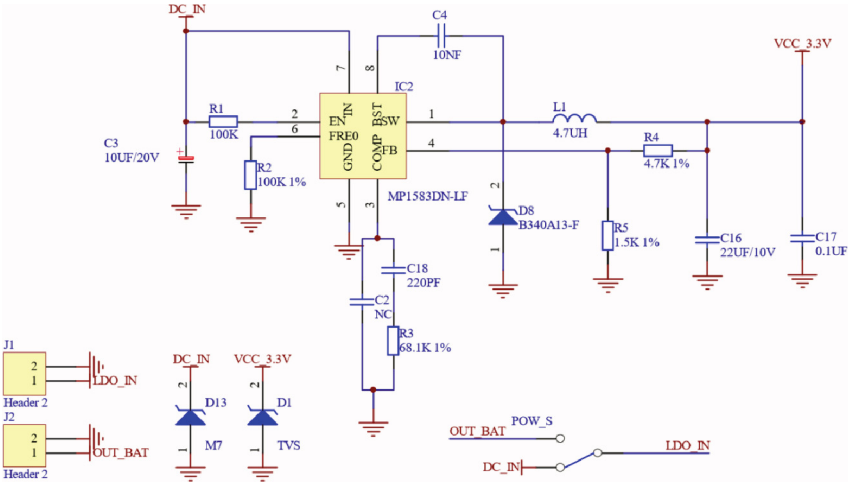



**Fig. 7.** STM32 MCU power module

## 5 Control System

This device uses STM32F103RBT6 chip of STM32F1 series. The characteristics of the chip are consistent with the function of the device, which can ensure the normal and stable operation of the device. The device system uses 220 V AC power supply to power, and then step-down, rectifier, filter voltage stabilization and other links to provide electric energy for other modules.STM32 single-chip microcomputer module adopts 3.3 V voltage (as shown in Fig. 7), other modules use 5 V voltage, and the stepping motor uses 12 V and 3.3 V voltage for power supply. The device uses DRV8825 chip as the drive chip of stepping motor. The chip provides a motor driven scheme for the automation of the inspection device. The chip has a special sleep mode, which allows the chip to be in a state of low power consumption as much as possible. Since the de-icing part is not used frequently, it can reduce the burden of power supply. At the same time, the chip can monitor the state of the chip in real time, and provide short-circuit protection,

over-current protection, over-temperature protection and under-voltage lock for the chip, so as to protect the motor and the chip [2].

## 6 Conclusions and Prospects

In this study, a patrol deicing device which works on a high voltage line is designed. The device can detect the working conditions along the line and find some early hidden dangers of the line, such as broken lines, rust, device loss and damage, etc. It can also detect temperature and humidity, the concentration of sulfur in the air, rainfall and other data along the route.

In this study, the de-icing part of the key research, the design and modeling of a convenient and reliable new de-icing knife; The de-icing knife adopts the physical de-icing method, which can effectively cut and shoot down the ice. At the same time, it has the advantages of simple structure, convenient disassembly and replacement. In this study, some advanced technologies are also integrated into the device, such as artificial intelligence, infrared detection technology, etc. These technologies have made some achievements in other fields, but they are seldom used in patrol inspection devices. It is hoped that better progress and promotion can be made in this regard.

Although the patrol deicing device designed in this study has made some progress, there are still many reasons that lead to the failure to make a breakthrough. It is a pity that the better results have not been shown due to the limitation of time. Later, we will continue to improve the existing deficiencies and break through more difficulties, hoping to provide some references for the intelligent patrol inspection.

## References

1. Dong, W., Chen, X., Fang, Y., Chen. J.: External insulation cleaning of power system should pay attention to matters. Technol. Outlook **26**,152 (2016)
2. Yanling, G., Jianxin, L., Jian, L.: Design of 3D printer system for wood plastic particles based on STM32. Modern Electron. Technol. **42**(18), 120–124 (2019)
3. Liang, X.,, Lian, Z., Lin, Y., Li, J.: Conception and structural design of a new type of ice cutter. Hubei Agricult. Mech. (06) (2020)
4. Lin, F., Chen, J., Zheng, W., Liu, X., Deng, L.: Damage and prevention of wire caused by seismic hammer under microtopography of transmission line. In: Proceedings of the 2018 Annual Conference of Fujian Electrical Engineering Society, 2018–10 (2018)
5. Li, Q., Fan, Z., Wu, Q., Gao, J., Su, Z., Zhou, W.: Investigation and accident analysis of national transmission line icing. Grid Technol. (**09**) (2008)
6. Li, Z., Liu, C., Yi, Y.: Application and prospect of intelligent technology in anti-ice and anti-ice work. Southern Power Grid Technol. **12**(06), 66–75 (2018)
7. Sun, F., Qi, J., Li, Y., Fan, J., Zhang, M: Transmission line video image recognition management system. China power, 2014–04 (2014)
8. Xu, Z., Sheng, S, Qin, C.: Solar charger based on STM32 microcontroller. Mech. Electr. Eng. Technol. (**11**), 144–146 (2019)
9. Zhao, L., Xiao, A., Huang, C., Du, X.: Design and mechanical analysis of passive obstacle crossing plate of high voltage line inspection robot. Forestry Mach. Woodworking Equip. **47**(02), 17–20+25 (2019)

10. Xiangqing, Z., Changyi, H., Wanhong, Z., Chuangping, Z.: Design of communication technology experiment system based on STM32 microcontroller. Experimental Technol. Manage. **36**(08), 81–84 (2019)
11. Qi, Z., Zhi, X., Zize, L., Jianlong, T., Dehui, S., Renxue, L.: Development and application of line patrol robot. Robot Appl. **02**, 37–42 (2007)

# Carrier Recovery Schemes of 16-QAM in Repeater-Less Transmission System

Yichao Zhang[1,2], Yupeng Li[1,2(✉)], Lei Li[1,2], Xiaoming Ding[1,2], and Xiaocheng Wang[1,2]

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
fx_lyp@163.com

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

**Abstract.** In this paper, an off-line processing scheme is proposed to solve the damage of 16-ary quadrature amplitude modulation (16-QAM) signal caused by laser linewidth and frequency deviation by adopting block approximate processing method without considering the reference value of ring segmentation. The simulation tool based on VPItransmissionMaker10.1 is equipped with the 16-QAM signal transmission system and the off-line DSP module is applied to MATLAB to prove the method proposed in this paper. The simulation results show that the proposed method can complete the compensation of frequency offset and phase offset, and recover 16-QAM signal with the symbol rate of 24-GBaud.

**Keywords:** 16-QAM · Viterbi algorithm · Symbol phase · Phase correction

## 1 Introduction

With the continuous emergence of new services such as cloud computing, big data and high-quality video, the pressure of network bandwidth is increasing, and operators have higher and higher requirements for single-fiber capacity. Compared with traditional optical communication (only intensity modulation of optical signal), coherent optical communication based on high-order modulation mode, coherent detection and digital signal processing (DSP) technology has become the main development direction of high-speed optical communication system because of its advantages such as high spectral efficiency and easy transmission damage compensation in the electrical domain.

In coherent optical communication system, there are many kinds of damage that affect optical signal. In order to transmit high quality information, it is necessary to compensate these damages. Carrier phase noise [1, 2] is one of the common damages of coherent optical communication system. For 16-ary quadrature amplitude modulation (16-QAM) signal, common compensation algorithms can be divided into two types: Viterbi based improved algorithm [3–6] and blind phase search (BPS) algorithm [7–9]. One representative of Viterbi based improved algorithms is carrier phase estimation and recovery through ring splitting, but it is easy to make mistakes in the steps of ring


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 259–265, 2023.
https://doi.org/10.1007/978-981-99-2362-5_32

splitting, resulting in inaccurate signal classification and affecting phase recovery. BPS algorithm uses the principle of enumeration method to rotate the phase, and obtains the noiseless signal closest to the ideal state through parameter setting. However, the accuracy is closely related to the complexity of the algorithm, and the higher the accuracy will increase the cost and power consumption of the receiver, so it is difficult to meet the requirements of real-time processing.

This paper presents an off-line processing scheme based on Viterbi algorithm to solve the signal damage caused by laser linewidth and frequency deviation in optical transmission system. The simulation results show that the damage compensation can be completed and 16-QAM signal (symbol rate is 24 GBaud) can be recovered by using the proposed scheme when the bit error rate (BER) is $1 \times 10^{-4}$.

## 2 Principle of 16-QAM

Figure 1 shows the constellation of 16-QAM signals to be generated. There are three electro-optic modulation schemes available: a cascade scheme based on I/Q modulator and phase modulator, a parallel scheme based on multiple I/Q modulators, and a scheme based on multi-base electric signals driving I/Q modulators [10–12]. The first two schemes are feasible in theory, but because the modulation scheme involves multiple electro-optic modulators, the clock synchronization between each modulator and the bit alignment between optical signal and driving signal will increase the complexity of system implementation. In addition, the linewidth of the carrier light source will introduce phase noise, and the phase noise damage based on the series or parallel 16-QAM generation scheme is more complex, which makes the digital signal processing at the receiving end unable to effectively compensate the damage. The more reliable generation scheme of 16-QAM modulation format is based on quad analog signal and achieved by I/Q modulator. This method involves fewer electro-optic modulators, low implementation complexity and has obvious advantages in signal quality. Therefore, this

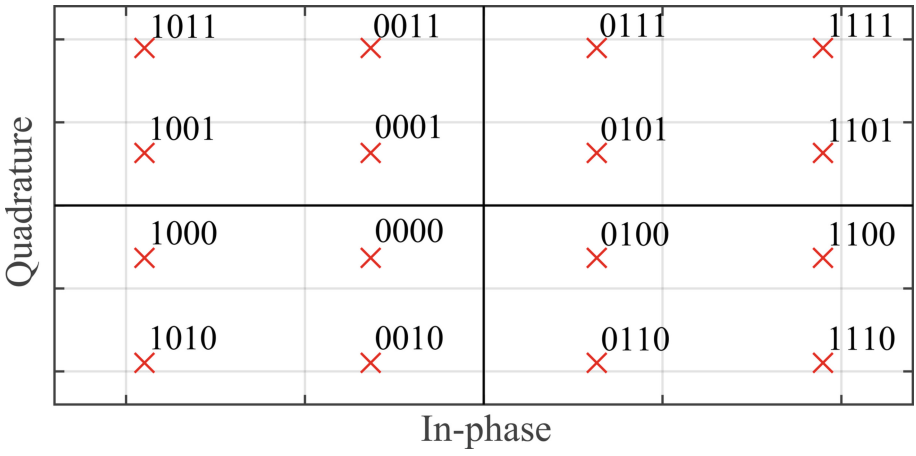



**Fig. 1.** 16-QAM signal constellation according to $\{\pm 1 \pm 1j, \pm 1 \pm 3j, \pm 3 \pm 1j, \pm 3 \pm 3j\}$.

paper uses the scheme based on multi-base electrical signals to drive the I/Q modulator to generate 16-QAM signal.

Figure 2 depicts the scheme of quadrature amplitude modulation based on I/Q modulator. For the quadrature phase shift keying (QPSK) modulation signal, the sub-bias bias point is at $V_\pi$, and the driving signal has a peak-to-peak value of $2V_\pi$ For 16-QAM signals, if the same setting is adopted: the sub-bias bias point is at $V_\pi$, and the peak-to-peak value of the driving signal is $2V_\pi$ The uneven distribution of 16-QAM signal constellation diagram, constellation points to concentrate in four locations, which is caused by transfer function curve is cosine type. If we want to solve this problem, we need to control the electric signal transfer function curve of the linear zone, or adjust the quaternary relative amplitude, the change of the analog signal level interval.

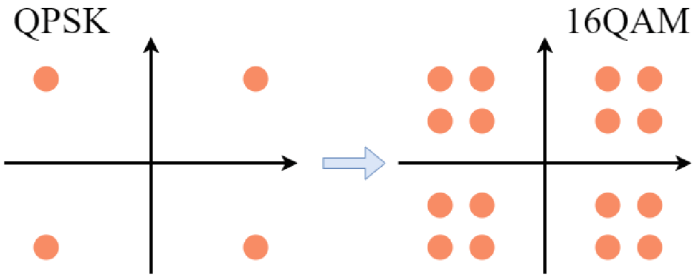



**Fig. 2.** QPSK and 16-QAM signal generation.

## 3 Compensation Principle of Carrier Recovery

In coherent optical communication system, the receiver uses the local oscillator (LO) laser to mix and balance the received signal to obtain baseband signal. Theoretically, the carrier frequencies of the two devices are required to be the same, but due to reasons such as optical device manufacturing, there is often a certain frequency deviation in practice. At the same time, the linewidth of the laser will introduce phase shift, which has a non-negligible impact on the signal. With the development of coherent optical communication technology, the above signal damage is generally compensated after signal polarization demultiplexing.

The following analysis is based on the previous signal damage compensation basically completed.

### 3.1 Frequency Offset Compensation

During off-line processing, after polarization demultiplexing is completed, we assume that the ***kth*** symbol sampling value of the received 16-QAM signal is

$$r(k) = A(k)\exp\{j(\theta_s(k) + 2\pi\,\Delta f kT + \theta_L(k) + \theta_n)\}. \quad (1)$$

$A(k)$ and $\theta_s(k)$ are the modulation information of the ***kth*** symbol. $\Delta f$ is frequency deviation. $T$ is the symbol period. $\theta_L(k)$ is the phase error introduced by laser linewidth. $\theta_n$ is phase noise.

The two symbols are processed by conjugate method as follows (ignore the amplitude):

$$r(k)r^*(k-1) = \exp\{j(\Delta\theta_s(k) + 2\pi\,\Delta fT + \theta_n)\}. \tag{2}$$

According to the characteristics of 16-QAM signal, it can be expressed as the following formula:

$$\Delta\theta_s(k) = \begin{cases} k \cdot \pi/4 \\ k \cdot \pi/4 \pm j \cdot \theta_{rot} \end{cases} \tag{3}$$

where $j$ and $k$ are integers, $\theta_{rot} = \pi/4 - \arctan(1/3)$.

Using the property that the mean of $\theta_n$ is zero and the average of the data before and after, we can get the following equation:

$$\sum_N r(k)^K r^*(k-1)^K/N \approx \exp\{j(K2\pi\,\Delta fT)\}. \tag{4}$$

This is because $K = 4n$ ($n$ is a positive integer), and if $N$ is a sufficiently large number, the following equation can be approximately equal

$$\Delta\theta(k) + \theta_n \approx \theta_n. \tag{5}$$

where, $\Delta\theta(k) = K\Delta\theta_s(k)$, the frequency offset estimation can be obtained after the angle calculation.

### 3.2 Phase Offset Compensation

On the basis of frequency offset estimation, the symbol sampling value can be expressed as

$$r'(k) = A(k)\exp\{j(\theta_s(k) + \theta_L(k) + \theta_n)\}. \tag{6}$$

According to the characteristics of uniform distribution of 16-QAM signals and block $M = 4m$ ($m$ is a positive integer), data processing here is similar to frequency bias, as shown in the following formula.

$$\arg\left\{\sum_M r'(k)^K/M\right\}/K = \theta_L. \tag{7}$$

After obtaining $\theta_L$, the problem of phase jump needs to be considered. Figure 3 illustrates the principle of detection by using the jump detection method [13]. $\Delta\theta_C$ is the phase difference between two data blocks. The phase jump correction is carried out by judging whether the phase of two data blocks exceeds the specified value.

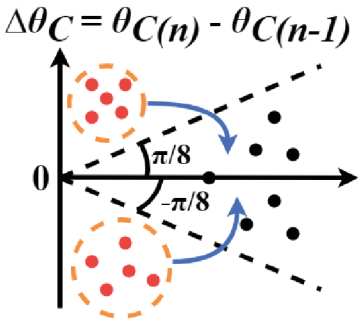



**Fig. 3.** Phase jump correction of symbols. $\theta_C(n)$ and $\theta_C(n-1)$ are the phases of adjacent data.

## 4 Simulation Results and Discussion

The simulation tool based on VPItransmissionMaker10.1 is equipped with the 16-QAM signal transmission system according to Fig. 4 (a), and the off-line DSP module is applied to MATLAB to prove whether the method proposed in this paper can compensate the signal damage caused by the frequency offset and linewidth of the lasers. The data is generated by pseudo-noise code to ensure the randomness of the data, and then encodes in 16-QAM modulation format to get the data. We set the signal symbol rate to be 24-GBaud.

After the coding is completed in MATLAB, the data is sent to the optical modulator to modulate the signal. Mach-Zehnder modulator (MZM) working mode is set to push-pull. The imbalance angle of I and Q channels is set to 20°. The linewidth and frequency offset of the lasers are set to 100 kHz and 500 MHz respectively. The effective area of the fiber sets as 80.0e-12 $m^2$, optical fiber attenuation is set to 0.2 dB/km. The scheme proposed in this paper mainly completes carrier recovery. In order to facilitate simulation experiment, we set the influence of chromatic dispersion (CD), polarization mode dispersion (PMD) and nonlinear effect to 0. The transmission distance is set to 100 km.

Figure 4 (b) shows the influence of the above signal damage parameters on 16-QAM signal. In off-line processing, we apply the method mentioned in this paper and set the values of K and M to 8 and 64, respectively. Figure 4 (c) shows constellation diagram of X-Pol after operation of the proposed scheme. According to the signal constellation diagram in Fig. 1, it can be found that data recovery is completed at this time.

In order to simulate attenuation damage and obtain different received optical power (ROP), variable optical attenuator (VOA) can achieve this effect by setting different values. We set 9 different groups of VOA parameters (20 dB to 28 dB). Figure 4 (d) shows the change of BER when ROP is from −27 dBm to −18 dBm and the laser linewidth is from 100 kHz to 300 kHz after using the compensation scheme mentioned in this paper. Through BER and constellation diagram of X-Pol state after data recovery, the effectiveness of the proposed scheme is proved.

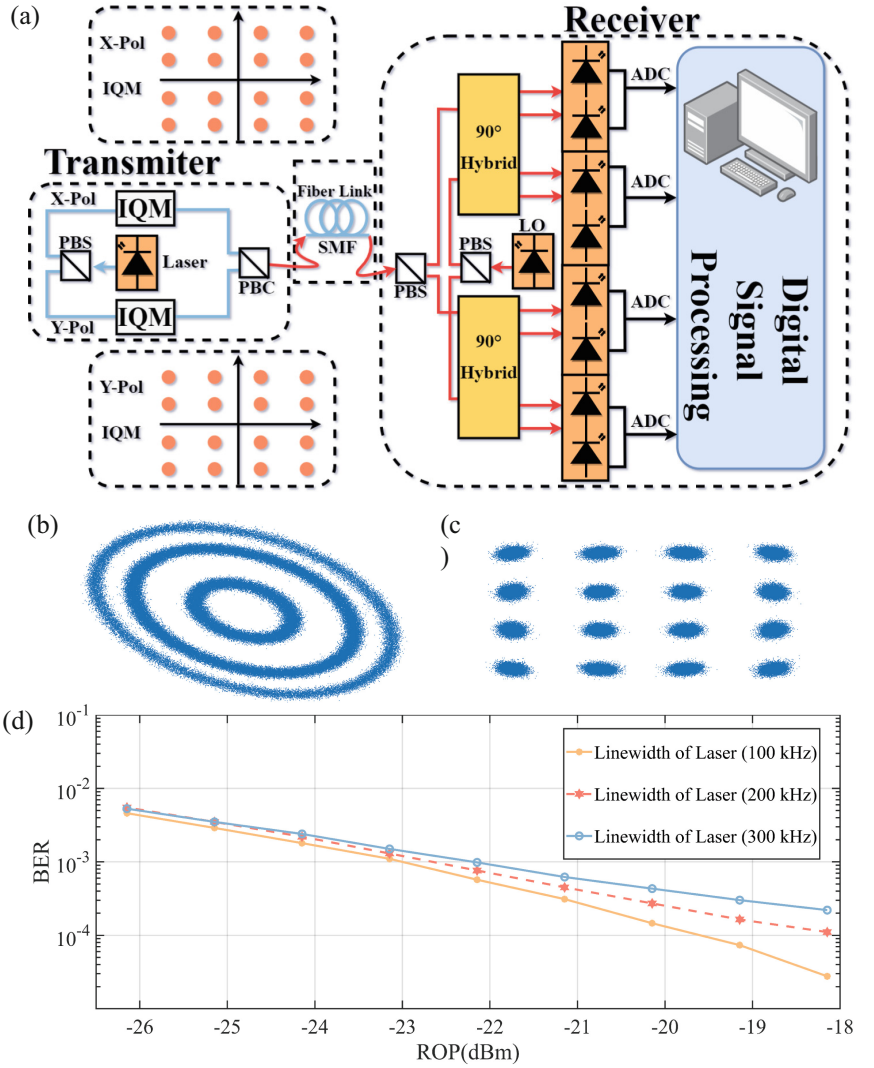



**Fig. 4.** (a) 16-QAM signal transmission system (transmitter, fiber link and receiver). (b) X-polarization (X-Pol) constellation of the receiver. (c) Constellation diagram of X-Pol after operation of the proposed scheme. (d) bit-error-rate (BER) and received optical power (ROP) with different linewidth of the lasers (100 kHz to 300 kHz).

## 5 Conclusion

In the transmission process of 16-QAM signal, the line width of the laser at the receiving and sending end is compensated. By using the compensation scheme mentioned in this paper, the linewidth of the laser at the receiving and sending is from 100 kHz to 300 kHz and the frequency deviation is 500 MHz to complete the data recovery, and the symbol

rate of 24-GBaud data transmission can be realized. In the case of more transmission bits and moderate frequency and phase deviation, the proposed scheme can be more practical than the traditional BPS algorithm and reduce the requirements on hardware.

**Acknowledgment.** This work was supported in part by the Natural Science Foundation of China under Grant 61901301, 62001328, 62001327.

## References

1. Zhang, S., Kam, P.Y., Yu, C., Chen, J.: Decision-aided carrier phase estimation for coherent optical communications. J. Lightwave Technol **28**, 1597–1607 (2010). https://doi.org/10.1109/JLT.2010.2048198
2. Yu, C., Zhang, S., Kam, P.Y., Chen, J.: Bit-Error rate performance of coherent optical M-ary PSK/QAM using decision-aided maximum likelihood phase estimation. Opt. Express **18**, 12088–12103 (2010). https://doi.org/10.1364/OE.18.012088
3. Fatadin, I., Ives, D., Savory, S.J.: Laser linewidth tolerance for 16-QAM coherentoptical systems using QPSK partitioning. IEEE Photonics Technol. Lett. **22**, 631–633 (2010). https://doi.org/10.1109/LPT.2010.2043524
4. Gao, Y., Lau, A.P.T., Yan, S.: Low-complexity and phase noise tolerant carrier phase estimation for dual polarizatin 16-QAM systems. Opt. Express **19**, 21717–21729 (2011). https://doi.org/10.1364/OE.19.021717
5. Zhong, K.P., Ke, J.H., Gao, Y., Cartledge, J.C.: Linewidth-tolerant and low-complexity two-stage carrier phase estimation based on modified QPSK partitioning for dual-polarization 16-QAM sSystems. J. Lightwave Technol. **31**, 50–57 (2013). https://doi.org/10.1109/JLT.2012.2227457
6. Fatadin, I., Ives, D., Savory, S.J.: Carrier-phase estimation for 16-QAM optical coherent systems using QPSK partitioning with barycenter approximation. J. Lightwave Technol. **32**, 2420–2427 (2014). https://doi.org/10.1109/JLT.2014.2326434
7. Li, J., Li, L., Tao, Z., Hoshida, T., Rasmussen, J.C.: Laser linewidth tolerant feed forward carrier phase estimator with reduced complexity for QAM. J. LightwaveTechnology **29**, 2358–2364 (2011). https://doi.org/10.1109/JLT.2011.2159580
8. Zhou, X.: An improved feed-forward carrier recovery algorithm for coherent receivers with M-QAM modulation format. IEEE Photonics Technol. Lett. **22**, 1051–1053 (2010). https://doi.org/10.1109/LPT.2010.2049644
9. Martins, C.S., Guiomar, F.P., Pinto, A.N.: Hardware optimization of dual-stage carrier-phase recovery for coherent optical receivers. OSA Continuum **4**, 3157–3175 (2021). https://doi.org/10.1364/OSAC.438524
10. Yu, J., Zhou, X., Gupta, S., Huang, Y., Huang, M.: A novel scheme to generate 112.8-Gb/s PM-RZ-64QAM optical signal. IEEE Photonics Technol. Lett. **22**, 115–117 (2010). https://doi.org/10.1109/LPT.2009.2036737
11. Lu, G.W., et al.: 40-Gbaud 16-QAM transmitter using tandem IQ modulators with binary driving electronic signals. Opt. Express **18**, 23062–23069 (2010). https://doi.org/10.1364/OE.18.023062
12. Winzer, P.J., Gnauck, A.H., Doerr, C.R., Magarini, M., Buhl, L.L.: Spectrally efficient long-haul optical networking using 112-Gb/s polarization-multiplexed 16-QAM. J. Lightwave Technol. **28**, 547–556 (2010). https://doi.org/10.1109/JLT.2009.2031922
13. Zhang, Y., Li, Y.: Modulation Schemes of 6PolSK-QPSK in repeaterless transmission system. In: Asia Communications and Photonics Conference (ACP), paper T4A.46 (2021) https://doi.org/10.1364/ACPC.2021.T4A.46

# Hybrid Attention Module Based on YOLOv5 for Foreign Object Debris Detection

Huan Lu[1], Tangyou Liu[1(✉)], and Jiafeng Zhang[2]

[1] College of Information Science and Technology, Donghua University, Shanghai, China
liuty@dhu.edu.cn

[2] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin, China

**Abstract.** Foreign object debris detection plays an important role in aircraft safety. Because the detection target is tiny and has limited feature information, foreign object debris detection is quite challenging. In this paper, we propose a new model called HAM-YOLOv5 for foreign object debris detection. Specifically, HAM-YOLOv5 is based on YOLOv5 to design hybrid attention module (HAM) for learning contextual information. For foreign object debris detection task, we add one prediction head for tiny object detection. Furthermore, we release Donghua University Foreign Object Debris Detection Dataset (DHFOD) to supplement the lack of current data samples. We conduct a serial of experiments on DHFOD. The experimental results show that our method outperforms other methods, which demonstrates the effectiveness of the proposed HAM-YOLOv5.

**Keywords:** Foreign object debris · CNN · Detection · Foreign object debris detection dataset

## 1 Introduction

With the development of aviation technology, the safety of aircraft is widely concerned. Foreign object debris is the foreign object on the aircraft runway, i.e. stones, metal pieces, paper products. It is sucked into the aircraft engine,, which results in indelible damage to the aircraft [1,2]. In order to improve aircraft safety, we need to detect FOD on aircraft runway. Nowadays, FOD detection relies on humans, which wastes human resources. Moreover, because of the small size of the FOD, human detection is often not accurate enough. Hence, automatic detection FOD methods are essential.

Because of the excellent characterization ability of convolutional neural network [3,4], the CNN-based methods become mainstream for detection. CNN-based detection methods are mainly categorized as one-stage [5,6] methods and two-stage methods [7]. In two-stage methods, Girshick [7] proposed the R-CNN,


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 266–272, 2023.
https://doi.org/10.1007/978-981-99-2362-5_33

which applies selective search to generate region of interest (ROI), and then utilizes the CNN to extract features for localization and classification on each ROI. Because of the high complexity of the model, two-stage detection methods are difficult to achieve real-time. In one-stage methods, YOLO series [8,9] plays an important role. One-stage detection methods achieve high real-time performance, but the detection accuracy needs to be improved.

There are two issues with directly using existing detection methods [10–12] for FOD detection, as shown in Fig. 1. First, FOD is learned with less feature information, which is interfered from the background for detection. Second, Foreign object debris belongs to small targets. The current detection methods are difficult to handle such objects because of the large difference in resolution between the detection target and the input image.

In this paper, we propose an improved method named HAM-YOLOv5, which is based on YOLOv5 to solve the above two issues for foreign object debris detection. We apply CSPDark-net53 [13] and PANet [14] as the backbone and neck of HAM-YOLOv5, respectively. For the FOD detection task, we add one prediction head for tiny object detection. Furthermore, we propose the hybrid attention module (HAM) in the head part, which can hybridize pixel attention and channel attention to enhance the discrimination of details.

**Fig. 1.** Examples to explain the challenge of foreign object debris detection

Our contributions can be summarized as follows:

1) We propose HAM-YOLOv5, which utilizes hybrid attention module to enhance the discrimination of details for accurate foreign object debris detection.
2) We release a new foreign object debris detection dataset DHFOD to address the lack of available data.
3) We conduct a series of experiments on DHFOD and the experimental results demonstrate the effectiveness of the proposed HAM-YOLOv5.

## 2 Approach

The framework of HAM-YOLOv5 is shown in Fig. 2. We modify YOLOv5 to yield HAM-YOLOv5 for FOD detection. YOLOv5 offers four different models depending on the number of model parameters, i.e., YOLOv5s, YOLOv5m, YOLOv5l

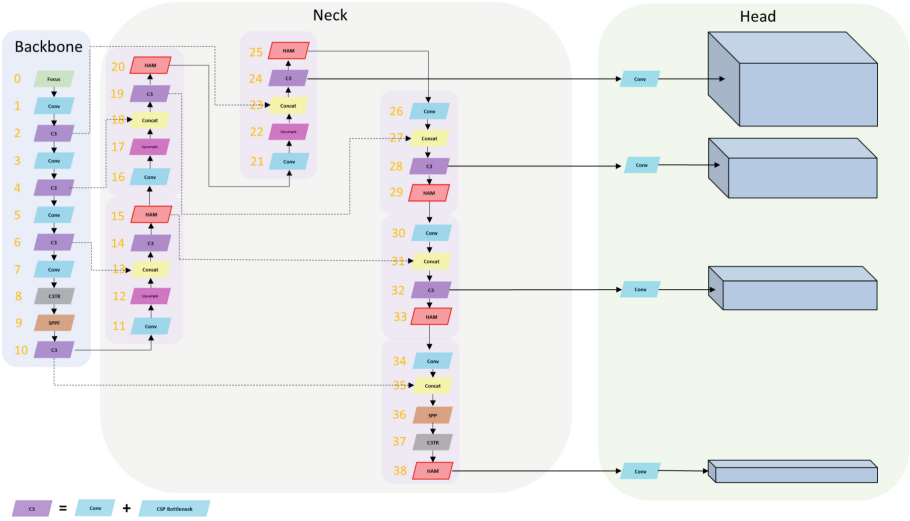



**Fig. 2.** The framework of the proposed HAM-YOLOv5.

and YOLOv5x. To balance accuracy and real-time, the HAM-YOLOv5 adopts the framework of YOLOv5l.

HAM-YOLOv5 consists of backbone, neck and head. Specifically, HAM-YOLOv5 uses CSPDarknet53 as the backbone for feature extraction. Afterwards, PANet is applied as Neck for feature fusion. Finally, we propose hybrid attention module to predict the detection results.

We analyze the dataset and find that foreign object debris are tiny objects. To enlarge the receptive field, feature maps tend to reduce the resolution as the network layers deepen. These tiny objects tend to lose information with reducing resolution, so we add one additional prediction head for tiny objects. From the figure, we can see that the additional prediction head is produced by shallow level features, which is more sensitive to tiny objects. Combined with the original three prediction heads, this captures information at different scales for accurate detection results.

### 2.1 Hybrid Attention Module

We replace some of the original convolutional layers in YOLOv5 with hybrid attention modules, which learns contextual information. Moreover, we only use the output from hybrid attention module for detection in the head part.

The structure of HAM is shown in Fig. 3. HAM considers the spatial and channel dimension information of feature maps and fuses them to obtain contextual information. Specifically, the input feature maps $H \times W \times C$are fed into two branches, where $H$, $W$ and $C$ are the height, the width and the channel number, respectively. Each branch has global pooling layers for high or wide

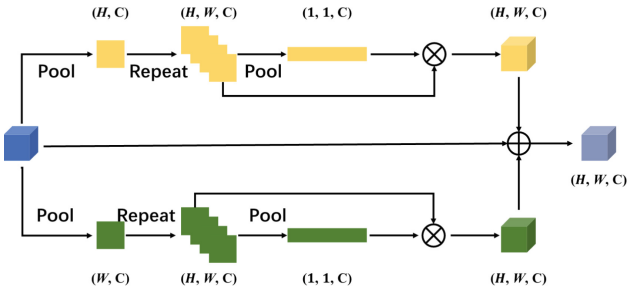



**Fig. 3.** The flowchart of the proposed hybrid attention module.

dimensions. Each with global pooling layers for high and wide dimensions, which yields two feature maps. For each feature maps, we repeat it to obtain feature maps with the same size as the input. Afterwards, we use channel attention to capture the channel information. Finally, these feature maps are fused, which produces the output of the hybrid attention module.

## 3 Foreign Object Debris Detection Dataset

Data samples are critical to the deep learning model for training, which determines the generalization of the model to the real world. Currently, foreign object debris detection methods are limited by the number of datasets. Hence, we release a new foreign object debris detection dataset DHFOD. DHFOD is available to download at https://github.com/Kinoyeye/DHU-Large-Scale-Foreign-Object-Debris-Detection-Database. DHFOD consists of 224 images, which are collected by technicians in Shanghai. DHFOD is randomly distributed into 160 training images and 64 test images. All images are captured on DHFOD by the on-board industrial camera and stored in BMP format with the resolution of $1920 \times 1080$. We illustrate some sample images as shown in Fig. 4

## 4 Experiment

In this section, we evaluate the performance of the proposed HAM-YOLOv5 on DHFOD. Specifically, we first present experimental in details. Afterwards, we conduct a series of experiments to demonstrate the effectiveness of the proposed HA-YOLOv5.

### 4.1 Implementation Details

In our experiments, data enhancement is performed before the input images are fed into the model. The data enhancement includes random flip, illumination distortion, geometric distortion, etc. We use stochastic gradient descent (SGD) [15] as the optimizer. Furthermore, some experimental hyperparameters are shown in Table 1.

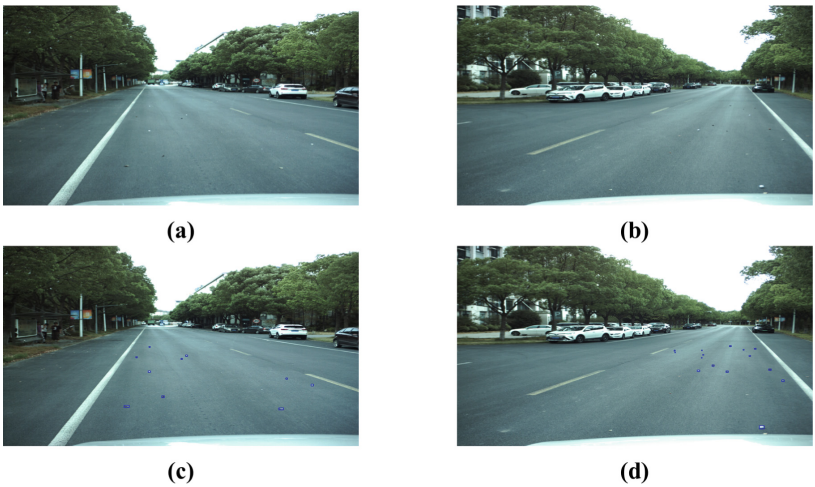



**Fig. 4.** Some foreign object debris detection images and annotated images with ground truth.

**Table 1.** Some Experimental Hyperparameters

| **Hyperparameters** | **Value** |
|---|---|
| Epoch | 300 |
| Batch Size | 2 |
| Initial Learning Rate | 0.01 |
| Momentum | 0.937 |
| Weight Decay | 0.0005 |

### 4.2 Comparisons with State-of-the-Art Methods

We compare the proposed HAM-YOLOv5 with other methods. The experimental results are shown in Table 2. Several conclusions are concluded from the table. Firstly, our method achieves the best results, which demonstrates the effectiveness of HAM-YOLOv5. Secondly, both HAM-YOLOv5 and TPH-YOLOv5 are better than YOLOv5, which indicates the added prediction head is beneficial for small target detection. Finally, yolov5-based methods outperform YOLOv3, proving the effectiveness of YOLOv5. As shown in Fig. 5, some visualization results of the proposed HAM-YOLOv5 on DHFOD.

**Table 2.** Comparison of Different Methods Performance on DHFOD

| Methods | mAP |
|---|---|
| YOLOv3 [5] | 42.09 |
| YOLOv5 | 47.52 |
| TPH-YOLOv5 [8] | 52.85 |
| HAM-YOLOv5 | 56.75 |

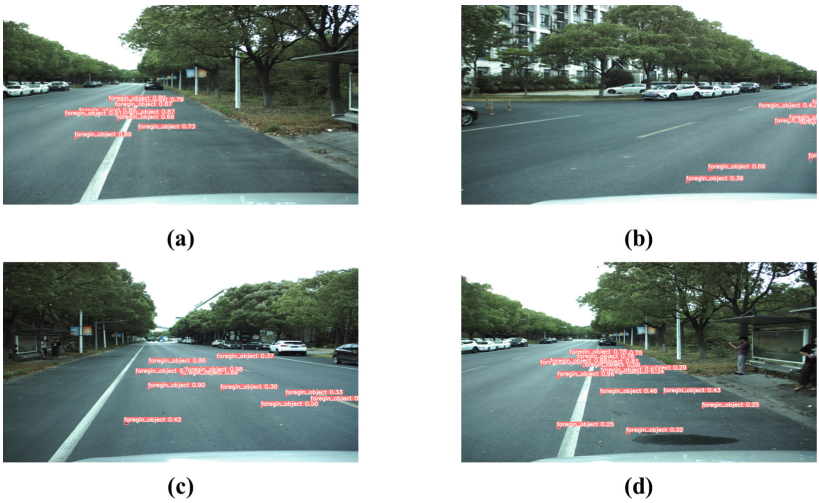



**Fig. 5.** some visualization results of the proposed HA-YOLOv5 on DHFOD.

## 5 Conclusion

In this paper, we have proposed HAM-YOLOv5 for foreign object debris detection. Specifically, we use backbone to extract features from the input image. Afterwards, we apply the neck to fuse the features from the backbone output. We propose hybrid attention module (HAM) in the head, where HAM can comprehensively fuse channel information and pixel information for accurate detection. We additionally add one prediction head for fine object detection. Furthermore, we release Donghua University Foreign Object Debris Detection Dataset (DHFOD). Finally, we have conducted some experiments on DHFOD, and the experimental results show the effectiveness of our method.

## References

1. Yigit, E., Demirci, S., Unal, A., Ozdemir, C., Vertiy, A.: Millimeter-wave ground-based synthetic aperture radar imaging for foreign object debris detection: experimental studies at short ranges. J. Infrared, Millimeter, Terahertz Waves **33**(12), 1227–1238 (2012)

2. Qunyu, X., Huansheng, N., Weishi, C.: Video-based foreign object debris detection. In: IEEE International Workshop on Imaging Systems and Techniques. IEEE 2009, pp. 119–122 (2009)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778 (2016)
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014)
5. Redmon, J., Farhadi, A.,: Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018)
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788 (2016)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587 (2014)
8. Zhu, X., Lyu, S., Wang, X., Zhao, Q.: Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2778–2788 (2021)
9. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
10. Cao, X., et al.: Region based cnn for foreign object debris detection on airfield pavement. Sensors **18**(3), 737 (2018)
11. Xu, H., Han, Z., Feng, S., Zhou, H., Fang, Y.: Foreign object debris material recognition based on convolutional neural networks. EURASIP J. Image Video Process. **2018**(1), 1–10 (2018). https://doi.org/10.1186/s13640-018-0261-2
12. Ni, P., Miao, C., Tang, H., Jiang, M., Wu, W.: Small foreign object debris detection for millimeter-wave radar based on power spectrum features. Sensors **20**(8), 2316 (2020)
13. Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H.: Cspnet: A new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 390–391 (2020)
14. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768 (2018)
15. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics, pp. 400–407 (1951)

# Sleep Quality Analysis Based on Machine Learning

Lifeng Tian(✉)

Honor Device Co. Ltd, Shenzhen 518000, China
1210333493@qq.com

**Abstract.** One third of a person's life is spent in sleep, sleep is a need of life, and the quality of sleep plays a vital role in human health. With the accelerated pace of life, the average sleep duration of human beings has continued to decline over the past time. Poor sleep quality has led to serious social problems. According to some studies, short-term lack of sleep can lead to distraction and lack of energy, while long-term lack of sleep can lead to an increased risk of many fatal diseases. However, today's monitoring of sleep quality is far from adequate. For example, sleep diaries are subjective and the polysomnography monitoring process is too cumbersome. Based on this, this paper proposes a convenient, fast and low-cost sleep quality detection system based on the Internet of Things. This paper investigates the relationship between cardiopulmonary coupling (CPC) signals and sleep stages and various sleep disorders. The algorithm is based on MATLAB programs such as constitutional neural networks and disease index prediction models.

**Keywords:** Sleep Quality · Cardiopulmonary Coupling · Constitutional Neural Networks

## 1 Introduction

Sleep is an important way for humans to maintain normal physiological functions. In particular, deep and steady sleep helps the body to recover and restore energy. In addition, sleep quality is closely related to both mental and physical health. Under normal circumstances, an adult needs no less than 7 h of sleep. However, recent surveys have found that higher than 35% of people sleep for less than seven hours [1]. If these sleep-deprived people work or study the next day, their risk of a car accident or mistakes is greatly increased.

In addition, patients with sleep disorders are also a large group of poor sleep quality [2]. According to statistics, in the United States, about 50 million to 70 million people suffer from sleep disorders. Obstructive sleep apnea (OSA) is considered a representative.

Mild sleep deprivation in humans can also lead to memory loss, affecting judgment and mood. In addition to psychiatric effects, chronic sleep deprivation is also associated with other serious physical conditions such as obesity and hypertension [3]. Additionally, obesity and high blood pressure can worsen sleep disturbances, creating a dangerous cycle for human health.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 273–279, 2023.
https://doi.org/10.1007/978-981-99-2362-5_34

Early sleep quality evaluation usually follows polysomnography to infer the patient's sleep, that is, collect the patient's physiological signals from multiple channels, and then conduct professional analysis by experts to assess sleep quality [4].

However, PSG has some drawbacks. First, it relies on an experienced technician spending hours manually identifying the characteristic waveforms on the EEG and stage sleep. Second, patients often spend considerable time waiting in line for PSG testing. Moreover, a complete PSG requires at least 11 channels and 22 wires to be connected to the patient, and a large number of recording wires in the whole body will make the patient feel very uncomfortable and affect the quality of sleep [5]. Finally, due to the influence of psychological factors, some patients may feel uncomfortable when sleeping in an unfamiliar environment, and the symptoms of the first night may affect the sleep quality of patients [6].

This report proposes a sleep analysis method based on CPC, which is also part of the Medical Internet of Things (M-IoT). This project aims to use a small wearable recorder to obtain the patient's Electrocardiography (ECG) signal. The patient only needs to wear the ECG acquisition sensor at the bedside, and the device can then measure the single-lead ECG signal at a certain sampling rate. The collected ECG signals can be uploaded to the mobile terminal through Bluetooth, and then converted into CPC signals on the cloud platform, which can be used to measure sleep stages and evaluate patients' potential sleep disorders/diseases. Finally, the patient will be given feedback on sleep health status and corresponding professional medical advice (e.g. report email sent to the patient's mobile terminal).

## 2 CPC

Sleep is a physiological process, and changes in physiological characteristics are relaxation of the brain (eg, muscle relaxation, lower heart rate, slower breathing, stabilization). The most representative of these is the rhythm characteristic of the heart, which changes with the different states of the human body. The electrical activity of the heart also exhibits different characteristics in different sleep states.

In 2015, the Harvard Sleep Analysis Group reported a special coupling between the human heart rate loop system and breathing changes during sleep. Based on this discovery, R.J. Thomas and the Harvard team first proposed the concept of cardiopulmonary coupling, called CPC [7].

As shown, CPC acquisition first uses the PanTopkin algorithm to detect QRS complexes. Then the RR interval sequence was extracted from the ECG signal, normal heartbeats and abnormal heartbeats were identified, outliers were removed, and then the normal sinus NN interval sequence was selected. Then, the respiration signal time series (EDR) and heart rate variability (HRV) time series was extracted from the ECG by the "area method", and the series was resampled to 4 Hz by cubic spline interpolation. Finally, a 5-min window is set, that is, a fast Fourier transform is performed on each sample window, the coherence and cross power spectrum of the two signals are calculated, and the product of the cross power spectrum and the coherence is used as the cardiopulmonary coupling value.

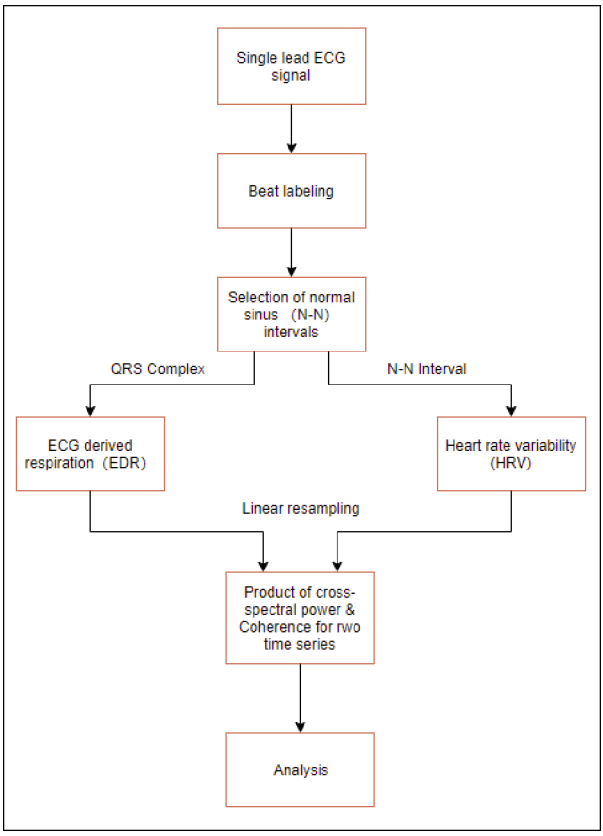



**Fig. 1.** CPC acquisition process

Heart rate variability (HRV) indicates the small fluctuations in instantaneous heart rate between consecutive sinus heartbeats [8], i.e., the physiological phenomenon regarding the time interval between heartbeats. The QRS complex reflects the changes of the left and right ventricular depolarization potential and time, and can be used for the calculation of EDR.

Breathing disorders during sleep can be divided into apnea and hypopnea. Apnea means that airflow from the nose and mouth stops for more than 10 s, and hypopnea means that ventilation is reduced by more than 50%. The AASM PSG manual states that sleep apnea/hypopnea patients have an average of 5 or more sleep apnea/hypopnea per night and can be diagnosed as sleep apnea syndrome. Sleep apnea syndrome can be divided into 2 categories. They are obstructive sleep apnea (OSA) and central sleep apnea (CSA).

OSA, the most common form of sleep apnea in the general population, increases the risk of stroke, cardiovascular disease and other complications. OSA occurs due to increased upper airway resistance [9].

CSA is when the brain temporarily stops sending signals to the muscles that control breathing, resulting in a weakening or lack of freezing of breathing, usually manifested in periodic drops of 10 to 30 s. CSA is usually caused by destabilization of the human feedback mechanism that controls breathing [10].

The Harvard Sleep Analysis team also compared apnea markers in the Physical Network Sleep Apnea Database with the Elevated low frequency coupling(e-LFC regions identified by the CPC algorithm. The results show that the e-LFC region in the CPC signal and the location in the database are highly consistent, so the two are significantly correlated. Therefore, e-LFC can be used as a physiological feature to judge sleep apnea. The e-LFC is further split into high and low frequency coupled narrowband (e-LFC_NB) and high and low frequency coupled broadband (e-LFC_BB). e-LFC_BB $> \geq 15\%$ indicates the existence of obstructive sleep disorder, while e-LFC_NB $> 2\%$ indicates the existence of central sleep-disordered breathing.

CPC has many advantages over PSG. CPC is based on an organic combination of heart rate variability during sleep and the extraction of respiratory information from ECG signals. The coupling strength of the two is used to quantitatively evaluate the breathing movement during sleep, which overcomes the shortcomings of the two when applied alone. In addition, the analysis method of CPC is simpler and more convenient than PSG. Unlike PSG, which needs to acquire multiple channels of physiological signals, CPC detection only requires the patient's single-lead ECG, which can greatly reduce the impact of the detection process on sleep, making it easier for patients to accept. Moreover, the price of CPC equipment is favorable. After purchasing an ECG signal capture device, it can also be reused for multiple sleep quality tests. Multiple collections of sleep data mean that sleep data is objective and real, eliminating the effects of abnormal nighttime caused by other factors. In addition, multiple sleep quality tests mean that clinicians can update patients' sleep health status at any time and make reasonable diagnoses in a timely manner. Thereby, the accuracy of CPC is greatly improved.

## 3 CNN

This project first collects the CPC signal. The equipment used is shown in the Fig.1, a smart ECG sticker called "Real Nursing" produced by Beijing Real Vision Company. The device is small in size, with a thickness of 0.8 cm and a weight of 13 g. Patients need to connect the device to the official app in their mobile terminal via Bluetooth, wear the device while sleeping, and collect ECG signals, which are then uploaded by iRealcare Fig. 2.


**Fig. 2.** Irealcare ECG detecting sensor.

Then calculate the CPC spectrum. For cardiopulmonary coupling analysis, suppose the RR interval time series extracted from the ECG is x(t) and the respiratory signal time series is y(t), the correlation between two random signals can be described by a cross-correlation function. The correlation between sequence y(t) and x(t) at time t is determined by taking the limit of the average product, then the cross-correlation function of these two signals is expressed as:

$$R_{xy}(\tau) = \lim_{\tau \to \infty} \int_0^{\tau} x(t)y(t+\tau)dt, \tag{1}$$

Their cross-power spectrum function is the Fourier transform of the cross-correlation function:

$$S_{xy}(w) = \int_{-\infty}^{\infty} R_{xy}(\tau)e^{-jw\tau}d\tau, \tag{2}$$

If $R_{xy}(\tau)$ and $S_{xy}(w)$ are the energy power spectra of X and Y respectively, then the coherence coefficients of these two sequences is:

$$C_{xy} = \frac{|R_{xy}(w)|^2}{R_{xx}(w) * R_{yy}(w)}, \tag{3}$$

Apply the Cpsd function and Mscohere function in Matlab to calculate the cross power spectrum and coherence. Finally, the CPC spectrum is obtained.

$$\mathrm{CPC}(w) = C_{xy}(w) * \left[S_{xy}(w)\right]^2, \tag{4}$$

Finally process the data. The CNN toolbox used in this study was written in Matlab R2018. The model composition is shown in the Fig. 3.

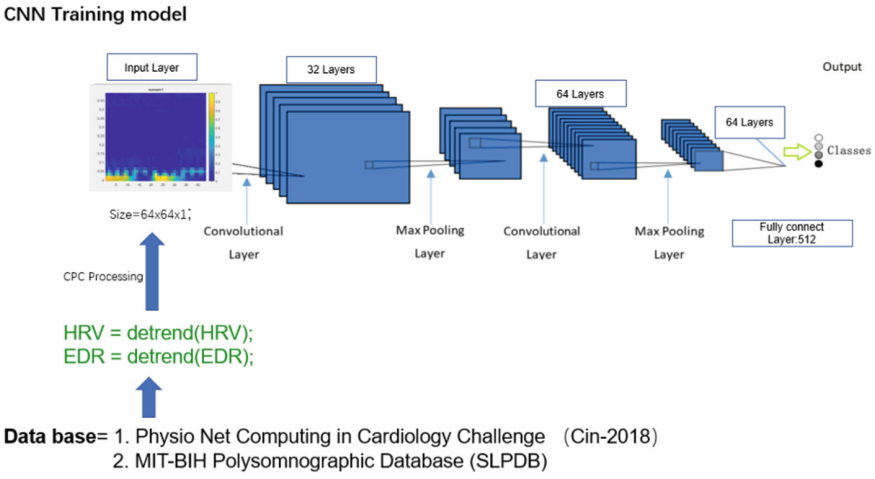



**Fig. 3.** Structure of CNN

The first is an input layer with an input data size of $64 \times 64 \times 1$. All cut data will be put into the input layer, which goes through three convolutional layers. Each convolutional layer is followed by a batch normalization layer, where the data is Gaussian. The third is a max pooling layer with stride 2 to increase the receptive field of the model. The max pooling layer downsamples the input twice in both directions, thereby compressing the data to 1/4 and preserving the most obvious categorical features. After the last max pooling layer, a fully connected network with a single hidden layer of 512 neurons is added to classify the extracted features. Finally, the output is mapped to the interval (0,1) by the SoftMax activation function, and the cumulative sum of all elements is 1. The SoftMax function obtains the probability of each category corresponding to the input data, and selects the category with the highest probability as the prediction result.

## 4 Results and Discussion

In the present report, two representative subjects A01 and B01 were selected for analysis. The label shows that subject A01 suffered from sleep disorders, while the subject B01 is in healthy condition. The obtained e-LFC ratio and the SAI percentage illustrate that the sleep indicator parameter of subject A01 has the feature of CSA and OSA, while the sleep indicator parameter of subject B01 is normal. That is same as diagnosed result in the database Fig. 4.

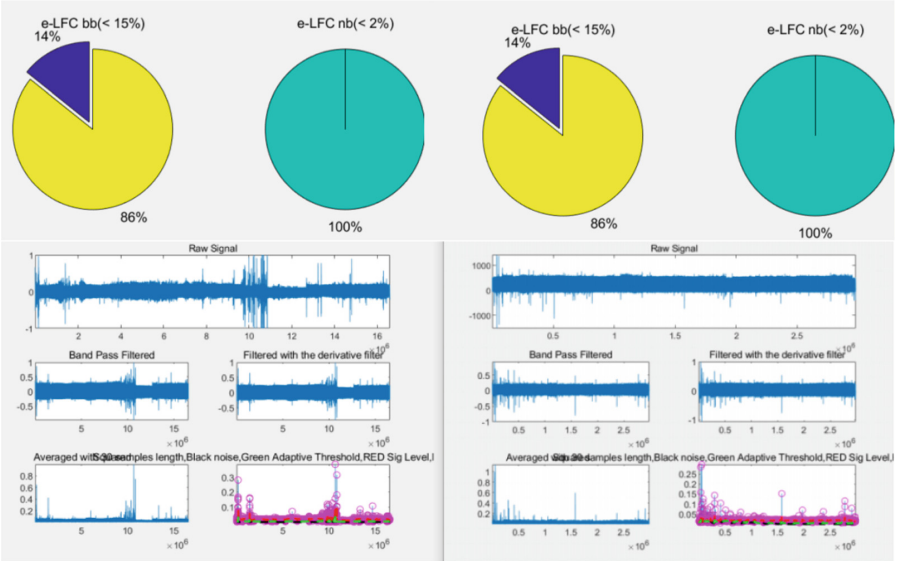



**Fig. 4.** comparison between A01 and B01.

## 5 Conclusion

This paper combines the CPC concept with the IoT system, which is a low-cost, high-accuracy detection solution. This scheme greatly improves the efficiency of sleep quality detection. This article introduces the importance of sleep quality, the current popular sleep quality detection schemes and the concept of CPC. In the design part, this paper proposes a new IoT-based sleep quality detection system. The system consists of ECG data acquisition and uploading part, data processing and data feedback. Finally, two subjects were selected and tested, and the results were in line with expectations.

## References

1. “CDC Press Releases”, CDC, 2020. https://www.cdc.gov/media/releases/2016/p0215-enough-sleep.html. Accessed 07 May 2020
2. Apa.org, 2020. https://www.apa.org/news/press/releases/stress/2013/sleep. Accessed 07 May 2020
3. Spira, A., et al.:Self-reported Sleep and β-Amyloid Deposition in Community Dwelling Older Adults",JAMA Neurology (2013). https://doi.org/10.1001/jamaneurol. 2013. 4258 Accessed 7 May 2020
4. Flemons, W.W., Douglas, N.J., Kuna, S.T., Rodenstein, D.O., Wheatley, T.: Access 51to diagnosis and treatment of patients with suspected sleep apnea. Amer. J. Respir. Crit. Care Med. **169**(6), 668–672 (2004)
5. "Aystesis Clinic | Polysomnography sleep disorders diagnosis | Costa Rica | AYSTESIS CLINIC", Aystesis.com, 2020. https://www.aystesis.com/polysomnography/. Accessed 24 May (2020)
6. Agnew, H.W., Webb, W.B., Williams, R.L.: The first night effect: an EEG study of sleep. Psychopharmacology **2**(3), 263–266 (1966)
7. Sannino, G., De Falco, I., De Pietro, G.: Monitoring obstructive sleep apnea by means of a real-time mobile system based on the automatic extraction of sets of rules through differential evolution. J. Biomed. Inform. **49**, 84–100 (2014)
8. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. **85**(2), 461–464 (2000)
9. Young, T., Peppard, P., Gottlieb, D.: Epidemiology of obstructive sleep apnea. Am. J. Respir. Crit. Care Med. **165**(9), 1217–1239 (2002)
10. Eckert, D., Jordan, A., Merchia, P., Malhotra, A.: Central sleep apnea. Chest **131**(2), 595–607 (2007)

# Measurements and Analysis for the Second-Order Statistical Properties of Time-Variant A2G Channels

Kai Mao[1,2], Taiya Lei[1], Yanheng Qiu[1], Qiuming Zhu[1(✉)], Maozhong Song[1], and Yang Miao[2]

[1] Key Laboratory of Dynamic Cognitive System of Electromagnetic Spectrum Space, College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
{maokai,leitaiya,qiuyanheng,zhuqiuming,smz108}@nuaa.edu.cn

[2] The Radio Systems, Faculty of Electrical Engineering Computer Science and Mathematics (EEMCS), University of Twente, Enschede, The Netherlands
y.miao@utwente.nl

**Abstract.** Channel measurements are the fundamental to accurately learn the realistic channel properties. In this paper, we demonstrate our latest UAV-aided air-to-ground (A2G) channel sounder. Several channel data processing methods, i.e., system response elimination (SRE), sampling time offset compensation (STOC), and multi-path component recognition (MPCR) are also proposed to increase the accuracy of measured channel properties. Moreover, the second-order statistical properties of channel small-scale fading (SSF) such as autocorrelation function (ACF), cumulative distribution function (CDF), and level crossing rate (LCR) are derived based on the measurement data. Finally, time-variant properties of second-order statistics are analyzed according to the measurement data obtained at 3.5 GHz in a near-urban scenario. The results can be used for validating the performance of channel models and assisting the design of A2G communication systems under the similar scenario.

**Keywords:** A2G communication · channel measurement · channel model · second-order statistical properties

## 1 Introduction

Unmanned aerial vehicle (UAV) has been widely used on many applications such as military reconnaissance, earthquake relief, forest fire detection, emergency rescue, etc., due to its flexible deployment, high mobility, and low cost. UAV communication technology becomes an important part in the beyond fifth generation (B5G) and sixth generation (6G) communication networks with the ability to build line-of-sight (LoS) links easily [1,2]. It is vital to establish a good communication link to ensure the effectiveness and reliability of the UAV communication system by well understanding the propagation channels. Compared with traditional terrestrial communications, the air-to-ground (A2G) communication


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 280–287, 2023.
https://doi.org/10.1007/978-981-99-2362-5_35

has more obvious time-variant channel properties [3]. Channel measurement is the most accurate way to capture the realistic channel. Therefore, it is crucial to carry out A2G channel measurements and analyze the channel properties.

The vector network analyzer (VNA) is a common instrument used to measure the frequency domain characteristics of wireless channel, and the channel impulse response (CIR) can be obtained by using inverse fast Fourier transform (IFFT) [4]. This method has high system reliability but is hard to be utilized to measure A2G channel due to the bulkiness of VNA and the limitation of measurement distance resulting from the wire connection between the link ends and the same VNA. Therefore, separate transceivers are more often used in the A2G channel measurements. Recently, A2G channel measurement campaigns have gained a lot of attentions [5–8]. For example, some portable A2G channel sounders were designed and developed for the platforms as the quadcopter UAV [5], hexacopter UAV [6,7], and octocopter UAV [8]. However, there are still many challenges needed to be tackled. Firstly, the sampling time offset (STO) of transceivers due to the deviation of crystal oscillators could lead to power loss on the sliding correlation results [9]. Secondly, a constant noise threshold is often used to extract the valid multi-path components (MPCs) from the continuous CIR, which is likely to remove low-power multi-paths or bring in some fake multi-paths [10,11]. The last but not least, only few research works analyzed the second-order statics of A2G channel small-scale fading (SSF) in detail [12,13], where there are still some gaps.

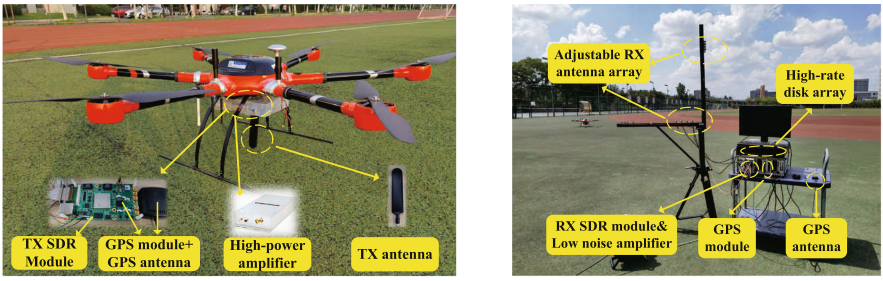
To address these issues and fill the gaps, an A2G measurement system with several channel data pre-processing methods, i.e., system response elimination (SRE), STO compensation (STOC), and MPC recognition (MPCR) is proposed and verified in this paper. Moreover, several second-order statistics of channel SSF are derived based on the discrete measurement data, and time-variant second-order statistics are analyzed in a near-urban measurement scenario. In Sect. 2, the developed channel measurement system and target measurement scenario are given. In Sect. 3, we propose several channel data processing methods and derive the second-order statistics of channel SSF. The time-variant properties of second-order statistics based on the measurement data are analyzed in Sect. 4. Section 5 draws the conclusions.

## 2 Measurements Setup

### 2.1 Measurement System

The developed channel measurement system consists of two parts, i.e., transmitter (TX) part and receiver (RX) part, as shown in Fig. 1. In the TX part, several components are mounted at the bottom of a hexa-copter UAV including a customized TX software-defined-radio (SDR) module with light weight and small volume, a high-power amplifier, an omnidirectional TX antenna, and a L1-band globe positioning system (GPS) module with GPS antenna. The Zadoff-Chu (ZC) sequence with good autocorrelation property is generated in the TX SDR module as the sounding signal. In the RX part, it consists of an adjustable antenna array, a group of low-noise amplifiers and NI SDR modules, a L1-band

GPS module with GPS antenna, and a high-rate disk array. The antenna array includes four omnidirectional antennas in the azimuth plane and four in the elevation plane, respectively. The high-rate disk array is used to store the raw channel data for subsequent data processing. Both the SDR modules in transceiver support maximum frequency band 6 GHz and maximum bandwidth 100 MHz. It should be mentioned that the GPS modules are not only used to record the locations of transceiver but also used to generate the triggering signal for time synchronization of transceiver.




(a) (b)

**Fig. 1.** (a) TX part and (b) RX part of A2G measurement system.

### 2.2 Measurement Scenario

As shown in Fig. 2, the measurement campaigns are carried out at 3.5 GHz in a near-urban scenario at the Jiangjun Road campus, Nanjing University of Aeronautics and Astronautics, Nanjing, China. In this measurement scenario, the average height of the buildings and trees are about 20 m and 5 m, respectively. The RX is put on the toproof of a building with height of 25 m, which is a typical height of base station. The UAV flies forward along a straight line with the velocity of 2 m/s at the height of 60 m, from the starting point to the end point as shown in Fig. 2. During the measurement, the line-of-sight (LoS) path always exists. The non-LoS (NLoS) propagation paths could also occur caused by the ground reflection or the scattering of surrounding buildings and trees.

## 3 Data Processing and Channel Statistical Properties

### 3.1 Channel Data Pre-processing

After obtaining the raw channel measurement data, the sliding correlation operation can be used to extract time-variant CIRs [7]. However, the original CIRs involves the effect of system response caused by the hardware modules, i.e., filters, amplifiers, and so on. Therefore, we obtain the system response by performing a back-to-back measurement before the real measurement, where the system response is denoted as $G(t)$ in frequency domain. Assuming the transmitted sounding signal and the received signal are denoted as $s(t)$ and $y(t)$ ,

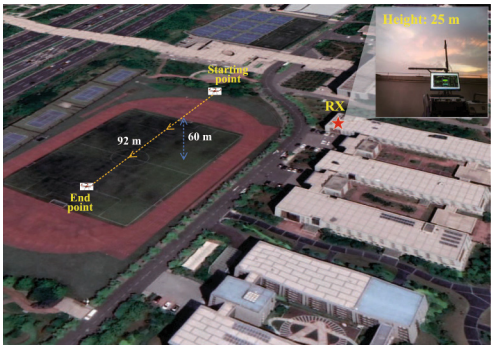



**Fig. 2.** Measurement scenario.

respectively, the CIR after SRE can be expressed as

$$h_{SRE}(t) = \text{IFFT}\left(\frac{Y(t) \cdot \hat{S}^*(t)}{G(t)}\right) \tag{1}$$

where IFFT $(\cdot)$ is the inverse Fourier transform, $Y(t)$ is the Fourier transform of $y(t)$. and $\hat{S}^*(t)$ is the Fourier transform of $\hat{s}^*(t)$, where $\hat{s}^*(t)$ is the conjugate reverse sequence of $s(t)$.

Moreover, we find that there are STOs between the transceiver due to the deviation of crystal oscillator, which would bring maximum cumulative error of 6 dB to the sliding correlation peaks (SCPs). In [9], a signal recovery method is proposed to compensate the error of SCPs caused by the STO. However, the dynamic range of measured CIR decreases a lot by using this method, where the measured CIR could lose some resolvable channel multipath. Therefore, we propose a modified STOC method as

$$h_{STOC}(t) = \int_{t-W_r}^{t+W_r} h_{SRE}(t) \cdot \text{sinc}\left(\frac{k}{I}\right) dk \tag{2}$$

where $W_r$ is the window size, and $I$ is the oversampling ratio of sinc filter.

After obtaining the CIR, another important thing is recognize the valid MPCs from the complete CIR. The constant noise threshold is normally used to recognize the MPCs but it could bring some fake paths or remove some valid paths. In this paper, the constant false alarm rate method with dynamic threshold is applied for MPCR, where the dynamic threshold can be expressed as

$$Thr(t) = \frac{\xi}{W_d - W_g}\left(\int_{t-\frac{W_d}{2}}^{t-\frac{W_g}{2}} h_{STOC}(w)\, dw + \int_{t+\frac{W_g}{2}}^{t+\frac{W_d}{2}} h_{STOC}(w)\, dw\right) \tag{3}$$

where $W_d$ and $W_g$ are the width of detection window and guard window, respectively, and $\xi = (W_d - W_g) \cdot \left(p^{-1/(W_d - W_g)} - 1\right)$ is the threshold factor, where $p$ is the probability of false alarm. Then the parameters of CIR peaks larger than the threshold are taken as the values of valid paths such as path amplitude $\alpha_n(t)$.

### 3.2 Second-Order Statistical Properties of Channel SSF

Second-order statistical properties of channel fading are common parameters for describing the statistical properties of fading channels and validating the performance of channel models [14,15], which are also important for the channel coding, channel estimation, interleave scheme in the communication system. Considering the channel measurement data in the channel sounder is discrete, we derive several second-order statistics of channel SSF in a discrete way in this section, i.e., ACF, CDF, and LCR. The channel fading can be seen as the superposition of large-scale fading (LSF) and SSF. Therefore, the discrete SSF can be expressed by removing the LSF from the envelope of received signal as

$$\gamma[m] = \frac{|y[m]|}{\sum_{n=1}^{N[\lceil m/l_s \rceil]} \sqrt{|\alpha_n[\lceil m/l_s \rceil]|^2}} \tag{4}$$

where $m$ is the diecrete time index, $y[m]$ is the discrete received signal, $\lceil \cdot \rceil$ is the floor function, and $l_s$ is the length of sounding sequence. $N[\lceil m/l_s \rceil]$ is the valid path number in $\lceil m/l_s \rceil$ th CIR, $\alpha_n[\lceil m/l_s \rceil]$ is the path amplitude of nth path in $\lceil m/l_s \rceil$ th CIR. ACF is an important second-order statistical property to reflect the similarity change of channel SSF along with time delay $\Delta t$. The ACF of channel SSF can be expressed as

$$R_{\gamma\gamma}[\Delta t; m] = \sum_{k \in [m, m+l_s]} \gamma[k]\gamma[k - \Delta t]. \tag{5}$$

It should be mentioned that the length of SSF signal in (5) should be infinite according to the original definition of ACF. However, due to the length limitation of real measurement data and the non-stationarity of A2G channel, the length of SSF signal is set as the length of sounding signal to perform the autocorrelation operation in this paper. The CDF with respect to specific threshold level $r$ can be expressed as

$$\begin{aligned} F_\gamma[r; m] &= \Pr(\gamma[k] \le r)_{k \in [m, m+K]} \\ &= \frac{\text{card}\{k | \gamma[k] \le r, k \in [m, m+K]\}}{K} \end{aligned} \tag{6}$$

where $K$ is the sampling number of one statistical interval, card $\{\cdot\}$ is the element number of a vector. LCR represents the times that channel SSF positively or negatively passes through a specific level in unit time. The LCR with respect to specific threshold level $r$ can be expressed as

$$L_\gamma[r; m] = \frac{\text{card}\{k | \gamma[k] \le r, \gamma[k] > r, k \in [m, m+K]\}}{KT_s} \tag{7}$$

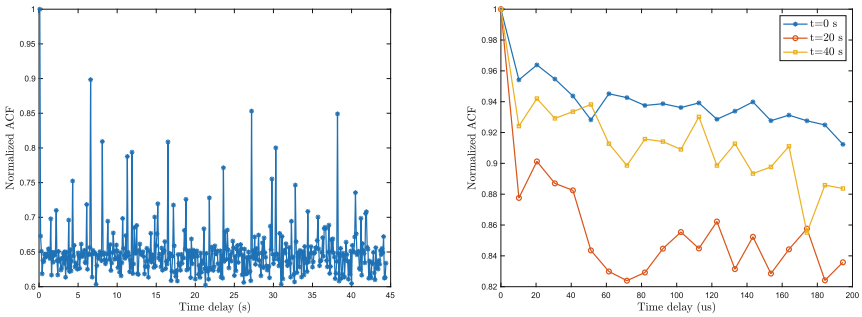
where $T_s$ is the sampling rate.

## 4 Measurement Results Analysis

For the measurement scenario in Sect. 2.2 we continuously record 20480 samples of raw received signals to calculate one slice of channel SSF. In total 460 slices

are recorded with time interval of 100 ms. The sampling rate of received signals is 100 Msps. Then the time-variant second-order statistics of channel SSF, i.e., ACF, CDF, and LCR are calculated and analyzed in this section.

As shown in Fig. 3(b), we present three ACFs for three slices of channel SSF at different time instants, i.e., $t$= 0 s, 20 s, and 40 s. Each ACF for different slice of channel SSF has different decreasing tendency, which means the ACFs are time variant. However, the ACF only decreases to about 0.8 in each slice duration. Therefore, we further calculate the ACF between the first slice and all the other slices of channel SSF along the whole measurement trajectory. It can be found that the lowest ACF is about 0.6, which means that the channel SSF always keeps high correlation in this measurement case.

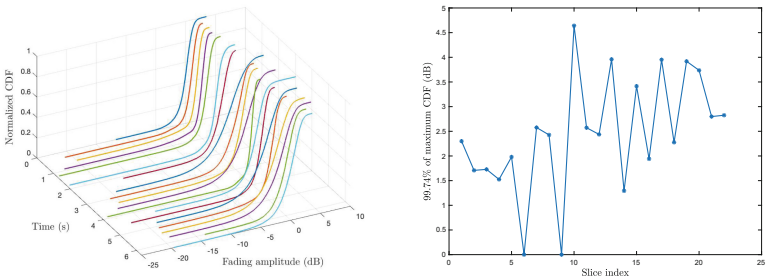



(a) (b)

**Fig. 3.** (a) ACF along the whole measurement trajectory and (b) time-variant measured ACFs within three different slices.

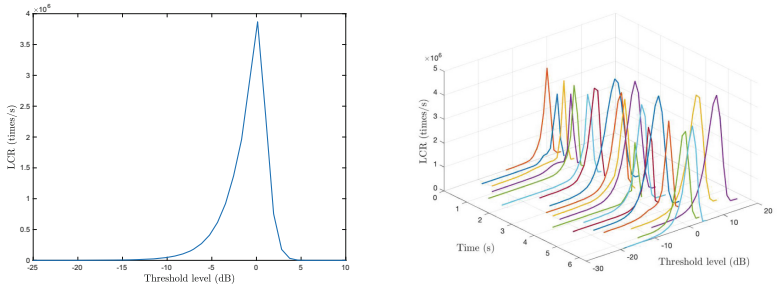
Similarly, it can be found from Fig. 4(a) that the CDFs are also time-variant. To clearly show the results, we only plot 20 slices with time interval of 0.3 s. As the CDF reflects the distribution of fading amplitude, 99.74% of maximum CDF value (denoted as 99.74% CDF) is an important parameter for calculating the fading depth. The fading amplitudes corresponding to each 99.74% CDF are shown in Fig. 4(b), which are time-variant as well and mainly distribute between 0–4.7 dB in this measurement scenario.

LCR is another second-order statistic relative to fading amplitude, which reflects the times that fading envelope positively or negatively pass through a specific threshold level in unit time. As shown in Fig. 5(b), the LCRs are also time-variant, but most of the threshold levels corresponding to the maximum values of LCR for each slice of channel SSF distribute between 0–3 dB, which does not vary too much from the 0 dB of the whole measurement trajectory. And the crossing times of different slice of LCR distribute between $3 \cdot 10^6$–$4.8 \cdot 10^6$, which is also close to $4 \cdot 10^6$ for the whole measurement. Therefore, the LCRs do not vary too much along the whole measurement, which means this measurement scenario is not highly non-stationary.




(a) (b)

**Fig. 4.** (a) Time-variant measured CDFs and (b) fading amplitude corresponding to 99.74% CDF.




(a) (b)

**Fig. 5.** (a) LCR for the whole measurement trajectory and (b) time-variant measured LCRs for different slices.

## 5 Conclusion

This paper has developed a UAV-aided channel sounder, where several optimized channel data processing methods, i.e., SRE, STOC, and MPCR have been proposed. Moreover, several second-order statistics of channel SSF have been derived and analyzed based on the measurement data. The results have shown that the second-order statistics of A2G channel are time-variant. The channel SSF can keep above 60% similarity in the whole measurement, the fading amplitudes corresponding to each 99.74% CDF distribute between 0–4.7 dB, and most of the threshold level corresponding to the maximum value of each LCR distribute between 0–3 dB. These results would be a good reference to the channel coding, channel estimation, interleave scheme of A2G communication system. In the future, we plan to develop more efficient real-time channel data processing methods to increase the sampling interval of CIRs.

**Acknowledgments.** This work was supported in part by the National Natural Science Foundation of China (No. 62271250), in part by China Scholarship Council, and

in part by Postgraduate Research & Practice Innovation Program of Jiangsu Province, No. KYCX22_0360.

## References

1. You, X., Wang, C.-X., Huang, J., Gao, X., Zhang, Z., et al.: Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts. Sci. China Inf. Sci. **64**(1), 1–74 (2021)
2. Zhu, Q., Bai, F., Pang, M., Li, J., Zhong, W., Chen, X., et al.: Geometry-based stochastic line-of-sight probability model for A2G channels under urban scenarios. IEEE Trans. Antennas Propag. **70**(7), 5784–5794 (2022)
3. Mao, K., Zhu, Q., Song, M., Li, H., Ning, B., et al.: Machine learning-based 3D channel modeling for U2V mmWave communications. IEEE Int. Things J. **9**(18), 17592–17607 (2022)
4. Lyu, Y., Kyösti, P., Fan, W.: Sub-THz VNA-based channel sounder structure and channel measurements at 100 and 300 GHz. In: Proceedings of PIMRC 2021, Helsinki, Finlan, pp. 1–5 (Sept 2021)
5. Khawaja, W., Ozdemir, O., Erden, F., Guvenc, I., Matolak, D.W.: Ultra-wideband air-to-ground propagation channel characterization in an open area. IEEE Trans. Aerosp. Electron. Syst. **56**(6), 4533–4555 (2020)
6. Rodríguez-Piñeiro, J., Domínguez-Bolaño, T., Cai, X., et al.: Air-to-ground channel characterization for low-height UAVs in realistic network deployments. IEEE Trans. Antennas Propag. **69**(2), 992–1006 (2021)
7. Ning, B., Li, T., Mao, K., et al.: A UAV-aided channel sounder for air-to-ground channel measurements. Physical Commun. **47**, 101366 (2021)
8. Geise, R., Weiss, A., Neubauer, B.: Modulating features of field measurements with a UAV at millimeter wave frequencies. In: Proceedings of CAMA 2018 Västerås, Sweden, pp. 1–4 (Sept 2018)
9. Burmeister, F., Jacob, R., Traßl, A., Schwarzenberg, N., Fettweis, G.: Dealing with fractional sampling time offsets for unsynchronized mobile channel measurements. IEEE Wirel. Commun. Lett. **10**(12), 2781–2785 (2021)
10. Meng, Y.S., Lee, Y.H.: Measurements and characterizations of air-to-ground channel over sea surface at C-band with low airborne altitudes. IEEE Trans. Veh. Technol. **60**(4), 1943–1948 (2011)
11. Cui, Z., Briso-Rodríguez, C., Guan, K., Güvenç, I., Zhong, Z.: Wideband air-to-ground channel characterization for multiple propagation environments. IEEE Antennas Wireless Propag. Lett. **19**(9), 1634–1638 (2020)
12. Simunek, M., Fontan, F.P., Pechac, P., Otero, F.J.D.: Space diversity gain in urban area low elevation links for surveillance applications. IEEE Trans. Antennas Propag. **61**(12), 6255–6260 (2013)
13. Cui, Z., Briso-Rodríguez, C., Guan, K., Zhong, Z., Quitin, F.: Multi-frequency air-to-ground channel measurements and Analysis for UAV communication systems. IEEE Access. **8**, 110565–110574 (2020)
14. Zhu, Q., et al.: Map-based channel modeling and generation for U2V mmwave communication. IEEE Trans. Veh. Technol. **71**(8), 8004–8015 (2022)
15. Chang, H., Wang, C.X., Liu, Y., et al.: A novel nonstationary 6G UAV-to-ground wireless channel model with 3-D arbitrary trajectory changes. IEEE Internet Things J. **8**(12), 9865–9877 (2021)

# Analysis Method of Flow Density Based on YOLOv4 Multi-feature Fusion

Youli Zhang, Zifei Yu(✉), Lin Wang, and Tianyi Gao

Beihai Communication Technology Company, Tianjin 300000, China
13652016083@163.com

**Abstract.** In order to solve the problem of low precision in calculating the pedestrian flow density of train carriages under complex background and mutual occlusion scenes, a method based on YOLOv4 multi-feature fusion model is proposed to estimate the pedestrian flow density. Firstly, the data set is labeled with head and body respectively to generate the head set and body set. Then use these two data sets to train YOLOv4 multi-feature fusion models YOLO-body and YOLO-head respectively. Finally, the two models are used to reason on the same test data set, and their output results are fused with maximum values. The results show that the method based on YOLOv4 multi-feature fusion model improves the accuracy by 8% compared with the original object detection method and the density map regression method while maintaining better robustness, and can better adapt to the scene of multi-occlusions carriage flow density detection.

**Keywords:** Flow density analysis · Yolov4 · multi-feature fusion

## 1 Introduction

With the development of urban transportation and the in-depth implementation of China's carbon emission reduction policy, subway travel has become the choice of more and more travelers with its advantages of low carbon, environmental protection, convenience and quickness. As important places for urban residents to travel and distribute, subways are highly enclosed and will be impacted by large passenger flows in different degrees, which may lead to chaos, crowding, pushing and other situations in the station, with certain potential safety hazards [1–3].

In view of the problems of low accuracy and poor stability of flow density estimation caused by factors such as high passenger flow density, uneven distribution, serious occlusion and fast flow velocity in the subway area, scholars at home and abroad have made certain research results. Hong et al. [4] proposed a method for estimating the flow density using IOT. By collecting the image population information at different times, and using the population image calculation compensation matching technology, they compensated for the difference in the number of people in the front and back frames, making up for the shortcomings of traditional methods, and achieved pretty good results. Ma et al. [5] introduced the mixed Gaussian model for modeling, described the human body contour with the help of the foreground edge detection of the image, and tracked


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 288–294, 2023.
https://doi.org/10.1007/978-981-99-2362-5_36

and extracted the features of the human body contour based on the Canny operator. Based on these contours, they realized efficient parallel distributed computing of flow density, with the missing detection rate reduced from 13.21% to 5.24%, and the prediction accuracy of flow reached 92.43%. CJ jin et al. [6] proposed a top view video training dataset for some pedestrian flow experiments based on UAV records. YOLOv3 was used to train the detection model and the detection results were good. When the pedestrian density was up to 9.0 ped/m2, its accuracy, recall and FI score might also be greater than 0.95. Haroon Idrees et al. [7] put forward a locally consistent form of scale prior, which captures the scale similarity in the local neighborhood and its smooth change on the image. According to the detection scale and confidence level obtained from potential human detectors, the scale and confidence prior can be inferred in Markov random field. This method uses local neighborhood correlation constraints to solve human detection and occlusion reasoning, so as to respect the interdependence characteristics between individuals, and conduct flow density analysis. The experimental results show that the accuracy of this method in flow density analysis has been improved significantly.

This paper creatively proposes a multi-feature fusion flow density image algorithm based on YOLOv4, which can well adapt to the natural application scenarios with large passenger flow density, uneven distribution and serious occlusion overlap. In the relevant subway scene, it has better effect than the single feature model, and can stably analyze the flow density of the real-time images obtained by the camera in the subway for a long time under the condition of ensuring high accuracy.

### 1.1 YOLOv4 Model Principle

The basic idea of YOLOv4 algorithm can be divided into two parts: firstly, generate a series of candidate regions on the image according to certain rules, and then mark the candidate regions according to the position relationship between these candidate regions and the real region of the object on the image. The part of the candidate area whose distance from the real box is less than the threshold value will be marked as a positive sample. At the same time, the position coordinate of the real box is taken as the target value of the position coordinate of the positive sample. The candidate areas with a large distance from the real box will be marked as negative samples, and the negative samples do not need to predict the location coordinates or category information [8–11].

Secondly, the convolutional neural network is used to extract the features of the image and predict the location coordinates and category information of the candidate regions. In this way, each predicted box can be considered as a sample and the labelled values can be obtained based on the location coordinates and category information of the real box relative to it. A network model is used to predict its position and category, and the network predictions are compared with the labelled values. This allows a loss function to be constructed for training. The thought of YOLOv4 algorithm is shown in Fig. 1.

The backbone network adopted by YOLOv4 is DarkNet-53. The DarkNet-53 network structure does not have a pooling layer. During forward propagation, the pooling layer is replaced by changing the step size of the convolutional kernel, and the feature extraction model uses many $3 \times 3$ and $1 \times 1$ convolutional layers, plus a total of 53 fully connected layers. After DarkNet-53 feature extraction, in order to improve the detection

accuracy of objects of different sizes, YOLOv4 detects objects on three different scales, each of which has three bounding boxes. Finally the target is predicted by the bounding value with the largest IOU to the real box. The structure of YOLOv4 is shown in Fig. 2. Resn in the figure represents a residual block, which contains n residual units; DBL is the basic component of YOLOv4, representing convolution + batch normalization (BN) + activation function leak ReLU.

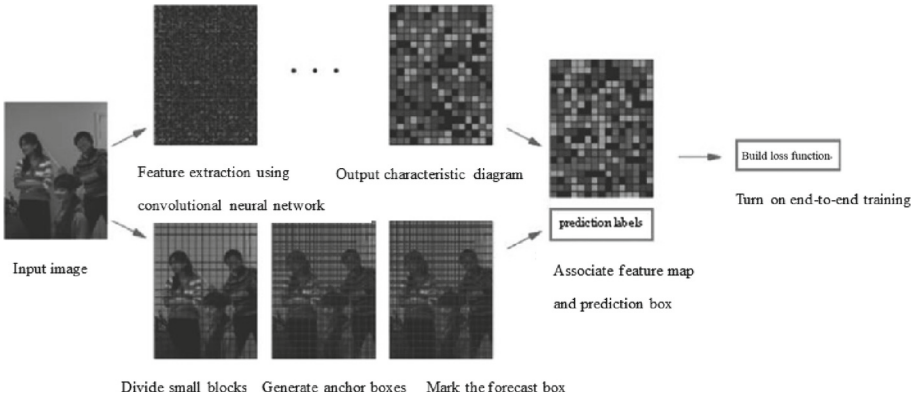



**Fig. 1.** The thought of YOLOv4 algorithm

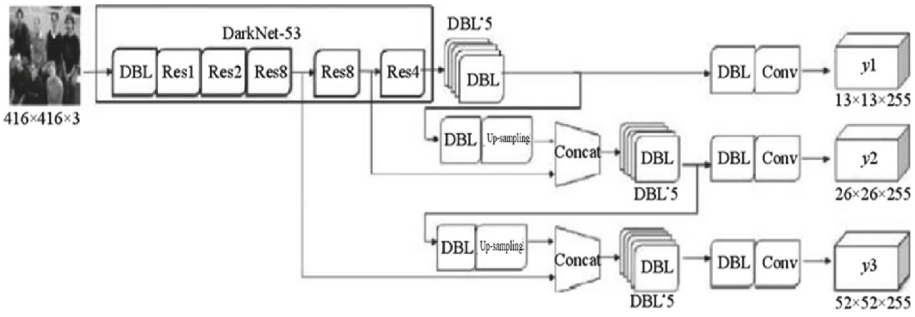



**Fig. 2.** YOLOv4 network structure diagram

## 2 Multi-feature Fusion Model Based on YOLOv4

In order to maximize the detection accuracy while maintaining the inference speed, the YOLOv4 multi-feature fusion model is improved on the basis of the original network as follows:

(1) The backbone network uses Res Net50-vd to replace the original Dark Net-53. ResNet-vd is a modified version of the ResNet family, which has almost the same number of parameters and computational effort as ResNet, but with a 2% increase

in accuracy. Although DarkNet-53 also uses a residual network as shown in Fig. 3, the ResNet50-vd has a speed and accuracy advantage over the DarkNet-53, and is more easily scalable with the ResNet family of networks. ResNet18, 50 and 101 can be flexibly selected as the backbone network for target detection according to different business requirements.

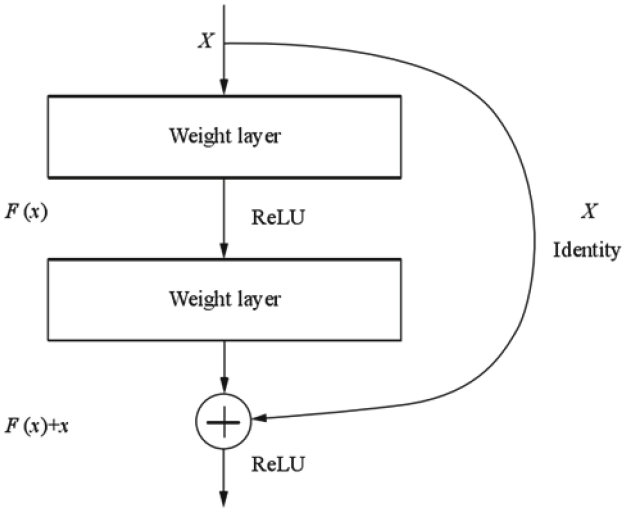



**Fig. 3.** Residual structure

(2) Deformable Convolution (DCN) is introduced as an alternative to the original convolution operation. Deformable Convolution has been extensively validated for vision tasks in various neighbourhoods. Considering the premise of maintaining the balance between speed and improving accuracy, the YOLOv4 augmentation model uses Deformable Convolution to replace the 3 × 3 convolution in the 5th stage of the backbone network.

(3) As YOLOv4 is a single-stage target detection model, it has a natural disadvantage in terms of localization accuracy compared with two-stage target detection models such as Faster RCNN and Cascade RCNN. Therefore, YOLOv4 is enhanced with an IoU loss branch, which can improve the localization accuracy of bounding boxes to a certain extent and narrow the gap between single-stage target detection networks and two-stage detection networks.

The traditional L2 loss is trained on the location coordinate information of the detection target as four variables independent of each other. The IoU loss uses the maximum overlap between the predicted bounding box and the underlying true value, and regresses all the bound variables as a whole, so that the position coordinate information is trained as a whole. Therefore, using IoU loss can obtain more accurate training effects and detection results. The L2 loss and IoU loss are illustrated in Fig. 4.

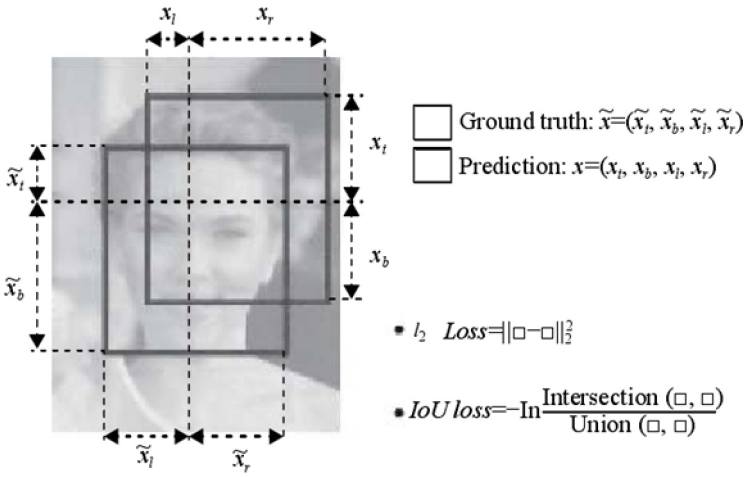


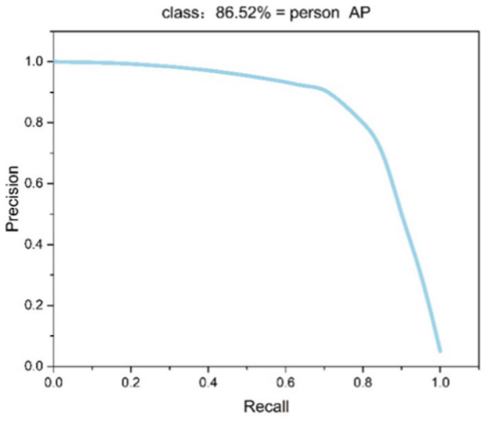

**Fig. 4.** L2 loss and IoU loss

## 3 Results of the Experiment

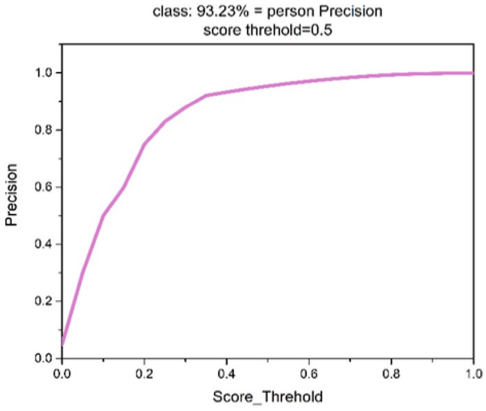
After training, a real-time detection speed of 2FPS was selected to conduct experiments on the YOLOv4 multi-feature fusion algorithm to implement human density detection in NVIDIA Jetson TX2. The variation of AP value, accuracy rate and recall rate when the model reached convergence during training are shown in Fig. 5(a), 5(b) and 5(c) respectively.

In Fig. 5, the YOLOv4 multi-feature fusion method was used to identify the flow density and was able to achieve an AP value (Average Precision) of 86.52% after training, a result that makes the recognition task more satisfactory; As people are often obscured from each other in real-life scenarios, this can cause very serious recognition errors. In this respect, we have improved the features for the dataset by changing the features to be extracted from the entire human torso to a combination of human head and torso features, which reduces to some extent the errors caused by the mutual occlusion of people. Figure 6 (a) and (b) show the experimental results on the NVIDIA Jetson TX2.

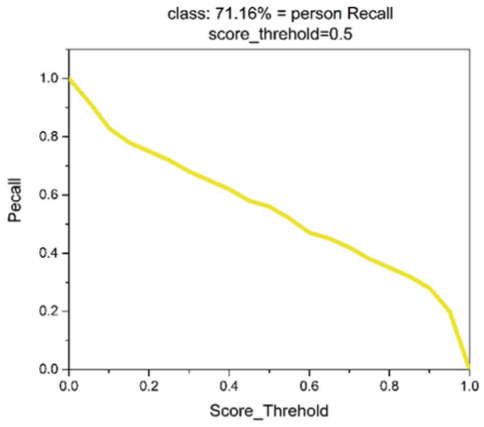
This paper compares the multi-feature fusion model approach, the original single-model detection approach and the Gaussian density map regression approach on the NVIDIA Jetson TX2 with average error rates (AER) of 0.353 for the VGG16 Gaussian density map regression, 0.186 for the MCNN Gaussian density map regression, 0.212 for the YOLO-body and 0.182 for the YOLO- head, and 0.127 for the multi-feature fusion model. The results show that the method based on YOLOv4 multi-feature fusion model has an accuracy improvement of 8% compared with the original target detection method and density map regression method, and has better robustness, which can better adapt to the multi-occluded carriage flow density detection scenario.




**(a)**




**(b)**




**(c)**

**Fig. 5.** Change in training recall rate


(a) (b)

**Fig. 6.** Feature view: (a) Feature view of the head, (b) Feature view of the torso

## References

1. Zheng, S., Wu, J., Liu, F., et al.: An Improved Crowd Counting Method Based on YOLOv3. In: International Conference on Machine Learning and Intelligent Communications, pp. 337-350. Springer, Cham (2022). https://doi.org/10.1007/s11036-022-02082-7
2. Zhang. H., Yang, X., Hu. Z., et al.: High-density pedestrian detection algorithm based on deep information fusion. Appli. Intell. 1–13 (2022)
3. Xue, Y., Ju, Z., Li, Y., et al.: MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection. Infrared Phys. Technol. **118**, 103906 (2021)
4. Hong, Y.H., et al.: Content networking technique in flow density estimation application, **30**(6), 4 (2013)
5. Ma, J.: Design and implementation of Hadoop-based crowd density prediction model for scenic spots **29**(18), 39–43 (2021). https://doi.org/10.14022/j.issn1674-6236.2021.18.009
6. Jin, C.J., Shi, X., Hui, T., et al.: The automatic detection of pedestrians under the high-density conditions by deep learning techniques **2021**(4), 1–11 (2021)
7. Idrees, H., Soomro, K., Shah, M.: Detecting Humans in Dense Crowds Using Locally-Consistent Scale Prior and Global Occlusion Reasoning (2015)
8. Elbishlawi, S., Abdelpakey, M.H., Eltantawy, A., et al.: Deep learning-based crowd scene analysis survey **6**(9), 95 (2020)
9. Shahbazi, M.J.B.D.: Computing C. Advances in convolution neural networks based crowd counting and density estimation. **5** (2021)
10. Jia, Q.U., Shi, Z.L., Yang-Dong, Y.: Unbalanced Crowd Density Estimation Based on Convolutional Features (2018)
11. Khodayari, A., Ghaffari, A., Kazemi, R., et al.: A modified car-following model based on a neural network model of the human driver effects. IEEE Trans. Syst. Man Cybern. Part A: Syst. Humans **42**(6), 1440–1449 (2012)

# A Segmentation Algorithm Based on Shallow Convolutional Neural Networks for Lung X-ray Images

Junjie Hu[1(✉)], Yan Wang[1], Xiaokai Liu[1], Heyu Zheng[1], Yuanmei Zhu[1], and Shiqiang Zhang[1,2]

[1] School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
hujunjie0512@126.com
[2] Hualu Zhida Technology Co., Ltd., Dalian 110623, China

**Abstract.** Considering the problem that the number of parameters of traditional convolutional neural networks is too large and the Chest X-ray (CXR) image dataset with manual labeling is limited, a CXR image segmentation algorithm based on Shallow Convolutional Neural Networks (SCNN) is proposed, including two stages of image preprocessing and segmentation. In the preprocessing stage, the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm is used to enhance the contrast of the image to make the lung boundary clearer. Based on this, the problem of conflicting pixel classification and pixel localization in semantic segmentation is appropriately handled by encoding-decoding network and attention mechanism. Experiments show that the algorithm has the advantages of reducing the number of parameters and high segmentation performance.

**Keywords:** Convolutional neural networks · Chest X-ray · Medical image segmentation

## 1 Introduction

CXR imaging is currently one of the most effective methods for screening pneumonia disease. The low cost and low radiation dose dominate all imaging modalities (computed tomography, MRI scans, *etc.*). In recent years, convolutional neural networks have been increasingly used in medical image segmentation. With the advantages of reducing labor costs and eliminating the need for manually designed features, convolutional neural networks have put traditional methods that rely on the manual production of lung features in a weak position. In the complex convolutional neural network environment, how to use convolutional neural networks for accurate segmentation of lung images is of great significance for diagnosis, treatment and surgical planning of lung diseases.

X. Liu—This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 83118038.


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 295–302, 2023.
https://doi.org/10.1007/978-981-99-2362-5_37

In the field of medical imaging, the encoder-decoder-skipping layer network structure of U-Net [1] has inspired a large number of improved medical image segmentation methods based on the U-Net structure to reduce the number of parameters and to improve the accuracy of detail recognition for network recognition.

To address the problem of saving parameters and computational power, Oktay *et al.* [2] proposed the Attention U-Net network to add the attention mechanism before the encoding and decoding of the network corresponding to feature splicing, which can automatically learn to distinguish the shape and size of the target and reduce the number of parameters. However, the network focuses only on the features on the channel. Zhao *et al.* [3] proposed a SCOAT-Net network with channel and spatial attention mechanisms on UNet++ to improve the network results. Yahyatabar *et al.* [4] proposed a deep CNN model called Dense-UNet, which increases the information flow of the whole network and reduces the network parameters by dense connections between the layers. However, the reuse rate of the Dense module is very low and has many redundant features. With the deepening of the network model, a small number of data sets are prone to overfitting.

To address the problem that features information is easily lost in the network, Xiao *et al.* [5] proposed Res-UNet inspired by ResNet, replacing each sub-module of U-Net with a residual connection in the form of reducing the problem of low detail discriminative ability respectively. Based on this, Jha *et al.* [6] proposed the Res-UNet++ network, which continues to introduce compressed attention mechanisms to make it applicable to a smaller number of images. Others, Azad *et al.* [7] proposed an SMU-Net network to generate mappings by merging low-level and high-level image structures, where low-level (e.g., grayscale and texture) and high-level (e.g., region-based objects) are generated by superpixel clustering and region grouping. Huang *et al.* [8] proposed a U-Net3+ network with deep supervision to maximize the use of full-size feature maps to enhance contextual information for accurate segmentation and reduce the number of parametric. By analyzing the above methods, from two perspectives of multi-scale information fusion and utilization of contextual information, this paper will adopt the network structure of the U-Net encoding-decoding skeleton that retains the advantages of high image resolution and combine with multi-layer fusion perception pyramid structure to enrich feature information. It will achieve accurate and realistic lung segmentation. Meanwhile, the attention mechanism is introduced in the skipping layer structure to enhance the critical feature weights and suppress irrelevant features.

## 2 Improved Algorithm Based on U-Net

### 2.1 Image Pre-processing

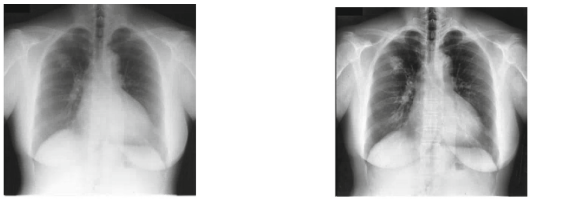
There are many images in the training set with blurred lung boundaries, unclear contrast, and other factors that make feature extraction difficult. To improve these problems, the images must be pre-processed before recognition. Using the

CLAHE algorithm to enhance the images to highlight the double lung portion, and scaling the image size to $256 \times 256$ to ensure that the images have sufficient visual detail of the lung and boundary structures while adapting to the size of the network input.

The main steps to improve the image contrast using the CLAHE algorithm are as follows.

(1) The image is filled in chunks.
(2) For each chunk, the mapping relationship is calculated, and the contrast limits are used to calculate the mapping relationship.
(3) A bilinear interpolation method is used to obtain the final enhanced image.

A comparison of the images after passing the pre-processing is shown in Fig. 1.



(a) original image (b) pre-processing image

**Fig. 1.** Comparison of images before and after the pre-processing

### 2.2 Attention Mechanism Module

In this paper, the attention mechanism is prior to encoding and decoding corresponding feature splicing, mainly the improved attention module and the spatial attention mechanism, which give different degrees of attention at channel locations and spatial locations and give more weight to important image locations. The structure of the attention mechanism is schematically shown in Fig. 2. The training makes the attention coefficient $\alpha \in [0, 1]$, so that the value of the target

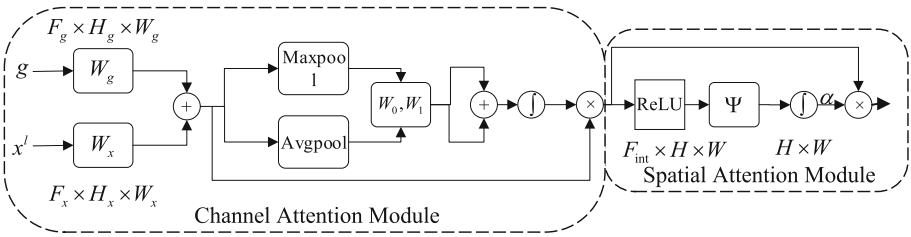



**Fig. 2.** Attention mechanism

region converges to 1 and the irrelevant region converges to 0. $x_{i,c}^{l_c}$ multiplying with $\alpha$, the result of multiplication will put the attention on the target region. The synthesis formula can be expressed as $\hat{x}_{i,c}^{l} = \alpha_i^l \cdot x_{i,c}^{l_c}$. In the Fig. 2, $x_i^{l_c}$ represents the $i$ feature map in the $l$ layer, $W_g$, $W_x$, $W_0$, $W_1$ denotes the $1 \times 1$ convolution operation, $F_g$, $F_l$, $F_{int}$, and $c$ represents the number of channels, $H$ represents the height, $W$ represents the width, and $\sigma$ is the *sigmoid* function.

### 2.3 SCNN-UNet Model

After the image preprocessing is completed, the image feature extraction is using SCNN-UNet for the problem of conflicting pixel classification and localization in semantic segmentation. The SCNN-UNet model is shown in Fig. 3.

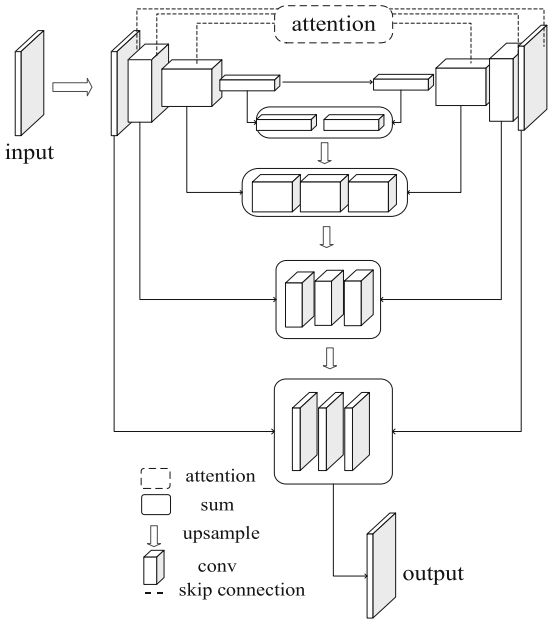



**Fig. 3.** SCNN-UNet model

The network consists of four encoder blocks (four convolutional layers after the input) and four decoder blocks (four deconvolutional layers in a symmetric structure with respect to the encoding network) connected by a skip connection. Each encoder block consists of two consecutive $3 \times 3$ convolutional layers and a maximum pooling layer. Each convolutional layer is followed by a BN layer and a ReLU layer.

SCNN-UNet obtains the image features $x_{i,c}^{l}$ through the encoding network and the feature map $g_i$ with pixel-level classification information through the decoding network, where $g_i$ denotes the decoding network layer feature map

corresponding to the $l$ layer. The attention mechanism is introduced in the skipping structure to make the network focus on the more prominent feature regions of the target category and effectively sense the lung location in CXR images. And upsampling feature extraction of the decoding network can be more effectively performed. Thus improving the final lung region segmentation effect. The encoding network feature map $x_{i,c}^{l}$ is passed through the attention mechanism to obtain the feature map $\hat{x}_{i,c}^{l}$. It is then stitched with the feature map of the decoded network to maintain the resolution of the image and to perform activation, as represented by Eq. 1.

$$g_{i,c}^{n-l+1} = \sigma(C(\hat{x}_{i,c}^{l}, g_{i,c}^{n-l+1})), \tag{1}$$

where $C$ denotes the splicing operation and $\sigma$ denotes *sigmoid* activation. Second, a modified multiscale fusion pyramid structure is incorporated under the above conditions, where the output of each layer uses bilinear interpolation so that the feature map recovers features of the same size as the previous layer, as expressed in Eq. 2.

$$x_{i,c}^{l_1} = U(x^{l_1-1}), \tag{2}$$

where $U$ denotes bilinear interpolation to get richer texture and semantic information.

Finally, feature maps with the same scale in the overall network structure are connected in the channel dimension and downscaled using pooling operations for feature fusion, as shown in Eq. 3.

$$x_{i,c}^{l_1} = f(x_{i,c}^{l} + g_{i,c}^{n-l+1} + x_{i,c}^{l_1}), \tag{3}$$

where $f$ is the $1 \times 1$ convolution, $x_{i,c}^{l-1}$ denotes the feature map after feature mapping of the layer $l-1$ coding network, and the specific structure is shown in Fig. 4. In multi-scale feature fusion, point-by-point summation and feature mapping are used to obtain richer texture and feature information with guaranteed resolution.

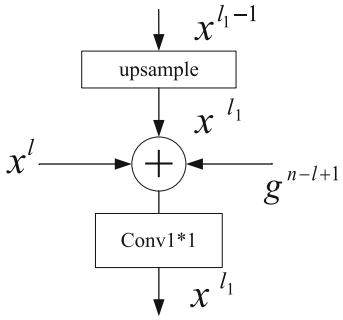



**Fig. 4.** The Structure of Mapping network

**Table 1.** Comparison of accuracy rates before and after pre-processing.

| **Methods** | Mean MIoU (%) | Mean Dice (%) |
|---|---|---|
| SCNN-UNet | 82.11 | 87.01 |
| SCNN-UNet (CLAHE) | 84.65 | 89.44 |

## 3 Experiment and Analysis

### 3.1 Data Set and Experimental Environment

The dataset used in this study is a mixed Montgomery and JSRT dataset, which is later expanded by using the CLAHE algorithm to enhance the images. The training set contains 600 images, and the validation set includes 170 images. The experiments in this paper were conducted under graphics card model TU106 [GeForce RTX 2070], Linux operating system, and use stochastic gradient descent method with momentum set to 0.9 and learning rate is initialized to 0.001.

### 3.2 Results and Analysis

Table 1 shows the comparison of the accuracy of the original image and the image after using the CLAHE algorithm. From Table 1, with the same network structure, increasing the image contrast can improve the accuracy of the network.

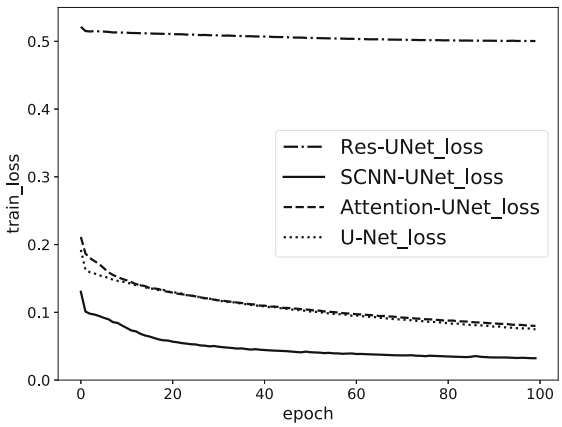



**Fig. 5.** Plot of the number of iteration steps versus loss for each network

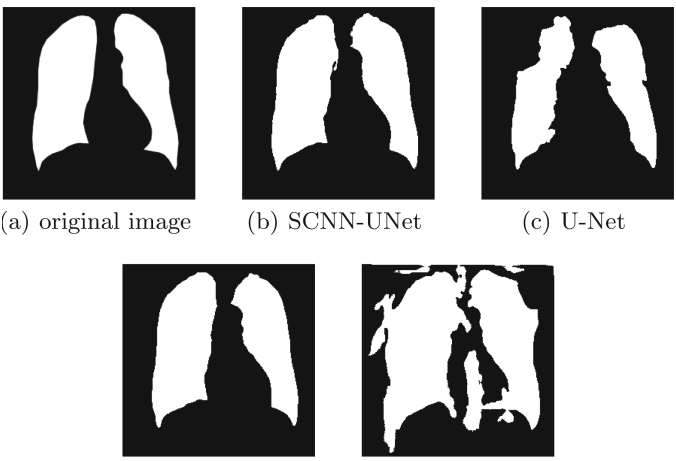
Figure 5 shows the loss function plots for each network, which Res-UNet++ and Dense-UNet with less accurate are not compared here. After image preprocessing, the enhanced images are fed into the network which involved in Table 2 for training and verifying. This experiment uses the sum of two loss

**Table 2.** Comparison of performance of different networks after CLAHE enhancement.

| Methods | Mean MIoU (%) | Mean Dice (%) | Parameters (M) |
|---|---|---|---|
| U-Net (5-layer) | 59.35 | 68.21 | 13.40 |
| Attention-UNet (5-layer) | 86.48 | 90.60 | 34.88 |
| Dense-UNet | – | – | 12.27 |
| Res-UNet | 66.82 | 79.86 | 13.04 |
| Res-UNet++ | – | – | 14.48 |
| SCNN-UNet (4-layer) | **84.76** | **89.05** | **8.82** |

functions, Diceloss and Bceloss, to guide the image segmentation, and uses two accuracy metrics, the average MIoU and the average Dice, of all the test images to evaluate the goodness of the network model.

Table 2 shows the performance comparison of different networks after enhancement. Table 2 shows that the MMIoU metrics of SCNN-UNet can reach about 85%, and MDice metrics can reach about 89%. Compared with U-Net, adding the attention mechanism SCNN-UNet has higher accuracy and proves the effectiveness of the attention mechanism. Compared with the Attention-UNet 5-layer network, the number of parameters required for the SCNN-UNet network, which is a 4-layer convolution with the highest channel count of 512, can be reduced to 25% of the Attention-UNet network, which is a 5-layer convolution with the highest channel count of 1024. SCNN-UNet network has a minimum number of parameters (8.82M). Figure 6 shows the results of each network.



(a) original image (b) SCNN-UNet (c) U-Net

(d) Attention-UNet (e) Res-UNet

**Fig. 6.** Experimental results for each network

It shows that SCNN-UNet and Attention-UNet are comparable, and SCNN-UNet is slightly better in detail processing; Res-UNet and U-Net are not good enough in enriching contextual information and detail processing. In summary, this paper proposes a shallow convolutional neural network-based CXR image segmentation algorithm, which has better performance in reducing the number of parameters and improving the accuracy rate.

## 4 Conclusions

In this paper, we propose a segmentation algorithm for lung X-ray images based on shallow convolutional neural networks. The method preprocesses the images and then scales them to an appropriate size to send them to the SCNN network model for recognition segmentation. After comparing and analyzing other algorithms, this algorithm has the advantages of reducing the number of parameters and high segmentation performance.

## References

1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
2. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv. 1804.03999 (2018)
3. Zhao, S., et al.: SCOAT-Net: a novel network for segmenting COVID-19 lung opacification from CT images. J. Pattern Recogn. **119**, 108109 (2021)
4. Yahyatabar, M., Jouvet, P., Cheriet, F.: Dense-Unet: a light model for lung fields segmentation in chest X-Ray images. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society. MBC, pp. 1242–1245. IEEE (2020)
5. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted Res-UNet for high-quality retina vessel segmentation. In: 2018 9th International Conference on Information Technology in Medicine and Education. ITME, pp. 327–331. IEEE (2018)
6. Jha, D., et al.: A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. IEEE J. Biomed. Health Inform. **25**, 2029–2040 (2021)
7. Azad, R., Khosravi, N., Merhof, D.: SMU-Net: style matching U-Net for brain tumor segmentation with missing modalities. arXiv preprint arXiv. 2204.02961 (2022)
8. Huang, H., et al.: UNet 3+: a full-scale connected UNet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1055–1059. IEEE (2020)

# Target Intention Reasoning Algorithm Based on Multi-agent Network and Adaptive Genetic Bee Colony

Huiya Zhao [1] and Yaping Wang[2(✉)]

[1] Beihang University, Beijing 100191, China
[2] Nanjing University of Science and Technology, Nanjing, China
19422876@qq.com

**Abstract.** Aiming at the problem that the analysis of target intention is difficult to apply to real environment and the reasoning efficiency is low, this paper constructs a target intention reasoning algorithm based on multi-agent network, which represents the target information and the relationship between them. Firstly, the problem of target intention inference is transformed into the problem of optimal inference path search in complex networks. Secondly, the idea of genetic mutation crossover is introduced into the artificial bee colony algorithm to improve the food source adaptation strategy and improve the solving ability and convergence speed of the global optimal solution. Finally, the experimental results show that the proposed method is significantly improved in reasoning accuracy and average time consumption compared with genetic algorithm and artificial bee colony algorithm.

**Keywords:** Intentional reasoning · multi-agent network · genetic reasoning algorithm · genetic bee colony

## 1 Introduction

Traditional goal intention reasoning algorithm [1] generally compares the situation information obtained from situation awareness with the data in the information base to find the closest situation pattern and get the reasoning result. When the surrounding environment is relatively simple, the number of targets is small, and the behavior is relatively simple, this method can achieve good reasoning effect. However, the real test environment is extremely complicated, the real-time information of situation changes rapidly, the test actions are abundant, the relationship between them is complicated, and the knowledge base and rule base needed for realizing intention reasoning are also very large. At this time, using traditional methods will make the reasoning efficiency low, and it is easy to get wrong reasoning results.

The multi-agent theory has been applied to solve the problem of intention reasoning for the target [2]. According to the relevant theories of multi-agent theory, the multi-agent intention reasoning network of the target is firstly established, and the key parameters in the network are explained, so that the complex problem of objective intention reasoning


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 303–312, 2023.
https://doi.org/10.1007/978-981-99-2362-5_38

is transformed into a complex network optimization problem. The basic enumeration method is not suitable for complex network solution, and other traditional algorithms have some shortcomings such as slow speed and great chance to fall into an endless loop when solving the best path search of multi-level, multi-relationship and multi-loop. Aiming at the above problems, this paper proposes an adaptive genetic bee colony algorithm based on artificial bee colony intelligent algorithm, combined with genetic algorithm (GA) and adaptive strategy, and applied to multi-agent intention reasoning of target intention.

## 2 Construction of Multi-agent Reasoning Network

Multi-agent network is established through knowledge base, which contains situation information and rules. In practical application, the inference network is activated by the input environmental information, and the inference calculation of intention is carried out. The multi-agent intention reasoning network can be established through the following steps:

(1) Determination of network nodes.

Each node in the network represents an entity goal or behavior state, which depends on the level and relative importance of the node.

(2) Determination of the directed edge of the network.

The directed edge in the network represents the relationship information between entities and behaviors, which is dynamic rule information.

(3) Determination of node weight and confidence of directed edge reasoning.

According to the above steps and operations, an abstract multi-agent intent reasoning network can be constructed, as shown in Fig. 1.

In this reasoning network, the bottom-up nodes 1–6 are the agent nodes of the attribute feature layer, 7–13 are the agent nodes of the rule reasoning layer (which is divided into two sub-layers including 7–10 and 11–13 respectively), and 14–15 are the agent nodes of the intention result layer. After the mapping of knowledge base, reasoning rule base and multi-agent network structure are realized, so the reasoning calculation of target intention can be carried out according to the above-mentioned network. Eventually, the reasoning problem of cluster target intention with complex collaborative relationship has been successfully transformed into a search problem of optimal reasoning path.

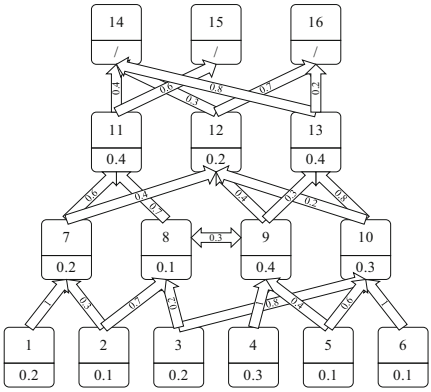



**Fig. 1.** Example of multi-agent intent reasoning network

## 3 Adaptive Genetic Bee Colony Algorithm Based on Multi-agent Network

Artificial bee colony algorithm (ABCA) is an intelligent optimization algorithm first proposed by Karaboga [2] in 2005 to simulate the behavior of bees seeking food sources, inspired by the foraging behavior of bees.This method has some disadvantages such as slow convergence speed, precocious puberty, poor local search ability and small diversity of bee colony. Therefore, the idea of genetic mutation crossover in genitic algorithm can be introduced, and the improved adaptive strategy can be combined to make full use of the advantages of the two which simutaneously complements the disadvantages, and has the ability of fast and accurate global optimization.

In the ABCA, the new food source is discovered by hiring bees or following bees to do neighborhood search near the original food source. It is unknown whether the profitability of the new food source is better than that of the original food source. If the profitability of the new food source is lower, the bees will stay at the original food source until they find a new food source with higher profitability, which makes the swarm more possibly stay generations. In order to improve the development ability and convergence speed of the algorithm, crossover and mutation operations in GA can be introduced.

In this algorithm, the GA and bee colony algorithm are combined, and the adaptive selection strategy of food source is introduced into the genetic bee colony algorithm. After the integration of the GA and the bee colony algorithm, the development ability of the algorithm is enhanced. At this time, the adaptive selection strategy of food sources and the adaptive factor $\omega$ is introduced into the genetic bee colony algorithm, so that the algorithm can achieve a balance between exploration and development. The adaptive factor $\omega$ is related to the number of iterations of the algorithm, and its expression is as follows:

$$rand(0, 1) < \omega = \frac{1}{iter\_\max} * iter \tag{1}$$

where *iter* is the current iteration number and *iter*_ max is the maximum iteration number. In the crossover operation, the food source that follows the bee and the food source

with global or local optimal solution are used as two parents for crossover operation. When the number of iterations satisfies the condition of Eq. (1), the food source and the global optimal solution are selected as parents in the crossover operation. If Eq. (1) is not satisfied, the food source and the local optimal solution are selected as parents. According to the above operation, the parent of the optimal solution of the algorithm can be transformed from local optimal to global optimal with the increase of iteration times, which can not only prevent the algorithm from falling into local optimal at the initial stage, but also enable the algorithm to quickly search for the global optimal solution at the later stage, accelerating the convergence speed and improving the overall performance of the algorithm.

## 4 Algorithm Implementation

The specific steps of the adaptive genetic bee colony reasoning algorithm are as follows:

Step1: Initialize the bee population, with the number of hired bees and follower bees equal to $SN$, search threshold as $limit$, initial iteration number as $iter = 0$, maximum iteration number as $iter\_\max$, crossover probability $P_c$, mutation probability as $P_m$ and adaptive factor as $\omega = {}^{iter}/_{iter\_\max}$;

Step2: Randomly generate $SN$ initial food sources $X_i$, that is, randomly generate $SN$ initial reasoning paths, and calculate the profitability of each food source;

Step3: Hire bees to search for new food sources $V_i$ near food sources and calculate their profitability;

Step4: Compare the profitability $V_i$ of new and old food sources, replace $X_i$ if the profitability is higher, otherwise keep $X_i$;

Step5, Calculate the probability $P_i$ of following bees to choose food sources;

Step6: Follow the bees to choose food sources according to the probability $P_i$;

Step7: Follow the bees to search for new food sources: if $P_m < \varphi_1 < P_c$ cross-operate; If $\varphi_1 < P_m$, perform mutation operation. When crossover operation is performed, if $\varphi_2 < \omega$, the global optimal solution is selected as the parent; If $\varphi_2 > \omega$, select the local optimal solution as the parent. Where, $\varphi_1$ and $\varphi_2$ are random numbers between $[-1, 1]$. Calculate the profitability $V_i$ of new food sources;

Step8: Same as Step4;

Step9: Judge whether there is food source whose search times exceed the search threshold $limit$. If so, generate new food source according to formula (2) to replace it;

Step10: Judge whether the loop termination condition is met, that is $iter > iter\_\max$, if the condition is met, the current optimal solution, that is, the final reasoning path, is output; otherwise, return to Step3, and set $iter$=$iter + 1$.

## 5 Simulation Analysis

Assuming that the UAV test consists of a formation of three UAVs, the intention reasoning analysis of the formation target is carried out. Firstly, according to prior knowledge and tactical rule base, a multi-agent intention reasoning network is established: the agent nodes of attribute feature level in the network structure represent the situation awareness information of three UAVs; The agent nodes in the rule reasoning level are

abstract representations of various states of the formation target obtained by analysis and fusion of perceived information. And the reasoning, causality and other relationships among different states are represented by directed edges between agent nodes, in which complex cooperative relationships among different agent nodes in the same layer can be represented by lateral directed edges. The agent node of the intention level represents the final reasoning result.

In order to verify the efficiency of AGBCA algorithm in solving the above-mentioned model, GA, ABCA and the proposed algorithm are used as comparisons for simulation experiments. Among them, in ABCA and AGBCA, the corresponding relationship between honeybee colony behavior and multi-agent intent reasoning network elements is shown in Table 1.

**Table 1.** The corresponding relationship between honey collecting behavior of bee colony and network elements of multi-agent intention reasoning

| Bee honey collecting behavior | Network elements of multi-agent intention reasoning |
|---|---|
| Food source location | Bottom-up reasoning path, that is, a complete reasoning path to get the formation intention result by reasoning from the attribute feature information of three aircraft targets |
| Food source profitability | Comprehensive trust degree of a reasoning path |
| Speed of finding food sources | The time required to obtain the intended result according to a certain reasoning path |
| High-yield food source | The optimal reasoning path, that is, the reasoning path with the greatest comprehensive trust |

Based on the above analysis, a multi-agent reasoning network is constructed as shown in Fig. 2, and the network is searched for the optimal path.

As shown in Fig. 2, the number below each node indicates the weight of the node in its hierarchy (sub-hierarchy), and the number on the directed edge indicates the confidence of the reasoning rule. Among them, the bottom-up nodes 1–3 are the agent node of the attribute feature layer, which represents the situation awareness information of three aircraft, 4–24 are the agent nodes of the rule reasoning layer (divided into several sub-layers), and 25–30 are the agent nodes of the intention result layer, which represents the final intention reasoning result of the formation.

For GA, the experiment sets population size of 50, iteration number of 30, crossover probability of 0.8 and mutation probability of 0.05. For ABCA and adaptive genetic bee colony algorithm AGBCA, the experiment sets the bee population size of 30, including 15 for hiring bees and 15 for following bees, the maximum iteration number $iter\_max$ of 100, the crossover probability $P_c$ of 0.8, and the mutation probability $P_m$ of 0.05. In addition, the search threshold for AGBCA *limit* is set to 30. For the same algorithm, the experiment adopts the control variable method to change the population size and iteration times, and observes its accuracy and convergence.

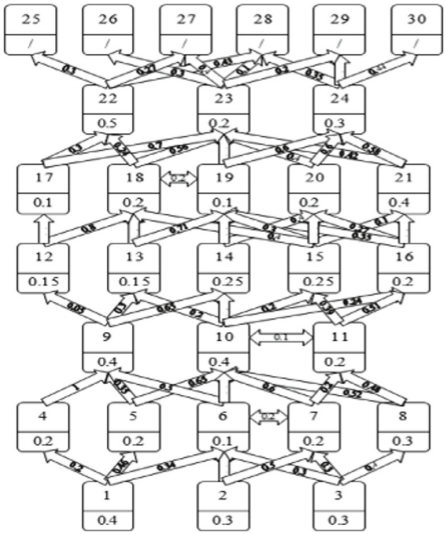



**Fig. 2.** Multi-agent intention reasoning network

The experiment adopts parallel processing architecture, based on 4 CPUs, 128 cores, 2.2 GHz, 128 GB memory and 100 GB capacity. 50 independent experiments are selected for statistics to eliminate randomness. Among them, the experimental results of GA, ABCA and AGBCA algorithms to solve the network intention reasoning problem of cluster target multi-agent are shown in Table 2.

Because the algorithms can reason for cluster targets at the same time, the total time of each algorithm for the same reasoning network is the time required for reasoning the target with the longest time. The average total time of solving the model by the above three algorithms is shown in Table 2.

The AGBCA algorithm is compared with particle swarm optimization algorithm and improved evolutionary algorithm. The three algorithms all adopt the reasoning network structure and parameters shown in Fig. 2, and the comparison results are shown in Table 3.

In the above experiments, GA, ABCA, adaptive genetic bee colony algorithm AGBCA, particle swarm algorithm PSO and improved evolutionary algorithm are applied to the multi-agent intent reasoning network as shown in Fig. 2. By analyzing and comparing the experimental results, we can find that:

(1) GA, ABCA and AGBCA can improve the accuracy of solving problems by adjusting parameters of increasing population size and iteration times, but the corresponding time is longer;
(2) There is horizontal reasoning between layers in the reasoning network structure. The above-mentioned intelligent algorithms can all get effective solutions when dealing with the network structure with loops, and realize algorithm convergence, which shows that they can effectively solve the problem of cluster target intention reasoning based on multi-agent networks.

**Table 2.** Experimental results of GA, ABCA, AGBCA algorithm for solving multi-agent network intention reasoning problem

| Cluser target | Optimal Reasoning Path and Its Comprehensive Trust | algorithm | Experimental serial number | Experimental parameters | | | | | Optimal path times/independent test times | Reasoning accuracy (%) | Average time of multiple tests (s) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Population size | Iterations | Cross probability | Variation probability | Search threshold | | | |
| 1 | 1–4-9–14-19–18-22–28, 0.9430 | GA | 1–1 | 50 | 30 | 0.8 | 0.05 | / | 26/50 | 54 | 0.218 |
| | | | 1–2 | 100 | 30 | | | / | 34/50 | 70 | 0.452 |
| | | | 1–3 | 50 | 60 | | | / | 39/50 | 80 | 0.560 |
| | | ABCA | 1–4 | 50 | 30 | | | / | 30/50 | 60 | 0.246 |
| | | | 1–5 | 100 | 30 | | | / | 43/50 | 86 | 0.550 |
| | | | 1–6 | 50 | 60 | | | / | 42/50 | 85 | 0.520 |
| | | AGBCA | 1–7 | 50 | 30 | | | 30 | 40/50 | 82 | 0.202 |
| | | | 1–8 | 100 | 30 | | | 30 | 48/50 | 96 | 0.344 |
| | | | 1–9 | 50 | 60 | | | 30 | 50/50 | 100 | 0.365 |
| 2 | 2–6-7–10-11–15-20–24-30, 0.7043 | GA | 2–1 | 50 | 30 | | | / | 24/50 | 50 | 0.292 |
| | | | 2–2 | 100 | 30 | | | / | 31/50 | 63 | 0.495 |
| | | | 2–3 | 50 | 60 | | | / | 37/50 | 74 | 0.565 |
| | | ABCA | 2–4 | 50 | 30 | | | / | 32/50 | 65 | 0.241 |
| | | | 2–5 | 100 | 30 | | | / | 34/50 | 78 | 0.457 |
| | | | 2–6 | 50 | 60 | | | / | 40/50 | 82 | 0.519 |

(*continued*)

**Table 2.** (*continued*)

| Cluser target | Optimal Reasoning Path and Its Comprehensive Trust | algorithm | Experimental serial number | Experimental parameters | | | | | Optimal path times/independent test times | Reasoning accuracy (%) | Average time of multiple tests (s) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Population size | Iterations | Cross probability | Variation probability | Search threshold | | | |
| | | AGBCA | 2–7 | 50 | 30 | | | 30 | 38/50 | 77 | 0.214 |
| | | | 2–8 | 100 | 30 | | | 30 | 48/50 | 98 | 0.313 |
| | | | 2–9 | 50 | 60 | | | 30 | 50/50 | 100 | 0.385 |
| 3 | 3–8-10–11-15–20-24–30, 0.6903 | GA | 3–1 | 50 | 30 | | | / | 29/50 | 60 | 0.230 |
| | | | 3–2 | 100 | 30 | | | / | 39/50 | 80 | 0.351 |
| | | | 3–3 | 50 | 60 | | | / | 42/50 | 85 | 0.388 |
| | | ABCA | 3–4 | 50 | 30 | | | / | 32/50 | 66 | 0.288 |
| | | | 3–5 | 100 | 30 | | | / | 40/50 | 80 | 0.440 |
| | | | 3–6 | 50 | 60 | | | / | 46/50 | 92 | 0.469 |
| | | AGBCA | 3–7 | 50 | 30 | | | 30 | 39/50 | 80 | 0.194 |
| | | | 3–8 | 100 | 30 | | | 30 | 49/50 | 98 | 0.264 |
| | | | 3–9 | 50 | 60 | | | 30 | 50/50 | 100 | 0.316 |

**Table 3.** Average total time consumption of GA, ABCA and AGBCA algorithms

| Algorithm | Average total algorithm time under different parameters (s) | | |
|---|---|---|---|
| | Population size: 50<br>Number of iterations: 30 | Population size: 100<br>Number of iterations: 30 | Population size: 50<br>Number of iterations: 60 |
| GA | 0.296 | 0.456 | 0.585 |
| ABCA | 0.342 | 0.668 | 0.527 |
| AGBCA | 0.232 | 0.340 | 0.381 |

**Table 4.** Performance comparison of AGBCA, PSO and improved evolutionary algorithm

| algorithm | Experimental results when the number of individual evaluations is 3000 | | | |
|---|---|---|---|---|
| | Average algorithm time (s) | Accuracy of algorithm (%) | | |
| | | Goal 1 | Goal 2 | Goal 3 |
| AGBCA | 0.383 | 100 | 100 | 100 |
| PSO | 0.440 | 93 | 89 | 89 |
| Improved evolutionary algorithm | 0.433 | 100 | 96 | 100 |

(3) By comparing different algorithms, it can be found that under the same experimental parameters such as population size and iteration times, the adaptive genetic bee colony algorithm is superior to the GA and ABCA in accuracy and time consumption, which shows that the adaptive genetic bee colony algorithm has better performance. At the same time, under the same number of individual evaluations, the performance of adaptive genetic bee colony algorithm is also better than that of particle swarm algorithm and improved evolutionary algorithm. This is because the adaptive genetic bee colony algorithm combines the advantages of GA and ABCA, and makes adaptive improvement, which makes the algorithm solve the contradiction between global search ability and fast convergence ability, and realize the balance between exploration and development, achieving global optimization at the initial stage and fast convergence at the later stage, and better accuracy and convergence.

## 6 Conclusion

In the analysis of target intention, the situation relationship is complicated and difficult to predict. This chapter first draws lessons from the multi-agent thought, reasonably maps the situation information, reasoning rules and object attributes in the multi-agent network, and constructs a multi-agent-based intent reasoning network that can represent the target information and the relationship between them, transforming the problem of target intent reasoning into the problem of searching the optimal reasoning path in

complex networks. Then, in order to solve the above problems, based on the basic ABCA, the idea of crossover of genetic variation in GA is introduced, and an adaptive genetic bee colony reasoning algorithm is proposed combined with the improved adaptive strategy of food origin. Finally, through the simulation example, it is verified that the adaptive genetic bee colony reasoning algorithm can quickly and accurately solve the problem of intent reasoning of cluster targets in multi-agent networks, and has better accuracy, convergence and performance than other algorithms.

## References

1. Fischer, Y., Bauer, A., Beyerer, J.: A conceptual framework for automatic situation assessment. In: IEEE First International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, pp. 234–239. IEEE (2011)
2. He, M.: Research on a Situation Reasoning Method Based on Intelligent Computing. Xidian University (2017)
3. Karaboga, D.: An idea based on honey bee swarm for numerical optimization. Erciyes University, Kayseri, Turkey (2005)
4. Karaboga, D.: A powerful and efficient algorithm for numerical function optimization : Artificial bee colony (ABC) algorithm. J. Global Optim. **39** (2007)
5. Smith, J., Jones, M., Jr., Houghton, L., et al.: Future of health insurance. N. Engl. J. Med. **965**, 325–329 (1999)
6. Icirelli, F., Giordano, A., Nigro, L.: Distributed simulation of situated multi-agent systems. In: IEEE/ACM, International Symposium on Distributed Simulation and Real Time Applications, pp. 28–35. IEEE (2011)

# A Method of UAV Cluster Track Planning Based on Improved PSO Algorithm

Guanghua Ni(✉), Yufeng Li, and Tianzhi Xie

China Electronics Technology Group Corporation Network Communication Research Institute, Shijiazhuang 050081, Hebei, China
genius-ni@163.com

**Abstract.** Aiming at the problem that the complex constraints of UAV cluster offline path planning are easy to lead to no solution in calculation, a method of UAV cluster offline path planning based on improved particle swarm optimization (PSO) is proposed. The method first constructs a cost function that combines UAV energy consumption, high threat, cluster collision avoidance and communication restrictions. Then, the acceleration coefficient and the maximum speed in the PSO algorithm are adaptively changed linearly to obtain the global optimal solution with fewer iterations, and the mutation strategy is updated for the cost function value of particles, so that the influence of beneficial particles is strengthened and the influence of unrelated particles is weakened. The final simulation results show that compared with the traditional PSO algorithm, this method can effectively reduce the number of algorithm iterations and improve the success rate of UAV cluster offline path planning.

**Keywords:** UAV cluster · Offline track planning · Particle Swarm Optimization · Mathematical

## 1 Introduction

With the vigorous development of UAV field, UAV cluster plays an increasingly important role in various scenarios, such as environmental monitoring, agricultural irrigation, auxiliary communication, etc. However, in the complex environment, the flight path of UAV cluster is always the key problem to be solved.

In recent years, swarm intelligence optimization technology has been widely used in UAV route planning. Reference [2] uses improved genetic algorithm to achieve track planning, which solves the problem that it is difficult to accurately plan a large number of control points in a flight mission, and smoothes the track through Bezier curve to generate a feasible track that meets the mission needs. Reference [3] uses B-spline curve to plan UAV path, and uses ant colony algorithm to calculate control vertices of spline curve. In reference [4] and [5], UAV target location is modeled as ant colony food source, hostile defense area is modeled as ant colony search area, and UAV path is optimized through path length and regional threat as cost function. Although the ant colony algorithm has strong robustness, its convergence speed is slow due to lack of


Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 313–322, 2023.
https://doi.org/10.1007/978-981-99-2362-5_39

early information. Reference [6] studied the improvement of multiple particle swarm optimization (PSO) algorithms, and improved PSO algorithms from three aspects: factor shrinkage, integral weight and Gaussian mutation. Reference [7] proposed a real-time path planning method based on improved differential evolution algorithm, which uses B-spline curve to plan UAV path, and the control vertices of the spline curve are optimized by differential evolution algorithm, effectively improving the search efficiency of the algorithm.

Among various UAV path planning algorithms based on swarm intelligence optimization, PSO algorithm has simple implementation principle, strong operability, good compatibility with complex constraints, and can be globally optimized through all particles, so it is often used in UAV path planning. Reference [8], the author improved PSO algorithm and combined sensors to achieve the positioning of UAV in the task. For the task requirements of UAV search and tracking in unknown environments, Reference [9] studied an efficient hybrid method of local particle swarm structure based on PSO algorithm. Reference [10], the author proposed a new idea to independently evaluate and evolve the waypoints in 2D UAV path planning. Reference [11] proposed a particle swarm inertia coefficient adjustment method based on Logistic function, which solved the task allocation problem of UAV cluster. Reference [12] proposed a PSO optimization algorithm based on cooperation and competition, focusing on solving the problem of energy consumption of UAVs during search tasks, and realizing the balance of communication consumption and energy between UAV clusters. In reference [13], the author integrated chaos mapping and Gaussian mutation mechanism into PSO algorithm, which also solved the problem that PSO algorithm is easy to fall into local minima and premature. Reference [14] and [13] have similar improvement purposes, but the authors of reference [14] focus more on solving the problem of falling into local optimum by improving the speed update rule.

This paper proposes an offline path planning method for UAV cluster based on improved PSO algorithm. The innovative work of this paper includes: First, a cost function considering the limitations of UAV itself and UAV cluster is constructed. Then, an adaptive linear change strategy is proposed for the acceleration coefficient and maximum speed in the PSO algorithm to obtain the global optimal solution with fewer iterations. A mutation update strategy is proposed for the cost function value of particles. The influence of beneficial particles is strengthened, The influence of unrelated particles is weakened. At last, the simulation results show that this method can not only meet the strong compatibility of PSO algorithm itself with complex constraints and is easy to solve, but also improve the problem that PSO algorithm itself is easy to fall into local optimization, and achieve offline path planning of UAV with a high success rate.

## 2 Problem Modelling

### 2.1 Cost Function

The track cost of UAV is the evaluation of track quality. A low surrogate value indicates that the planned track is better, while a high surrogate value indicates that the corresponding track is unavailable. The track cost acts on each track point and each segment

of track. Track planning is to find an effective path between the starting point and the end point to minimize the track cost. The cost function of a single UAV built in this paper is:

$$f_i = k_1 f_L + k_2 \sum_{n=1}^{N} f_{H_n} + k_3 f_{\text{dis}} + k_4 f_{\text{com}} \tag{1}$$

where, $f_L$ is the track energy consumption cost, $f_{H_n}$ is the high threat cost, $f_{\text{dis}}$ is the UAV cluster collision cost, $f_{\text{com}}$ is the UAV cluster communication cost, $k_1$-$k_4$ is the corresponding cost factor, and $N$ is the number of track points. For the corresponding UAV cluster, its cost function is the sum of the cost functions of a single UAV, that is:

$$F = \sum_{i=1}^{M} f_i \tag{2}$$

where, $M$ is the number of UAVs.

**Track Energy Consumption Cost**

Track energy consumption cost refers to the energy consumption during UAV flight. Since the energy that UAV can carry is limited, it is necessary to consider the energy consumption when planning the track. The energy consumption cost formula in this paper is:

$$f_L = \sum_{i=1}^{N-1} a_i L_i \tag{3}$$

where $\alpha_i$ is the energy consumption coefficient, and $L_i$ is the length of the track in the ith segment.

**High Threat Cost**

For the safe flight of UAV, the planned path should not cross the terrain and must avoid all mountains and threat areas in the environment. Therefore, the high threat cost in this paper is:

$$f_{H_n} = \begin{cases} \frac{H_n - H_{\min}}{H_{\max} - H_{\min}} & H_{\min} < H_n < H_{\max} \\ I & other \end{cases} \tag{4}$$

where, $I$ is the penalty value when the flight altitude does not meet the requirements, $H_{\min}$ is the minimum flight altitude, and $H_{\max}$ is the maximum flight altitude.

**UAV Cluster Collision Avoidance Cost**

If there are $M$ UAVs in the UAV cluster, the distance between the i-th UAV and the jth UAV is $d_{ij}$, and the minimum safe collision avoidance distance of the UAV is set to $d_{\text{safe}}$, then $d_{ij}$ must meet the requirements of $d_{ij} \geq d_{\text{safe}}$. When the distance between any two UAVs is $d_{ij} < d_{\text{safe}}$, there is a great possibility of collision. Therefore, the distance between UAVs must meet the limit of safe collision avoidance. The distance cost formula can be obtained as follows.

$$\begin{aligned} f_{\text{dis}} &= \sum_{i=1}^{M-1} \sum_{j=2}^{M} C_{i,j} \\ C_{i,j} &= \begin{cases} 1 & d_{ij} \leq d_{\text{safe}} \\ 0 & other \end{cases} \end{aligned} \tag{5}$$

where, $C_{i,j}$ indicates whether the safety distance limit is met between the i-th UAV and the jth UAV. When the distance between any two UAVs is less than $d_{\text{safe}}$, $C_{i,j}$ is 1, and when the distance limit is met, $C_{i,j}$ is 0.

**UAV Trunking Communication Cost**
During the flight, UAV clusters need to keep the communication link smooth to facilitate information interaction. The communication cost formula is as follows:

$$f_{\text{com}} = \sum_{i=1}^{M} \sum_{i \neq j}^{M} D_{i,j}$$
$$D_{i,j} = \begin{cases} 0 \ \textit{Communicable} \\ 1 \ \textit{Incommunicable} \end{cases} \tag{6}$$

Among them, $D_{i,j}$ indicates whether the i-th UAV can communicate with the jth UAV. When any two UAVs cannot communicate, $D_{i,j}$ is 1, and when they can communicate, $D_{i,j}$ is 0. Whether communication is possible is judged by the maximum communication distance.

## 3 UAV Cluster Track Planning Method Based on Improved PSO

### 3.1 Improved PSO Algorithm

PSO algorithm is one of swarm intelligence optimization algorithms, which has strong compatibility for complex optimization problems. For the traditional PSO algorithm, it usually includes the following steps.

1. Initialization of parameters and particle positions;
2. Calculation of local and global optimal solutions.
3. Update of particle positions.
4. Update of particle velocities.
5. Update of local and global optimal solutions.
6. Repeat 3–5 until the iteration stop condition is met.

When planning the path of the UAV cluster, due to the complexity of the problem, the PSO algorithm is prone to fall into the local optimal situation, leading to premature particles, which leads to global error convergence, and can not get the optimal solution or only get the feasible solution. In this paper, we will solve these problems by improving the initial distribution of particles, adaptive adjustment of parameters, and improving the position update strategy.

**Particle Initialization Based on Chaos**
In order to enrich the diversity of particle initialization in UAV cluster route planning, this paper chooses logical mapping to improve particle initialization. Logic mapping is a one-dimensional discrete time mapping, which has a simple form and unexpected complexity. Through logic mapping, complex chaotic oscillation behavior can be obtained. It is defined as:

$$W_{n+1} = \lambda W_n (1 - W_n) \tag{7}$$

Among them, when $\lambda = 4$, the logic mapping can uniformly generate chaotic signals to meet the requirements of diversified initial values of particles. $W_n$ is the nth particle, that is, the collection of all UAV track points.

**Fixed Parameter Adaptive Adjustment**

In the PSO algorithm, there are three important parameters involved, namely, inertial coefficient $\omega$, individual learning factor $c_1$ and social learning factor $c_2$. Among them, $\omega$ is mainly used to balance the speed of local search and global search. In the early stage of PSO algorithm search, it focuses on the increase of population diversity, and in the late stage of iteration, it focuses on the convergence of the objective function. Therefore, in order to improve the search accuracy in the early stage and the convergence speed in the later stage, it is possible to conduct linear changes to the inertia coefficient, that is:

$$\omega = \omega_{\max} - \frac{t}{T}(\omega_{\max} - \omega_{\min}) \tag{8}$$

where, $\omega_{\max}$ and $\omega_{\min}$ are the maximum and minimum values of the inertia coefficient respectively, $T$ is the number of iterations, and $t$ is the tth iteration.

Similarly, individual learning factor $c_1$ acts on local search, and social learning factor $c_2$ acts on global convergence, making $c_1$ play a major role in the first half and $c_2$ play a major role in the second half. Therefore, the changes of $c_1$ and $c_2$ can be designed as follows:

$$c_1 = c_{\max} - \frac{(c_{\max} - c_{\min})t}{T} \tag{9}$$

$$c_2 = c_{\min} + \frac{(c_{\max} - c_{\min})t}{T} \tag{10}$$

wherein, $c_{\max}$ and $c_{\min}$ correspond to the maximum and minimum values of learning factors.

### 3.2 UAV Cluster Route Planning Method

When applying the improved PSO to the path planning of UAV clusters, it is assumed that there are $K$ particles in total, and the kth particle is the set of all UAV track points, that is

$$W_k = (w_1, w_2, ..., w_M)^{\mathrm{T}} \tag{11}$$

where, $w_m$ is the set of track points of the mth UAV. The local optimal solution of the kth particle and the global optimal solution of the entire particle swarm are recorded as $\mathbf{p}_k$ and $\mathbf{g}$ respectively. Then the offline path planning method flow of UAV cluster based on improved PSO is shown in the following table (Table 1).

**Table 1.** UAV Cluster Track Planning Based on Improved PSO

Input: terrain parameters, cost function related parameters, improved PSO algorithm related parameters;

Output: $W_k^{\text{best}}$ track point;

Initialize all particle$W_k$ according to formula (3.1);

$p_k$ and g are determined by the cost function (1.2);

while $t \le T$ do

The particle position is updated by $W_k = W_k + v_k$;

Recalculate particle position

The particle velocity is updated by $v_k = \omega v_k + c_1 r_1(\mathbf{p}_k - W_k) + c_2 r_2(\mathbf{g} - W_k)$, $r_1$ and $r_2$ obey 0-1 uniform distribution

Update PSO algorithm parameters adaptively according to (3.2) - (3.5)

Update $\mathbf{p}_k$ and $\mathbf{g}$

if $|F_{\min}(t) - F_{\min}(t-1)| \le \tau$

break;

end if

end while

Output $W_k^{\text{best}} = \mathbf{g}$ as UAV cluster track;

## 4 Simulation Result

### 4.1 Simulation Parameter Setting

In order to verify the performance of the proposed UAV cluster track planning algorithm based on improved PSO, this paper compares the proposed algorithm with the UAV cluster track planning algorithm based on traditional PSO under various conditions. The relevant parameters of UAV cluster are shown in the following table:

| | |
|---|---|
| Number of UAVs $M$ | 3 |
| Number of track points $N$ | 15 |
| Safe distance $d_{\text{safe}}$ (km) | 0.005 |
| Flight altitude $H_{\max}/H_{\min}$ (km) | 0.15/0.05 |
| Energy consumption coefficient $\alpha_i$ | 1 |

The parameters of the improved PSO algorithm are shown in the following table:

| | |
|---|---|
| Inertia coefficient threshold $\omega_{max}/\omega_{min}$ | 0.9/0.4 |
| Learning factor threshold $c_{max}/c_{min}$ | 3.5/0.5 |
| Maximum speed threshold $V_{max}/V_{min}$ | 0.8/0.3 |
| Maximum Iterations $T$ | 500 |

The adopted map size is 10 km × 10 km, 10 peaks in total, randomly generated at peak location, height and range.

### 4.2 Comparison of Route Planning Simulation Results

Figures 1, 2, 3, 4, 5 and 6 shows the results of UAV cluster track planning based on improved PSO algorithm and traditional PSO algorithm and the corresponding convergence curve. It can be seen from the planning results that all tracks do not collide with each other and are kept within a certain distance, which can meet the constraints of UAV cluster collision avoidance and communication. When the population size is 1000, it starts to converge after 20 iterations. Compared with traditional PSO, the objective function value of the proposed algorithm is obviously lower after convergence, and the traditional PSO algorithm starts to converge after 300 iterations.

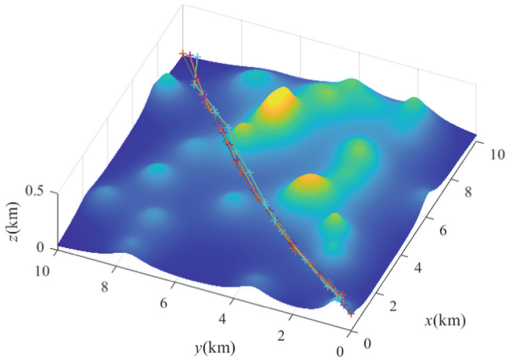



**Fig. 1.** Three dimensional view of UAV cluster route planning results based on improved PSO (PSO population size 1000)

Below is a comparison of the success rate of UAV cluster track planning algorithms based on improved PSO and traditional PSO. The map is generated randomly. The convergence precision of the cost function is set to $\tau = 0.06$, and the results after 1000 simulations are shown in the following table. If 16 UAVs are tested, the success rate of the traditional PSO algorithm with a population size of 1000 is 31.4%, while the success rate of the improved algorithm is 91.8%. It can be seen from the comparison of success rates that the success rate of UAV cluster offline route planning based on improved PSO algorithm is higher. Algorithm parameters Before algorithm improvement.

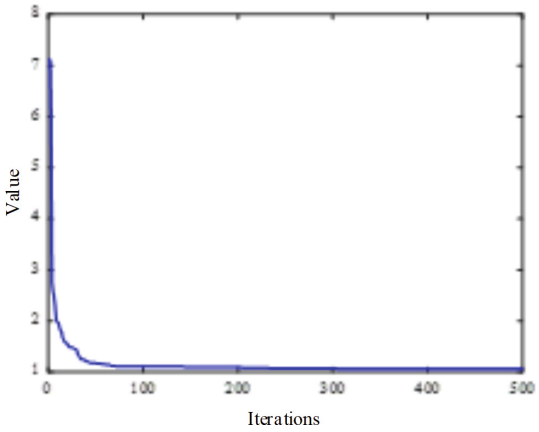



**Fig. 2.** Route planning convergence curve based on improved PSO (PSO population size 1000)

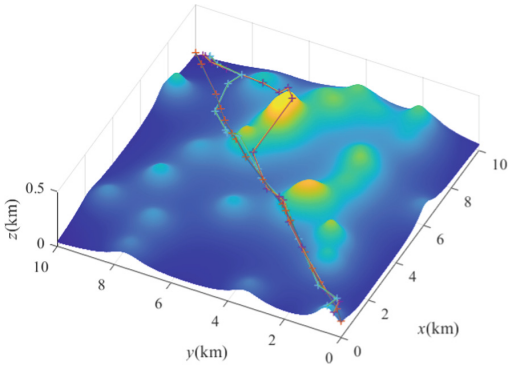



**Fig. 3.** Three dimensional view of UAV cluster route planning results based on traditional PSO (PSO population size 1000)

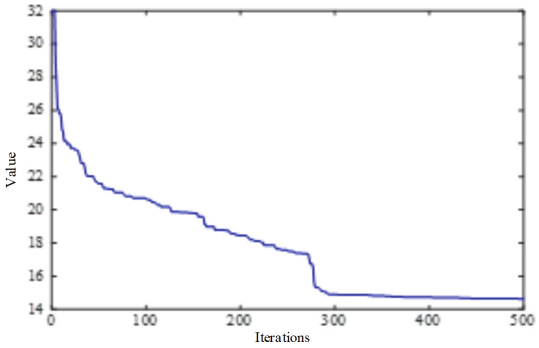



**Fig. 4.** Convergence curve of route planning based on traditional PSO (PSO population size 1000)

| Algorithm parameters | Before algorithm improvement | After algorithm improvement |
|---|---|---|
| $K = 500, M = 3$ | – | 87.2% |
| $K = 1000, M = 3$ | 72.8% | 93.6% |
| $K = 500, M = 16$ | – | 84.2% |
| $K = 1000, M = 16$ | 31.4% | 91.8% |

## 5 Conclusion

This paper improves the traditional PSO algorithm, which is easy to fall into the local optimal problem, and applies it to the offline path planning of UAV clusters. Firstly, based on the principle of route planning, the improvement of PSO algorithm is studied, including particle initialization, parameter adaptive processing and position update strategy. In addition, this paper designs the specific calculation steps of offline path planning of UAV cluster based on improved PSO algorithm. Finally, the proposed UAV cluster route planning algorithm is simulated. The results show that the target function value of the UAV cluster route planning algorithm based on improved PSO is significantly lower than that of the traditional PSO algorithm after convergence, and the success rate of the improved PSO algorithm is better than that of the traditional PSO algorithm.

## References

1. Zhang, H., Xin, B., Dou, L., et al.: A review of cooperative path planning of an unmanned aerial vehicle group. Front. Inf. Technol. Electron. Eng. **21**(12), 1671–1694 (2020)
2. Sahingoz, K.O.: Generation of Bezier curve-based flyable trajectories for multi-UAV systems with parallel genetic algorithm. J. Intell. Rob. Syst. **74**(1–2), 499–511 (2014)
3. Cheng, C.T., Fallahi, K., Leung, H., et al.: Cooperative path planner for UAVs using ACO algorithm with Gaussian distribution functions. In: IEEE International Symposium on Circuits & Systems, pp. 173–176 (2009)
4. Zhang, C., Zhen, Z., Wang, D., et al.: UAV Path planning method based on Ant Colony Optimization. In: Chinese Control & Decision Conference, pp. 3790–3792 (2010)
5. Zhang, D., Xian, Y., Li, J., et al.: UAV path planning based on chaos Ant Colony Algorithm. In: International Conference on Computer Science and Mechanical Automation (CSMA), pp. 81–85 (2016)
6. Pehlivanoglu, Y.V.: A new particle swarm optimization method for the path planning of UAV in 3D environment. J. Aeronaut. Space Technol. **5**(4), 1–14 (2012)
7. Zhang, X., Chen, J., Xin, B., et al.: Online path planning for UAV using an improved differential evolution algorithm. World Congr. **44**(1), 6349–6354 (2011)
8. Na, H., Yoo, S.: PSO-based dynamic UAV positioning algorithm for sensing information acquisition in wireless sensor networks. IEEE Access **7**, 77499–77513 (2019)
9. Saadaoui, H., Bouanani F., Illi, E.: Information sharing based on local PSO for UAVs cooperative search of moved targets. In: 2018 International Conference on Advanced Communication Technologies and Networking (CommNet), pp. 1–6 (2018)
10. Yang, P., Tang, K., Lozano, J.A., et al.: Path planning for single unmanned aerial vehicle by separately evolving waypoints. IEEE Trans Robot. **31**(5), 1130–1146 (2015)

11. Gou, Q., Li, Q.: Task assignment based on PSO algorithm based on logistic function inertia weight adaptive adjustment. In: 2020 3rd International Conference on Unmanned Systems (ICUS), pp. 825–829 (2020)
12. Saadaoui, H., El, F.: Communication and energy optimization of local PSO-assisted Multi-UAVs for moving targets exploration. In: 2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet), pp. 1–7 (2021)
13. Tian, D., Shi, Z.: MPSO: modified particle swarm optimization and its applications. Swarm Evol. Comput. **41**, 49–68 (2018)
14. Xia, X., Xing, Y., Wei, B., et al.: A fitness-based multi-role particle swarm optimization. Swarm Evol Comput. **44**, 349–364 (2019)

# Author Index





Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 323–324, 2023.
https://doi.org/10.1007/978-981-99-2362-5