tech.zte.com.cn

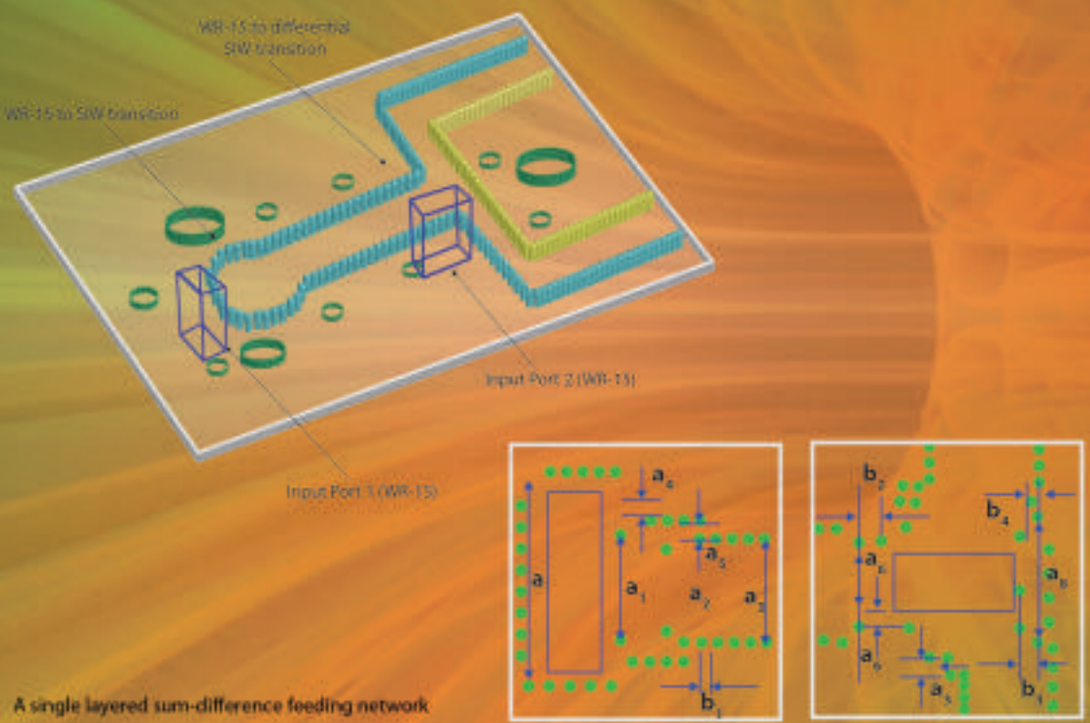
ISSN 1673-5188
CN 34-1294/ TN
CODEN ZCTOAK

ZTE COMMUNICATIONS

An International ICT R&D Journal Sponsored by ZTE Corporation

December 2016, Vol. 14 S1

SPECIAL TOPIC:
Multi-Gigabit Millimeter-Wave Wireless Communications



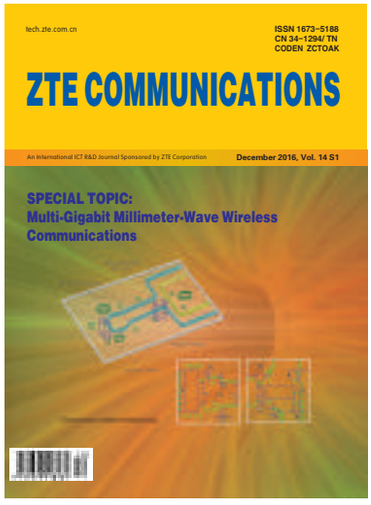
A single layered sum-difference feeding network

9 771673 518161 12

ZTE中兴 ZTE COMMUNICATIONS
VOLUME 14 NUMBER S1 DECEMBER 2016

ZTE Communications Editorial Board

CONTENTS



Special Topic: Multi-Gigabit Millimeter-Wave Wireless Communications

Review

CONTENTS

ZTE COMMUNICATIONS
Vol. 14 No. S1 (Issue 54)
Quarterly
First English Issue Published in 2003

Supervised by:
Anhui Science and Technology Department

Sponsored by:
Anhui Science and Technology Information Research Institute and ZTE Corporation

Staff Members:
Editor-in-Chief: CHEN Jie
Executive Associate Editor-in-Chief: HUANG Xinming
Editor-in-Charge: ZHU Li
Editors: XU Ye, LU Dan, ZHAO Lu
Producer: YU Gang
Circulation Executive: WANG Pingping
Assistant: WANG Kun

Editorial Correspondence:
Add: 12F Kaixuan Building,
329 Jinzhai Road,
Hefei 230061, P. R. China
Tel: +86-551-65533356
Fax: +86-551-65850139
Email: magazine@zte.com.cn

Published and Circulated (Home and Abroad) by:
Editorial Office of
ZTE Communications

Printed by:
Hefei Tiancai Color Printing Company

Publication Date:
December 25, 2016

Publication Licenses:

ISSN 1673-5188
CN 34-1294/ TN

Advertising License:
皖合工商广字0058号

Annual Subscription:
RMB 80

Research Paper

Roundup

Guest Editorial

ZHANG Yueping, GUAN Ke, and WANG Junjun

Multi-Gigabit Millimeter-Wave Wireless Communications

▶ ZHANG Yueping

ZHANG Yueping is a professor of electronic engineering at Nanyang Technological University, Singapore. His current research interests include the development of antenna-on-chip technology and characterization of chip-scale radio propagation channels at terahertz for wireless chip area network. Prof. ZHANG is a member of the Field Award Committee of the IEEE Antennas and Propagation Society (AP-S). He received the Sino-British Technical Collaboration Award in 1990 for his contribution to subsurface radio. He received the William Mong Fellowship from the University of Hong Kong in 2005. He received the IEEE AP-S Schelkunoff Prize in 2012. He was selected in 2012 by the Recruitment Program of Global Experts of China as a Qianren Scholar at Shanghai Jiao Tong University. Prof. ZHANG was elevated to IEEE Fellow in 2009 for contributions to integrated antennas and subsurface radio.

▶ GUAN Ke

GUAN Ke received BE and PhD degrees from Beijing Jiaotong University, China in 2006 and 2014, respectively. He is an associate professor in State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. In 2015, he has been awarded a Humboldt Research Fellowship. He was the recipient of 2014 International Union of Radio Science (URSI) Young Scientist Award. His paper received the honorable mention in the third International URSI student prize paper competition in 2014 URSI GASS. From 2011 to 2013, he has been a research scholar at the Institut fur Nachrichtentechnik (IfN) at Technische Universitaet Braunschweig, Germany. From September 2013 to January 2014, he was invited to conduct joint research in Universidad Politecnica de Madrid, Spain. His current research interests are in the fields of measurement and modeling of wireless propagation channels, high-speed railway communications, and channel characterization for future millimeter wave and terahertz communication systems. He has authored/co-authored over 100 research papers in international journals and conferences.

▶ WANG Junjun

WANG Junjun received the BE degree from Shandong University of Technology, China, in 1999, ME degree from Shanghai University, China, in 2002, and PhD degree from School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2006, all in electronic engineering. From 2005 to 2008, she worked at Sony Electronics (Singapore) as a senior R&D engineer. Since 2009, she has been an associate professor with the School of Electronic and Information Engineering, Beihang University, China. Her research interests include design of integrated in-package and on-chip antennas, electromagnetic compatibility, wireless energy harvesting and radio frequency circuits.

The exponential growth of using wireless devices in recent years has motivated the exploration of the millimeter-wave (mmWave) frequency spectrum for multi-gigabit wireless communications. Recent advances in antenna technology, radio-frequency complementary metal-oxide semi-conductor (RF CMOS) process and high-speed baseband signal processing algorithms have promised the feasibility of millimeter-wave wireless communications. The multi-gigabit-per-second data rate of millimeter-wave wireless communication systems is leading to applications in many important scenarios, such as wireless personal area networks (WPAN), wireless personal area networks (WLAN), back-haul for cellular systems. The frequency bands include 28 GHz, 38 GHz, 45 GHz, 60 GHz, E-BAND, and even beyond 100 GHz.

The special issue aims to present some major achievements of the research and development in multi-gigabit millimeter-wave wireless communications. It includes four technical contributions from leading researchers in mmWave communications. The first paper entitled "Substrate-Integrated Waveguide-Based Monopulse Slot Antenna Arrays for 60 GHz Applications", co-authored by ZHU, XUE and LIAO, presents the substrate-integrated waveguide (SIW)-based monopulse slot antenna arrays for the application of 60 GHz monopulse tracking systems. The second paper entitled "Millimeter Wave and THz Propagation Channel Modeling for High-Data Rate Railway Connectivity — Status and Open Challenges" is co-authored by Kürner, GUAN, Molisch, AI, HE, LI, TIAN, DOU and ZHONG. In this paper, the authors provide elementary discussions on bandwidth requirements of high-data rate railway connectivity, and highlight the open challenges in terms of wave propagation, static channel, and dynamic channel.

Co-authored by XU, MENG, MA and YEO, the third paper entitled "State of the Art in Passive Bandpass Filter Solutions for 60-GHz Communications", reviews the state-of-the-art filter designs for 60-GHz applications. Design methodology, design technology, key performance parameters, similarities and differences, advantages and drawbacks, and future trends are explored and studied. The last (but not least) paper "Low-Power High-Efficiency Multi-Gigabit 60 GHz Transceiver Systems Routing in Vehicular Environments" is co-authored by Byeon and Park. It proposes low-power high-efficiency multi-gigabit 60 GHz transceiver systems for short-range communications. The antenna-in-package module with the transceiver demonstrates mobile-to-display 1080p Full-HD video transmission over a distance of 60 cm.

We would like to thank all the authors for choosing this special issue to publish their new research results and insights, all the reviewers for their valuable review comments which help to improve the technical quality and presentation of this special issue, and the editorial official of *ZTE Communications* for all the support and help during the editorial process of this special issue. We are sure this special issue will again be quite informative and a pleasure for you to browse through and read in depth.

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 001
http://www.cnki.net/kcms/detail/34.1294.TN.20161208.1003.002.html, published online December 8, 2016

Substrate Integrated Waveguide Based Monopulse Slot Antenna Arrays for 60 GHz Applications

ZHU Jianfeng[1, 2], XUE Quan[1, 2], and LIAO Shaowei[1, 2]
(1. State Key Laboratory of Millimeter Waves, City University of Hong Kong, Kowloon, Hong Kong SAR, China;
2. CityU Shenzhen Research Institute, Shenzhen 518057, China)

Abstract

Monopulse slot antenna arrays based on substrate integrated waveguide (SIW) are proposed for the application of 60 GHz monopulse tracking systems in this paper. The sum-difference monopulse comparator can provide a high amplitude and phase balance over wide frequency band and no phase delay technique is required for the difference channel. Resonant slot antennas are adopted as the radiating elements since they can be integrated with the sum-difference monopulse comparator in a single layer with a compact size. Two monopulse arrays with 2 × 4 and 4 × 4 slot elements are designed, fabricated, and measured. Measured results show that the proposed antenna arrays have wide bandwidth covering the unlicensed 60-GHz band. The peak sum beam gain is 13.85 dBi for the 2 × 4 element array and 16.24 dBi for the 4 × 4 element array. The peak difference beam gain is 11.20 dBi for the 2 × 4 element array and 12.11 dBi for the 4 × 4 element array and the maximum null depth can reach -40 dB.

Keywords

resonant slot antenna array; monopulse; substrate integrated waveguide (SIW); millimeter-wave (mmWave)

1 Introduction

The monopulse technique, which compares a signal received simultaneously from the sum (Σ) and difference (Δ) channels, is widely used in millimeter wave radar and satellite systems for target tracking [1]. Monopulse antennas can be simply realized by changing the phases of different radiating elements. The sum beam is achieved when radiating elements are all excited in phases, while two difference beams are obtained by employing an 180-degree phase-shifting to excite several elements. The former can generate one main lobe along the target direction, and the latter exhibits a null in the same direction.

Basically, the monopulse antenna consists of two parts: the monopulse comparator and the radiating elements. The comparator is utilized to generate the sum and difference patterns and the radiating antenna can be planar or three-dimensional (3-D). The 3-D monopulse antennas are usually in a form of reflector antennas, such as Cassegrain parabolic antennas and horn antennas with waveguide monopulse comparators [2], [3]. Though these antennas can obtain good characteristics, the antennas and corresponding monopulse comparator are usually complicated and bulky in structure. Microstrip-based antennas are attractive for monopulse applications due to their low cost and compact size. Various microstrip monopulse antennas have been proposed with good performance [4]–[7]. However, the microstrip technique is inappropriate for V-band applications as the loss of the microstrip structure cannot be neglected at high frequencies. Substrate integrated waveguide (SIW) technology has been demonstrated as a low-loss and high-integration technique for passive components. Till now, several SIW monopulse antenna arrays have been presented [8]–[11]. A substrate integrated monopulse antenna array for W-band was proposed in [8] with good performance, low cost and easy fabrication, but the design complexity is relatively high. LIU et al. [9] proposed a two dimensional SIW monopulse slot antenna array without any microstrip structures. However, the bandwidth of the array is narrow and it suffers from phase imbalance, which results in the deterioration of antenna performance. A SIW monopulse antenna array with filtering respond was proposed in [10], where the comparator was achieved by a square dual-mode SIW cavity resonating at its TE_{201} and TE_{102} modes. A low-profile antenna array combines the scanning capabilities of pillbox configurations and enhanced resolution of two-quadrant monopulse technique was realized in [11].

To overcome these challenges, a SIW-based monopulse slot

This project is supported by the National Basic Research Program of China ("973" Program) under Grant No. 2014CB339900 and the National Natural Science Foundation of China under Grant No. 61372056.

antenna array is proposed, which has good characteristics including wide bandwidth, deep null-depth, high radiation efficiency and high port isolation. Besides, there are two additional advantages of the novel design: 1) The phase delays of the feeding network are frequency independent as the SIW magic T of the comparator, which combines the E-plane and H-plane junctions, can be regarded as a 180-degree hybrid. This provides a constant phase difference between the output ports for the difference channel as well as high degree decoupling between the input ports with a wide frequency range. In addition, as the feeding network is entirely implemented with SIW technology, there is no radiation from feeding network. 2) As for fabrication, the radiating elements and the monopulse comparator are implemented in a single layer and can be realized with standard printed circuit board (PCB) technology.

2 Antenna Design

2.1 Design and Evaluation of Monopulse Comparator

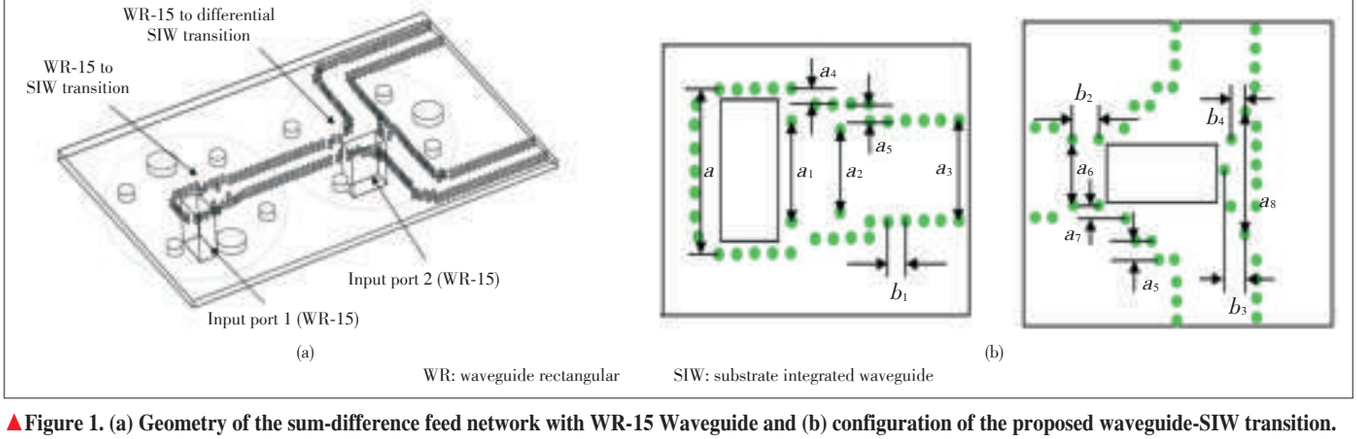
In literature, most of the SIW-based sum and difference feeding networks for mmWave frequencies are implemented by 3 dB directional couplers and 90-degree phase shifters. Although these structures can be well-integrated in a single layer, they can only operate in a rather narrow bandwidth. However, the T-junction shows broad impedance bandwidth, and the outputs of the H-plane junction exhibits high amplitude and phase balance while the E-plane junction can exhibit the same amplitude and reverse phase in a rather broad bandwidth. Inspired by this, a single layered sum-difference feeding network is proposed as shown in **Fig. 1**. In this design, Rogers 5880 PCB laminates with a thickness of 0.787 mm and a relative permittivity of 2.2 is used. Both the sum and difference channels are directly fed by the WR-15 waveguide so that additional radiation from the feed discontinuity can be avoided.

As can be seen in Fig. 1, the sum channel (when port 1 is excited) is based on H-plane Tee where the energy flows into the two arms of the SIW with the same amplitude and phase [12]. In order to obtain good impedance matching between the WR-15 waveguide and the SIW, the widths of the short end sections of SIWs are enlarged from 2.7 mm (a_3) to 4.44 mm (a). The width of the SIW is chosen as 2.7 mm (a_3) so that only the dominant mode exists in the SIW over the whole operating frequency band. For the difference channel (when port 2 is excited), the WR-15 waveguide fed is located in the center of the two arms of the SIW and the axis of the WR-15 is parallel to the plane of the E-field vector of the SIW. The energy is coupled from the WR-15 waveguide by means of E-field and then divided into two arms with the same amplitude but 180-degree phase difference naturally [13]. Due to the field symmetry, the isolation between the sum and difference ports is inherently high. It is worth noting that the presence of the cutting slot for the difference port will have little influence on the response of the previous sum channel because of the high isolation. **Fig. 2** shows the E-field current distributions of the comparator when port 1 and port 2 are excited. In order to obtain a good match of the waveguide-SIW transition of both the sum and difference ports, the posts adjacent to the ports need to be carefully optimized. Detailed configuration of the waveguide-SIW transition is shown in Fig. 1b and corresponding dimensions are shown in **Table 1**.

The sum-difference feed network is arranged back-to-back to evaluate its loss, as shown in **Fig. 3**. The simulated and measured S-parameters are shown in **Fig. 4**. As can been seen, both the simulated and measured reflection coefficients (S_{11}) are lower than −10 dB and the insertion loss is around 1 dB for the simulation and 3.6 dB for the measurement. The simulated and measured S_{22} are lower than −15 dB at most of the band. The simulated insert loss is around 0.6 dB and measured loss is about 1.3 dB. Therefore, the measured insert loss of the monopulse feed network is about 2.5 dB for port 1 and 0.65 dB for port 2.

2.2 Design and Measurement of 2 × 4 Slot Antenna Array

Based on the sum and difference feeding network presented in the previous section, one-dimensional monopulse antenna



▲**Figure 1. (a) Geometry of the sum-difference feed network with WR-15 Waveguide and (b) configuration of the proposed waveguide-SIW transition.**

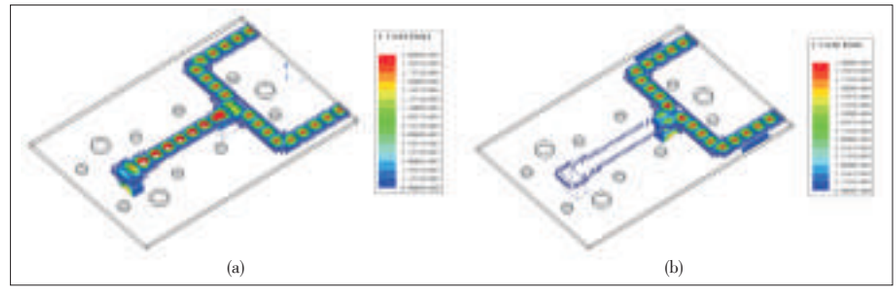
arrays with both wide impedance bandwidth and good radiation patterns are presented and discussed. The slot resonant antenna arrays are chosen as the radiating elements because the slot antennas can be well integrated with the feed network. The antennas are simulated by ANSYS High Frequency Structure Simulator (HFSS) [14]. After fabrication, an Agilent network analyzer (E8361A) measures its reflection coefficient. The broadside gain and radiation patterns are measured by an in-house far-field millimeter wave antenna measurement system. Due to the system limitation, only the radiation pattern in the upper half space is measured [15]. The geometry of the 2×4 slot antenna array is shown in **Fig. 5a**. The photograph of the fabricated prototype is shown in **Fig. 5b**.

The simulated and measured reflection coefficients of the 2×4-element slot antenna arrays are given in **Fig. 6**. The measured results show that the reflection coefficients of the arrays are lower than −15 dB at most of the unlicensed 60-GHz band. Moreover, the isolation is better than 20 dB.

The array generates the difference beam in E-plane (*yoz* plane) and the sum beam in both E-plane and H-plane (*xoz* plane) when ports 1 and 2 are excited, respectively. The simulated and measured normalized radiation patterns at 60 GHz are shown in **Fig. 7a** and the results agree well with each other. The null-depth is below −35 dB for the difference beam. The measured sum and difference radiation patterns at 58 GHz and 64 GHz are shown in **Fig. 7b**. The measured maximum gain is 13.85 dBi for the sum beam and 11.2 dBi for the difference beam while the minimum one is 12.1 dBi and 8.8 dBi for the sum beam and difference beam, respectively.

2.3 Design and Measurement of 4 × 4 Slot Antenna Array

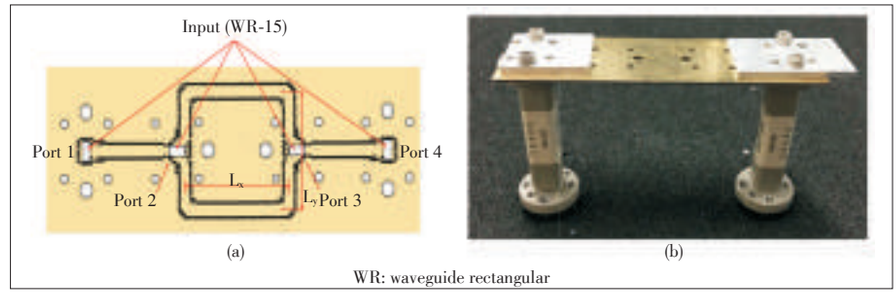
In order to further increase the gain, a 4×4 element array is designed with two two-way equal power dividers, as shown in **Fig. 8**. The two-way power dividers are crucial to the 4×4 antenna array as the



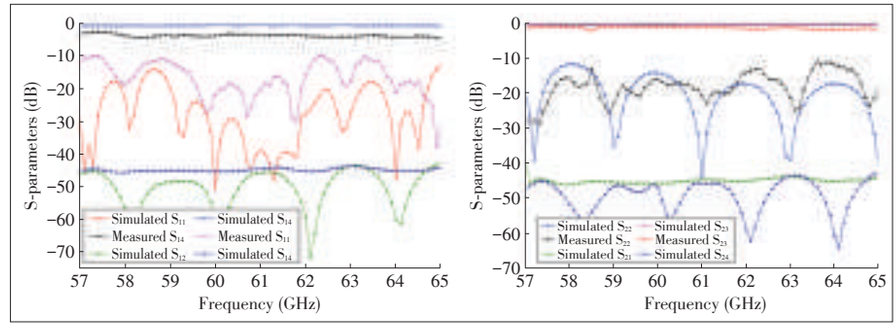
▲**Figure 2. E-field distributions of the comparator when (a) port 1 and (b) port 2 are excited.**

▼**Table 1. Dimensions of the waveguide: SIW transition**

Parameter	a	a_1	a_2	a_3	a_4	a_5	a_6	a_7
Value (mm)	4.44	2.55	2.32	2.70	0.81	0.45	2.10	0.34
Parameter	a_8	b_1	b_2	b_3	b_4			
Value(mm)	3.6	0.6	0.88	0.72	0.42			



▲**Figure 3. (a) Geometry of the back-to-back test of the simplified differential feeding network (L_x = 24.24 mm and L_y = 19.01 mm) and (b) fabricated prototype.**



▲**Figure 4. Simulated and measured S-parameters of the back-to-back test of the monopulse feeding network.**

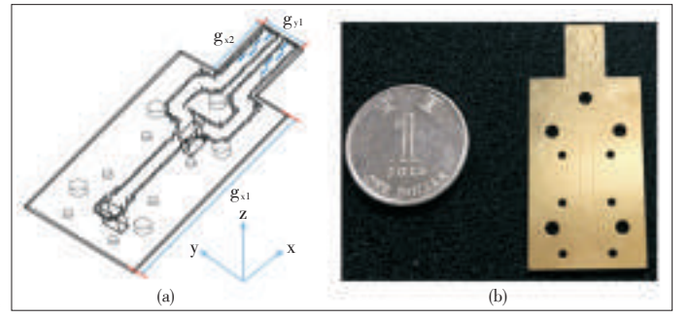
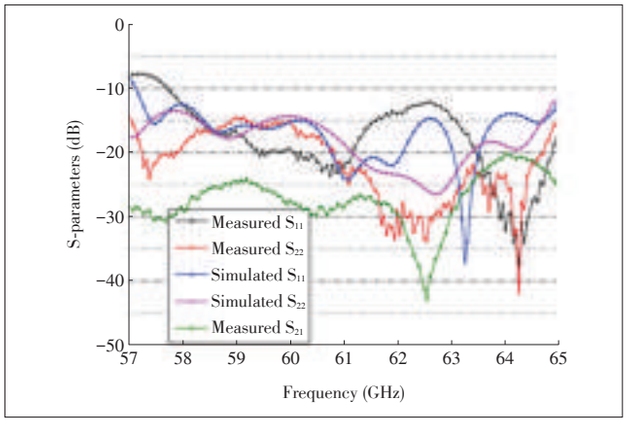
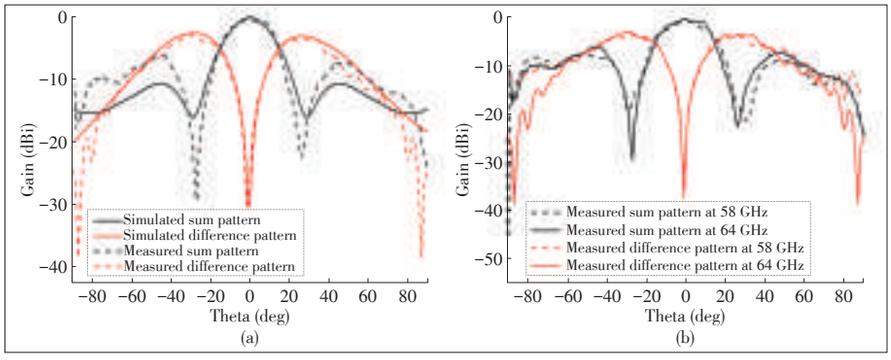


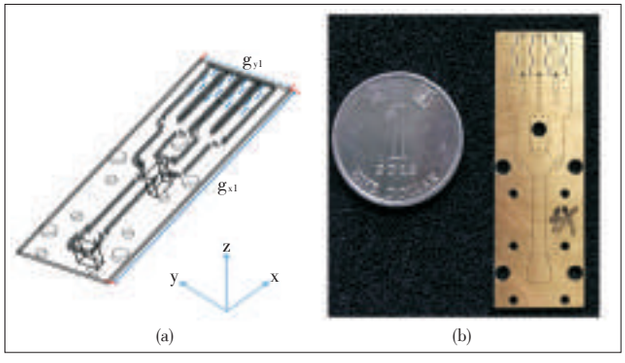
Figure 5. ▶ (a) Geometry of the 2×4 slot antenna array (g_{x1}=39.5 mm, g_{x2}=14.3 mm, g_{y1}=8.7 mm) and (b) the fabricated prototype.



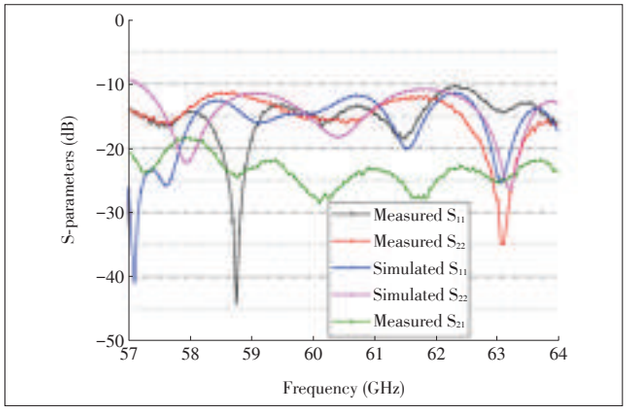
◀**Figure 6. Simulated and measured S-parameters of the 2×4 slot antenna array.**



▲**Figure 7. (a) Simulated and measured normalized radiation patterns at 60 GHz and (b) measured sum and difference radiation patterns at 58 GHz and 64 GHz.**



◀**Figure 8. (a) Geometry of the 4×4 slot antenna array (g_{x1}=52.4 mm, g_{y1}=17.7 mm) and (b) the fabricated prototype.**



◀**Figure 9. Simulated and measured S-parameters of the 4×4 slot antenna array.**

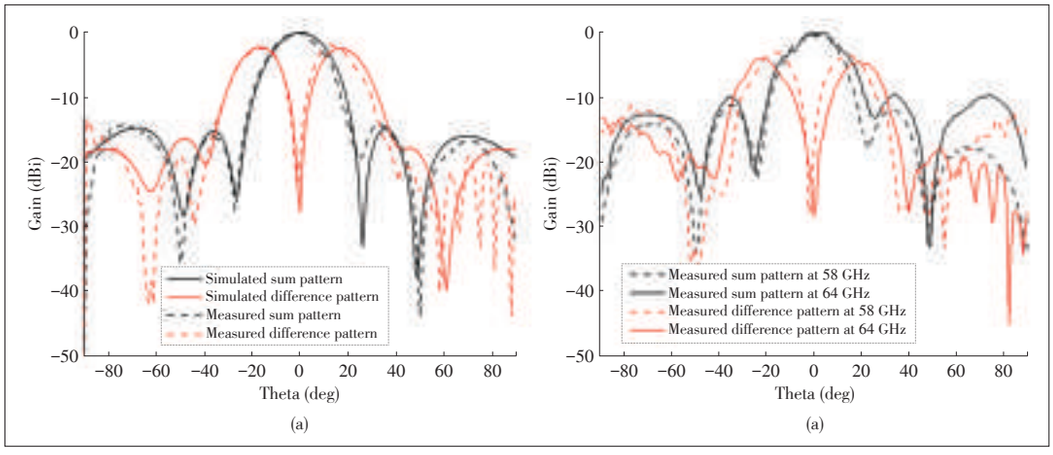
phase and amplitude imbalance may affect the radiation pattern and antenna gain. Therefore, the power dividers are designed with good impedance matching, high amplitude and phase balance as well as compact size. Simulated results show that from 57 GHz to 65 GHz the input reflection coefficient of the power divider is lower than −18 dB and the phase and amplitude imbalance of the output ports are less than 0.2 dB and 0.6°, respectively.

The simulated and measured reflection coefficients of the 4×4-element arrays are given in **Fig. 9**. The measured results show that the reflection coefficients of the arrays are lower than −10 dB in the entire unlicensed 60-GHz band and the isolation is better than 20 dB.

The 4×4 slot antenna array generates the difference beam in E-plane (*yoz* plane) and the sum beam in both E-plane and H-plane (*xoz* plane) when port 1 and port 2 are excited, respectively. The simulated and measured normalized radiation patterns at 60 GHz are shown in **Fig. 10a** and the measured sum and difference radiation patterns at 58 GHz and 64 GHz are shown in **Fig. 10b**. The null-depths are both below − 28 dB for the difference beam. The measured maximum gain is 16.24 dBi for the sum beam and 12.11 dBi for the difference beam while the measured minimum gain is 13.4 dBi and 9.89 dBi for the sum beam and difference beam, respectively.

3 Conclusions

In this paper, SIW-based monopulse slot antenna arrays are proposed for the 60 GHz applications. The monopulse comparator is realized by the combination of the H-plane and E-plane junctions in a single layer. This provides high amplitude and phase balance over a wide frequency band. Based on the monopulse comparator, the 2×4 and 4×4 element slot antenna arrays with compact size are designed, fabricated and measured. The simulated and measured results of the proposed antennas are in good agreement, indicating the antennas are good candidates for future millimeter-wave monopulse systems.



◀Figure 10.
(a) Simulated and measured sum and difference patterns at 60 GHz and (b) measured sum and difference patterns at 58 GHz and 64 GHz.

References

[1] S. M. Sherman and D. K. Barton, *Monopulse Principles and Techniques*. Norwood, USA: Artech House, 2011.

[2] C. E. Profera and L. H. Yorinks, "A high efficiency dual frequency multimode monopulse antenna feed system," *IEEE Transaction on Aerospace and Electronic Systems*, vol. AES - 2, no. 6, pp. 314– 322, Nov. 1966. doi: 10.1109/TAES.1966.4502022.

[3] K. M. Lee and R. S. Chu, "Design and analysis of a multimode feed horn for a monopulse feed," *IEEE Transaction on Antennas and Propagation*, vol. 36, pp. 171–181, Feb. 1988. doi: 10.1109/8.1094.

[4] H. Wang, D.-G. Fang, and X. G. Chen, "A compact single layer monopulse microstrip antenna array," *IEEE Transaction on Antennas and Propagation*, vol. 54, no. 2, pp. 503–509, Feb. 2006. doi:10/1109/TAP.2005.863103.

[5] Y. W. Wang, G. M. Wang, Z. W. Yu, J. G. Liang, and X. J. Gao,"Ultra-wideband e -plane monopulse antenna using vivaldi antenna," *IEEE Transaction on Antennas and Propagation*, vol. 62, no. 10, pp.4961–4969, Oct. 2014. doi:10.1109/TAP.2014.2342767.

[6] S.-G. Kim and K. Chang, "Low-cost monopulse antenna using bidirectionally-fed microstrip patch array," *Electronics Letters*, vol. 39, no. 20, pp. 1428–1429, Oct. 2003. doi: 10.1049/el:20030963.

[7] Z. W. Yu, G. M. Wang, and C. X. Zhang, "A broadband planar monopulse antenna array of c-band," *IEEE Antennas and Wireless Propagation Letters*, vol. 8, pp. 1325–1328, 2009. doi: 10.1109/LAWP.2009.2038077.

[8] Y. J. Cheng, W. Hong, and K. Wu, "94 GHz substrate integrated monopulse antenna array," *IEEE Transaction on Antennas and Propagation*, vol. 60, no. 1, pp. 121–129, Jan. 2012. doi: 10.1109/TAP.2011.2167945.

[9] B. Liu, W. Hong, Z. Kuai, et al., "Substrate integrated waveguide (SIW) monopulse slot antenna array," *IEEE Transaction on Antennas and Propagation*, vol. 57, no. 1, pp. 275–279, Jan. 2009. doi: 10.1109/TAP.2008.2009743.

[10] H. Chu, J. X. Chen, S. Luo, and Y. X. Guo, "A millimeter-wave filtering monopulse antenna array based on substrate integrated waveguide technology," *IEEE Transaction on Antennas and Propagation*, vol. 64, no. 1, pp. 316–321, Jan. 2016. doi: 10.1109/TAP.2015.2497351.

[11] K. Tekkouk, M. Ettorre, L. Le Coq, and R. Sauleau, "SIW pillbox antenna for monopulse radar applications," *IEEE Transaction on Antennas and Propagation*, vol. 63, no. 9, pp. 3918– 3927, Sept. 2015. doi: 10.1109/TAP.2015.2446996.

[12] A. S. Khan, *Microwave Engineering: Concepts and Fundamentals*. Boca Raton, USA: CRC Press, 2014.

[13] S. Liao, P. Wu, K. M. Shum, and Q. Xue, "Differentially fed planar aperture antenna with high gain and wide bandwidth for millimeterwave application," *IEEE Transaction on Antennas and Propagation*, vol. 63, no. 3, pp. 966–977, Mar. 2015. doi: 10.1109/TAP.2015.2389256.

[14] *HFSS Version 15.0.0*, Ansoft Corporation, 2012.

[15] S. Liao, P. Chen, P. Wu, and Q. Xue, "Substrate-integrated waveguide-based 60 -GHZ resonant slotted waveguide arrays with wide impedance bandwidth and high gain," *IEEE Transaction on Antennas and Propagation*, vol.63, no. 7, pp.2922–2931, 2015. doi: 10.1109/TAP.2015.2423696.

Manuscript received: 2016-06-15

Biographies

ZHU Jianfeng (zhujianfeng@bupt.edu.cn) received the BS degree in communication engineering from Beijing University of Posts and Telecommunications, China in 2013, where he is currently working toward the PhD degree. Since October 2015, he has been with City University of Hong Kong, China as a research assistant. His research interest is millimeter-wave antennas.

XUE Quan (eeqxue@cityu.edu.hk) received the BS, MS, and PhD degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), China, in 1988, 1990, and 1993, respectively. In 1993, he joined the UESTC, as a Lecturer. He became a Professor in 1997. From October 1997 to October 1998, he was a Research Associate and then a Research Fellow with the Chinese University of Hong Kong. In 1999, he joined the City University of Hong Kong where he is currently a Chair Professor of Microwave Engineering. He also serves the University as the Director of Information and Communication Technology Center (ICTC), the Deputy Director of CityU Shenzhen Research Institute, and the Deputy Director of State Key Lab of Millimeter Waves (Hong Kong). He was the Associate Vice President (Innovation Advancement and China Office) from June 2011 to January 2015. He has authored or co-authored over 260 internationally referred journal papers and over 100 international conference papers. His research interests include microwave passive components, active components, antenna, microwave monolithic integrated circuits (MMIC, and radio frequency integrated circuits (RFIC), etc. Professor XUE served the IEEE as an AdCom member of MTT-S from 2011-2013 and the associate editor of *IEEE Transactions on Microwave Theory* and Techniques (2010 - 2013), the associate editor of *IEEE Transactions on Industrial Electronics* (2010-present).

LIAO Shaowei (shaowei.s.liao@ieee.org) received the PhD in electromagnetic fields and microwave technology from UESTC, China in 2010. From October 2007 to September 2009, he was a research assistant in the Department of Electronics, Carleton University, Canada. From January 2011 to August 2011, he was with the School of Electronic Engineering, UESTC, as a lecturer. From September 2011 to July 2012, he served as a senior research associate in the Department of Electronic Engineering, City University of Hong Kong. From July 2012 to September 2013, he joined Bell Labs Research in China, Alcatel-Lucent Shanghai Bell, as a research scientist. Now, he is an engineer at State Key Laboratory of Millimeter Waves, City University of Hong Kong. His research interests include antennas (in particular, millimeter-wave antennas and GNSS antennas), electromagnetic simulation and computational electromagnetics.

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 002
http://www.cnki.net/kcms/detail/34.1294.TN.20161212.1535.002.html, published online December 12, 2016

Millimeter Wave and THz Propagation Channel Modeling for High-Data Rate Railway Connectivity—Status and Open Challenges

Thomas Kürner[1], GUAN Ke[2, 1], Andreas F. Molisch[3, 4], AI Bo[2], HE Ruisi[2], LI Guangkai[2], TIAN Li[5], DOU Jianwu[5], and ZHONG Zhangdui[2]

(1. Technische Universität Braunschweig, Braunschweig 38106, Germany;
2. Beijing Jiaotong University, Beijing 100044, China;
3. Cohere Technologies, Santa Clara CA 95051, USA;
4. University of Southern California, Los Angeles CA 90089, USA;
5. ZTE Corporation, Shanghai 201210, China)

Abstract

In the new era of railways, infrastructure, trains and travelers will be interconnected. In order to realize a seamless high-data rate wireless connectivity, up to dozens of GHz bandwidth is required. This motivates the exploration of the underutilized millimeter wave (mmWave) as well as the largely unexplored THz band. In this paper, we first identify relevant communication scenarios for railway applications. Then the specific challenges and estimates of the bandwidth requirements for high-data rate railway connectivity in these communication scenarios are described. Finally, we outline the major challenges on propagation channel modeling and provide a technical route for further studies.

Keywords

millimeter wave; radio channel; railway communications; THz communications

1 Introduction

In order to meet the goals with respect to efficiency, safety and convenience, rail traffic is expected to evolve into a new era where infrastructure, trains, travelers and goods will be increasingly interconnected [1]. To realize this vision (one part of the objective of "smart, green and integrated transport" supported by Horizon 2020 [1]), calls and initiatives, such as "Shift2Rail" [2], are inviting proposals concerning the following topics [3]: intelligent rail infrastructure [4], intelligent mobility management, smart rail services (seamless multimodal travel and logistic services), and a new generation of rail vehicles (trains with smart power and wireless technologies). All these specific topics ultimately impose requirements for a seamless high-data rate wireless connectivity in rail traffic. Correspondingly, railway communications are required to evolve from only the critical signaling applications, to various high-data rate applications, which need to be realized in five rail scenarios: train-to-infrastructure, inter-wagon, intra-wagon, inside station, and infrastructure-to-infrastructure. The huge bandwidth requirements—up to dozens of GHz—in these scenarios form a strong motivation for employing millimeter wave (mmWave) and THz communications, because they can offer orders of magnitude greater bandwidth than current spectrum allocations and enable very large antenna arrays which in turn provide high beamforming gains [5]. In order to effectively support the design, simulation, and development of the mmWave and THz communication systems, a thorough understanding of the propagation channel characteristics is critical.

This paper clarifies the bandwidth requirements of high-data rate railway connectivity by defining and analyzing the five communication scenarios in Section 2. Based on this, we identify the technical challenges and provide a technical route for further studies on mmWave and THz channels in Sections 3, 4

This work is supported by the NNSF of China under Grant 61501021, the ZTE Corporation, the Alexander von Humboldt Foundation, the Fundamental Research Funds for the Central Universities (No. 2014JBM075), State Key Lab of Rail Traffic Control and Safety Project under Grant RCS2016ZT021, Natural Science Base Research Plan in Shaanxi Province of China under Grant no. 2015JM6320, and Key Project from Beijing science and Technology Commission under Grant D151100000115004.

and 5. Conclusion and future work are drawn in Section 6.

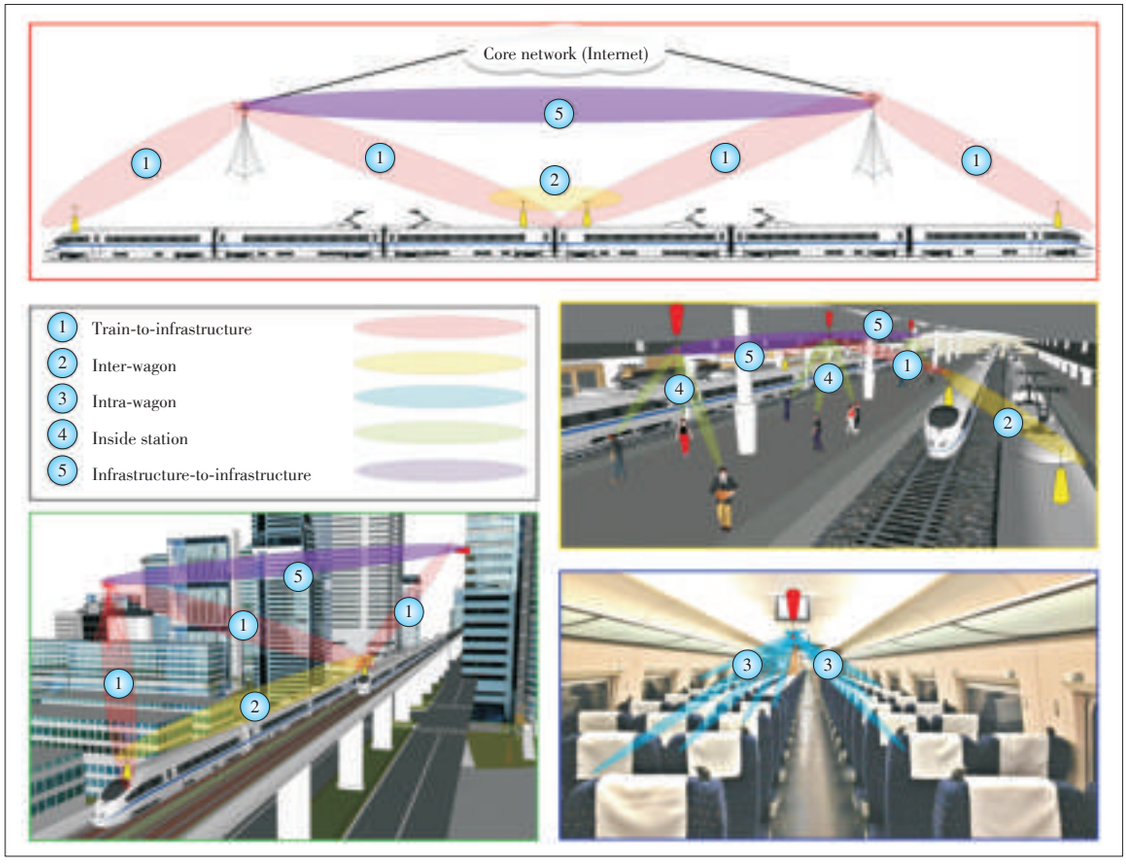
2 Bandwidth Requirements of High-Data Rate Railway Connectivity

In the future, railway communications are required to evolve to various high-data rate applications (e.g., [6], [7]): on-board and wayside high definition (HD) video surveillance that is critical for safety and security concerns (e.g., cars stuck on railway crossings, terrorist attacks, etc.); on-board real-time high-data rate connectivity for web browsing, video conferencing, video broadcast, etc.; train operation information that provides critical information regarding voice and control signaling, on-route train performance, and train equipment status; real-time train dispatching HD video between train and train control centers (TCCs) required for train dispatching and driverless systems; and journey information that dynamically updates journey information for all passengers via multimedia.

As shown in **Fig. 1**, the aforementioned applications can be realized in five communication scenarios [8]: train-to-infrastructure (T2I) (HD video and other information in real time transmitting among various infrastructures), inter-wagon (wireless network between wagons), intra-wagon (links between user equipment and access points of a wagon), inside station (links between access points (APs) and user equipment (UEs) in train/metro stations), and infrastructure-to-infrastructure (I2I) (HD video and other information in real time transmitting among various infrastructures).

The EU project METIS (which stands for Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society) [9] has identified four main challenges for 5G communications: very high data rate, very dense crowds of users, very low latency, and very high mobility. Since each of the mentioned communication scenarios in rail traffic meets the challenges similar to one or more test cases (TCs) defined in METIS, the corresponding TCs in METIS are cited here when analyzing the bandwidth requirements:

1) Train-to-infrastructure, corresponding to the combination of "TC12: traffic efficiency and safety", "TC8: real-time remote computing for mobile terminals" and "TC1: virtual reality office" in METIS, describes links between the infrastructures and the APs/transceivers of the wireless local network for the train. It requires bi-directional streams with very high data rates and low latencies (millisecond level), as well as robust links with low latencies together with an availability that is close to 100% while moving at a speed up to 500 km/h [10].
2) Inter-wagon requires a high-data rate and a low latency because the APs are arranged in every wagon and each AP serves as a client station for the APs in the other wagons



◀Figure 1. Five communication scenarios of high-data rate railway connectivity.

while also serving as the AP for all the stations within its wagon [11]. Therefore, this scenario requires two or more times (depending on how many wagons are connected) of bandwidth of the "intra-wagon" scenario.

3) Intra-wagon, corresponding to the "TC1: virtual reality office" in METIS, requires capability for real-time HD videos with low latencies. Considering using a 40 MHz channel to support compressed HD video and assuming 50% of the 180 passengers of a double-decker wagon want to use high-speed video, a total of 3.6 GHz bandwidth will be required for one wagon.
4) Inside station corresponds to "TC3: shopping mall" in METIS. In a shopping mall, users are strongly interested to get access to mobile broadband applications (e.g., 1 Gbps, which is expected to be supported by IMT-2020 for indoor users [10]), and the station will provide a fixed/wireless communication infrastructure to support general commercial as well as operational applications [9].
5) Infrastructure-to-infrastructure, corresponding to the "TC1: virtual reality office" in METIS, describes HD video and other information in real-time interaction among multiple cameras and APs, e.g., a high-date-rate wireless backhaul, supported by bi-directional streams with very high data rates and low latencies [9].

To sum up, for the scenarios "inside station" and "infrastructure-to-infrastructure", it is convenient to roughly estimate their bandwidth requirements (**Table 1**) by referring to their corresponding TCs in METIS. With the evaluation procedure already described in [12], the bandwidth requirements are from several hundred MHz to several GHz, depending on concrete conditions. For the scenarios "intra-wagon" and "inter-wagon", up to 3.6 GHz and up to dozens of GHz bandwidths will be required, respectively. In the scenario "train-to-infrastructure", the main interface between the network on a train and the fixed network transmits an aggregated stream of the inter/intra-wagon scenarios. Therefore, it requires the bandwidth from 7.2 GHz to dozens of GHz to realize over 100 Gbps transmission. Table 1 summarizes the communication scenarios, corresponding TCs in METIS, challenges (with detailed data from IMT 2020 defined by ITU-R [13]), and bandwidth requirements. Obviously, such high data rate and huge bandwidth requirements motivate the exploration of the underutilized mmWave and THz bands. Systems operated at these frequencies are referred as mmWave and THz communication systems.

3 Millimeter Wave and THz Communication Enabling High-Date Rate Railway Connectivity

Various technologies working at frequencies lower than 6 GHz, such as Long-Term Evolution Advanced (LTE-A) [14], WiMax [15] and LTE for Railway (LTE-R) [16], have been presented to realize the broadband wireless access in rail traffic. However, these existing technologies support data rates from several Mbps up to 100 Mbps, which are still up to three orders of magnitude lower than the desired throughput.

▼**Table 1. High-data rate railway connectivity scenarios, corresponding TCs in METIS, challenges, and bandwidth requirements**

Scenarios		Train-to-infrastructure	Inter-wagon	Intra-wagon	Inside station	Infrastructure-to-infrastructure
Corresponding TCs in METIS		TC12: traffic efficiency and safety; TC8: real-time remote computing for mobile terminals; TC1: virtual reality office		TC1: virtual reality office	TC3: shopping mall	TC1: virtual reality office
Applications	On-board and wayside HD video surveillance	√				√
	Train operation information	√	√	√		√
	Real-time train dispatching HD video	√			√	√
	On-board real-time high-data rate connectivity	√	√	√		
	Journey information	√	√	√	√	√
Challenges (detailed data from [13])	Very high data rate (peak data rate up to 20 Gbit/s and user experienced data rate up to 100 Mbit/s)	√	√	√		√
	Very dense crowds of users (up to 10^6 devices/km^2)		√	√	√	
	Very low latency (down to 1 ms)	√	√	√	√	√
	Very high mobility (up to 500 km/h)	√				
	Bandwidth requirements	7.2 GHz to dozens of GHz	Up to dozens of GHz	Up to 3.6 GHz	Up to several GHz	Up to several GHz

HD: high definition METIS: Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society TC: test case

MmWave communication for railway is not completely new but rather a long effort reaching back as far as 32 years ago (first literature in 1983 [17]). However, in the era of single-input and single-output (SISO) systems, the link distance and the mobility of the mmWave users were very limited. Nowadays, outstanding progress has been made toward the development of compact mmWave band transceivers providing high transmission power, high detection sensitivity, and low noise figures [18], [19]. Moreover, ultra-wideband (UWB) and multiple-input and multiple-output (MIMO) antennas have been designed for mmWave bands. It thus seems an opportune time to recall and initiate the mmWave communications enabling high-date rate railway connectivity. Today even frequencies as high as 300 GHz are considered for wireless communications as the so-called THz communications [20]–[23]. Recent advances in semiconductor technology have triggered even standardization activities. In IEEE 802.15, the first standard for fixed point-to-point links at 300 GHz is under development [24]. At the World Radio Conference in 2019 (WRC'19), spectrum allocations beyond 275 GHz for fixed and mobile applications will be investigated under agenda item 1.15 [25].

4 Related Work and Open Challenges

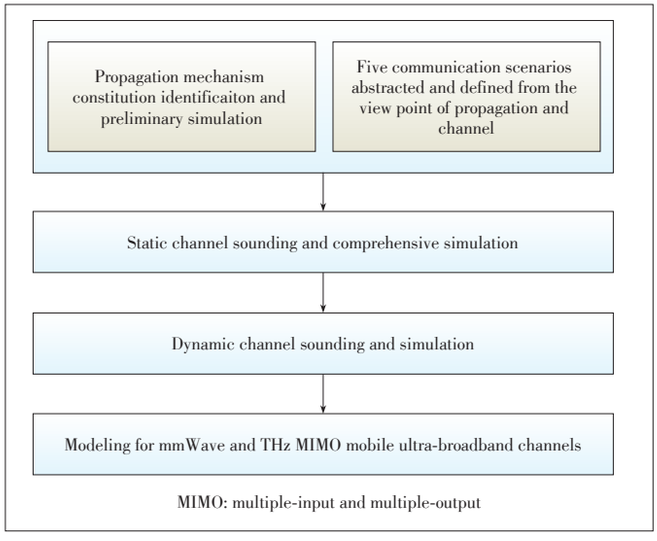
Whenever a wireless communication system is applied the first time in either a new environment or in a new frequency band, the propagation channel is subject to detailed investigations. This holds also for mmWave and THz communication systems in general and for the application of these systems in railway environments specifically. In the following the main challenges are described.

1) Wave propagation mechanisms: The synergism of susceptibility of molecular absorption, the changed relationships between wavelengths and dimensions of objects, and the ultra-broadband bandwidths, makes propagation in the mmWave band and THz distinct from microwave frequencies. Even though the main propagation mechanisms have been identified to some extent [26]–[28], more research efforts should be made on interpretation of the complex propagation phenomena, such as frequency-selective and distance-dependent behaviors, frequency dispersion, different shadowing effects, taking into account the main objects and geometries in railway environments.
2) Characterization of static channel: For indoor environments, MIMO mmWave channels were characterized in a range of environments [29]. The IEEE 802.11ad and 802.15 TG3c models were established for 60 GHz indoor communications [30]. The TG3c model covers the residential, office, library, desktop, and kiosk environments, whereas the TGad model covers the conference room, living room, and cubicle. A thorough review of mmWave propagation both indoor and outdoor can be found in [31]. For outdoor environments, there have been recent studies regarding the outdoor channel propagation characteristics that have shown the potential for utilizing the mmWave band for cellular communications [32], [33]. A detailed literature review is provided in [31] and [34]. First investigations on even more complex scenarios beyond 300 GHz have been published, focusing on static channels including both deterministic and stochastic channel models [19], [35]. A compilation of channel models for fixed-point-to-point applications at 300 GHz for various environments covering outdoor, indoor and even intra-device environments can be found in [36]. However, even though ITU-R recommends using the wide spectrum in above 6 GHz for supporting moving hotspot cell users such as high speed train moving 500 km/h, channel measurements in rail communication scenarios have been rarely reported in the mmWave and not at all in THz frequencies. Usually, railway appears as a small use case of "moving hotspot" in standard documents [37], without parameterized definition for details. How to include the railway channel features in even static channel models is an open challenge.
3) Dynamic channel modeling: Most of the existing measurements in mmWave band and all measurements in the THz band were done for static channels. In 2014, Samsung presented the world's first demonstration for 1.2 Gbps transmission at 28 GHz and 110 km/h cruising speed using dual-beamforming [38]. This evidence shows the strong potential of the mmWave band for mobile applications. As pointed out in [28] and [39], dynamic ray shadowing causes a temporal variation of the path losses. Reflections at persons may additionally be subject to Doppler shifts and entire Doppler spectra may result. The same holds if the transceiver units move relative to each other. The lack of insight into such dynamics inhibits the realization of the mmWave communications in any dynamic environment, particularly the five rail traffic scenarios.

5 Technical Route for Further Studies

In order to address the above open challenges, a technical route for further studies can be formed, as shown in **Fig. 2**.

To begin with, the propagation characteristics of the main objects in the five scenarios with various geometrical and physical configurations can be measured in the mmWave and THz bands using a Vector Network Analyzer (VNA) or even Time-Domain Spectroscopy (TDS). The frequency-dependent coefficients of every propagation mechanism can be derived empirically from the measurements or semi-empirically by considering some theoretical modeling as well. After evaluating the influence of various objects, the most significant objects should be modeled in detail to build the ray optical scenarios enabling the simulation of static channels. The propagation mechanism constitutions in every scenario can be quantitatively identified, which guides the design of measurement campaigns for more complex static scenarios. Then, as given by **Table 2**, the five



▲**Figure 2. Technical route for further studies on mmWave channel for high-data rate railway connectivity.**

scenarios can be defined from the viewpoint of propagation and channel. With various combinations of factors, such as link lengths, line of sight (LOS)/non-LOS (NLOS), and MIMO antennas, static measurements can be conducted in the rail scenarios. More detailed data can be found in [40], where six scenario modules for mmWave and THz train-to-infrastructure channels are defined and constructed for the first time. All the main objects, such as tracks, stations, crossing bridges, tunnels, cuttings, barriers, pylons, buildings, vegetation, traffic signs, billboards, and trains, are modeled according to the typical geometries and materials in reality. After calibration by the measurements, extensive comprehensive simulations can run to catch the frequency-dispersion effect of the ultra-broadband mmWave and THz channels in frequency, time, angular and polarization domains.

Afterwards, dynamic channel measurements can be designed and performed with various configurations, such as different moving transceivers and scatterers with variations of speed and motion direction. Similarly, dynamic simulations can be made by adding the same mobility to the corresponding scenarios. To perform measurements in dynamic scenarios a channel sounder (CS) operating in time-domain is necessary. Such measurements have been reported already at 60 GHz [41]–[45] but not at 300 GHz. An especially promising approach for future measurements is channel sounding employing orthogonal time frequency space (OTFS) waveforms, which are well suited for data transmission in quickly changing environments, and also provide a natural basis for identifying objects in the delay-Doppler domain [46], [47]. Based on both the measurement and simulation results, the correlations between the mobility and channel dynamics can be revealed. For instance, it is critical to evaluate the influence of the mobility and the variation of mobility on the first- and second-order statistics of the channel dynamic parameters.

Finally, based on extensive measurements and simulations done in the previous steps, all the static and dynamic channel parameters in full-dimensions can be extracted and modeled for channel realizations. This stochastic spatio-temporal model can provide a basis of the system design by generating the channels in the five rail traffic scenarios, without the need for in-depth understanding of mmWave propagation or ray tracing tools.

A general challenge for measurements using either a VNA or a CS is the limited distances (up to a few 10s of meters) between the transmitting and receiving units of the measurement system due to the requirement to connect the transmitter and receiver with cables to a central unit. This will make it especially challenging to perform measurements for the train-to-infrastructure and infrastructure-to-infrastructure scenarios for which specific measurement equipment has to be developed. As a "DFG State Funded Major Instrumentation" funded by the German federal government and the federal state of Lower Saxony, one custom-made mmWave and THz MIMO channel sounder was delivered to Technical University of Braunschweig in August 2016. This sounder supports 8 GHz bandwidth and uses high-gain (15 dBi–25 dBi) antennas. It sends a pseudo random sequence repeatedly with a very fast rate to realize fast time-domain measurements for mobile channel links up to

▼**Table 2. Scenario definition from the viewpoint of propagation and channel**

	Scenarios	Train-to-infrastructure	Inter-wagon	Intra-wagon	Inside station	Infrastructure-to-infrastructure
Setup	Indoor/outdoor	Outdoor	Outdoor	Indoor	Indoor	Outdoor
	Fix/mobile link	Mobile link	Mobile link	Mobile link	Mobile link	Quasi-fixed link
	Velocity of user	High, up to 500 km/h	Light, fluctuation of wagons	Low, pedestrian, ca. 1 m/s	Low, pedestrian, ca. 1 m/s	Light, antenna mispointing
	LOS/NLOS	LOS	LOS	LOS and NLOS	LOS and NLOS	LOS
	Weather	Dry and wet	Dry and wet	-	-	Dry and wet
Common	Scatters	Mobile scatters: pedestrians, passengers, moving cars, wind blades, and so on				
	MIMO antenna setup	From 2 × 2 to practical massive MIMO, setup linear/rectangular/cylindical array, with vertical/horizontal polarization				
	Frequency band	MmWave band, from 100 MHz to dozens of GHz bandwidth				

LOS: line of sight MIMO: multiple-input and multiple-output NLOS: non-LOS

several tens of meters for the 60 GHz and 300 GHz bands. Using correlation techniques, the receiver can extract the channel impulse response at a rate of more than 14,000 CIRs/s. These features enable the investigation on the channel dynamics with mobile transceivers and scatterers in the mmWave and THz bands.

6 Conclusions and Future Work

In this paper, we provide elementary discussions on bandwidth requirements of high-data rate railway connectivity, and highlight the open challenges in terms of wave propagation, static channel, and dynamic channel. More research efforts are expected to reveal the essence of complex propagation phenomena, such as frequency-selective and distance-dependent behaviors, frequency dispersion, different shadowing effects, for several GHz of bandwidths in the mmWave and THz bands, taking into account the main objects and geometries in rail scenarios. MmWave and THz channel models including railway features are still open issues, and therefore, a technical route for further studies on mmWave and THz propagation channel for high-date rate railway connectivity should be defined.

References

[1] European Commission. (2014, Jul.). HORIZON 2020 work programme 2014–2015: Smart, green and integrated transport revised. [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-transport_en.pdf
[2] European Commission. (2014). Shift2rail: driving innovation on railways. [Online]. Available: http://www.shift2rail.org
[3] A. Damiani. (2013). Transport research and innovation in Horizon 2020. [Online]. Available: https://unige.it/ricerca/news/documents/Damiani_Transport_Info_Day.pdf
[4] J. Josey. (2013, Aug.). Intelligent infrastructure for next-generation rail system. *Cognizant 2020 Insights*. [Online]. 35 (3), pp. 1–8. Available: http://www.cognizant.ch/InsightsWhitepapers/Intelligent-Infrastructure-for-Next-Generation-Rail-Systems.pdf
[5] M. R. Akdeniz, Y. Liu, M. K. Samim, et al., "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164– 1179, Jun. 2014. doi: 10.1109/JSAC.2014.2328154.
[6] J. Moreno, J. Riera, L. De Haro, and C. Rodriguez, "A survey on future railway radio communications services: challenges and opportunities," *IEEE Communications Magazine*, vol. 53, no. 10, pp. 62– 68, Oct. 2015. doi: 10.1109/MCOM.2015.7295465.
[7] B. Ai, K. Guan, M. Rupp, et al., "Future railway services-oriented mobile communications network," *IEEE Communications Magazine*, vol. 53, no. 10, pp. 78–85, Oct. 2015. doi: 10.1109/MCOM.2015.7295467.
[8] K. Guan, G. Li, T. Kürner, et al., "On millimeter wave and THz mobile radio channel for smart rail mobility," *IEEE Transactions on Vehicular Technology*, to appear, 2016.
[9] METIS. (2013, Apr.). Scenarios, requirements and KPIs for 5G mobile and wireless system. [Online]. Available: http://cordis.europa.eu/docs/projects/cnect/9/317669/080/deliverables/001-METISD11v1pdf.pdf
[10] ITU. (2015). IMT vision—framework and overall objectives of the future development of IMT for 2020 and beyond. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf
[11] D. Fokum and V. Frost, "A survey on methods for broadband internet access on trains," *IEEE Communications Surveys Tutorials*, vol. 12, no. 2, pp. 171–185, Second 2010.
[12] METIS. Description of the spectrum needs and usage principles. [Online]. Available: https://www.metis2020.com/wp - content/uploads/deliverables/METIS_D5.3_v1.pdf
[13] *IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, ITU-R M.2083-0, Sept. 2015.
[14] M.-C. Chuang and M. C. Chen, "A mobile proxy architecture for video services over high-speed rail environments in LTE-A networks," *IEEE Systems Journal*, vol. 9, no. 4, pp. 1264–1272, Dec. 2015. doi: 10.1109/JSYST.2014.2354435.
[15] M. Aguado, O. Onandi, P. Agustin, M. Higuero, and E. Jacob Taquet, "WiMax on rails," *IEEE Vehicular Technology Magazine*, vol. 3, no. 3, pp. 47–56, Sept. 2008.
[16] K. Guan, Z. Zhong, and B. Ai, "Assessment of LTE-R using high speed railway channel model," in *Proc. 3rd International Conference on Communications and Mobile Computing*, Qingdao, China, 2011, pp. 461– 464. doi: 10.1109/CMC.2011.34.
[17] H. Meinel and A. Plattner, "Millimeter-wave propagation along railway lines," *IEE Proceedings*, pp. 688–694, 1983.
[18] I. Akyildiz, J. Jornet, and C. Han, "Teranets: ultra-broadband communication networks in the Terahertz band," *IEEE Wireless Communications*, vol. 21, no. 4, pp. 130–135, Aug. 2014. doi: 10.1109/MWC.2014.6882305.
[19] E. Ojefors, J. Grzyb, Y. Zhao, et al., "A 820 GHz SiGe chipset for Terahertz active imaging applications," in *2011 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, USA, Feb. 2011, pp. 224– 226. doi: 10.1109/ISSCC.2011.5746294.
[20] J. Federici and L. Möller, "Review of terahertz and subterahertz wireless communications," *Journal of Applied Physics*, vol. 107, no. 11, pp. 1–22, 2010.
[21] T. Kleine-Ostmann and T. Nagatsuma, "A review on terahertz communications research," *Journal of Infrared, Millimeter and Terahertz Waves*, vol. 32, no. 2, pp. 143–171, 2011.
[22] H.-J. Song, S. Priebe, and T. Kürner, "Terahertz wireless communications," in *Handbook of Terahertz Technologies: Devices and Applications*, H.-J. Song and T. Nagatsuma, Ed. Boca Raton, USA:CRC Press, 2015, pp. 495–526.
[23] I. Kallfass, I. Dan, S. Rey, et al., "Towards MMIC-based 300 GHz indoor wireless communication systems," *IEICE Transactions on Electronics*, vol. E98-C, no. 12, pp. 1081–1090, 2015.
[24] IEEE. (2016). IEEE 802.15 working group for WPAN. [Online]. Available: http://www.ieee802.org/15/
[25] ITU, "Studies towards an identification for use by administrations for land mobile and fixed services applications operating in the frequency range 275-450 GHz," *The World Radiocommunication Conference*, Geneva, Switzerland, Tech. Rep., 2015.
[26] J. Jornet and I. Akyildiz, "Channel modeling and capacity analysis for electromagnetic wireless nanonetworks in the Terahertz band," *IEEE Transactions on Wireless Communications*, vol. 10, no. 10, pp. 3211–3221, Oct. 2011.
[27] C. Jansen, S. Priebe, C. Möller, et al., "Diffuse scattering from rough surfaces in THz communication channels," *IEEE Transactions on Terahertz Science and Technology*, vol. 1, no. 2, pp. 462–472, Nov. 2011.
[28] M. Jacob, S. Priebe, R. Dickhoff, et al., "Diffraction in mm and sub-mmWave indoor propagation channels," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 3, pp. 833– 844, Mar. 2012. doi: 10.1109/TMTT.2011.2178859.
[29] C. Gustafson, K. Haneda, S. Wyne, and F. Tufvesson, "On mm-wave multipath clustering and channel modeling," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 3, pp. 1445–1455, Mar. 2014.
[30] H. Harada and H. Sawada, "Millimeter-wave channel modelling for IEEE802 standardizations," in *8th European Conference on Antennas and Propagation (EuCAP)*, The Hague, The Netherlands, Apr. 2014, pp. 644–645. doi: 10.1109/EuCAP.2014.6901841.
[31] K. Haneda, "Channel models and beamforming at millimeter-wave frequency bands," *IEICE Transactions*, vol. 98-B, no. 5, pp. 755–772, 2015.
[32] T. Rappaport, S. Sun, R. Mayzus, et al., "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.
[33] T. Rappaport, F. Gutierrez, E. Ben-Dor, et al., "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.
[34] A. F. Molisch, A. Karttunen, R. Wang, et al., "Millimeter-wave channels in urban environments," in *Proc. European Conference on Antennas and Propagation (EUCAP)*, Davos, Switzerland, Apr. 2016. doi: 10.1109/EuCAP.2016.7481098.
[35] S. Priebe, M. Kannicht, M. Jacob, and T. Kürner, "Ultra broadband indoor channel measurements and calibrated ray tracing propagation modeling at THz frequencies," *Journal of Communications and Networks*, vol. 15, no. 6, pp. 547–558, Dec. 2013. doi: 10.1109/JCN.2013.000103.
[36] IEEE. (2016). IEEE P802.15 TG3d channel modeling document. [Online].

Available: https://mentor.ieee.org/802.15/dcn/14/15-14-0310-18-003d-channelmodeling- document.docx
[37] ITU, "Technical feasibility of IMT in bands above 6 GHz," Tech. Rep. ITU-R M.2376-0, 2015.
[38] Samsung. (2014). Samsung electronics sets 5G speed record at 7.5Gbps, over 30 times faster than 4G LTE. [Online]. Available: http://global.samsungtomorrow.com/?p=43349
[39] S. Collonge, G. Zaharia, and G. E. Zein, "Influence of the human activity on wide-band characteristics of the 60 GHz indoor radio channel," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 2396–2406, Nov 2004.
[40] K. Guan, X. Lin, D.P. He, et al., "Scenario modules and ray-tracing simulations of millimeter wave and terahertz channels for smart rail mobility," to appear, *11th European Conference on Antennas and Propagation (EuCAP2017)*, Paris, France, Apr. 2017.
[41] T. Zwick, T. J. Beukema, and H. Nam, "Wideband channel sounder with measurements and model for the 60 GHz indoor radio channel," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 4, pp. 1266–1277, Jul. 2005.
[42] J. Kivinen, "60-GHz wideband radio channel sounder," *IEEE Transactions on Instrumentation and Measurement*, vol. 56, no. 5, pp. 1831–1838, Oct. 2007.
[43] M. Peter, M. Wisotzki, M. Raceala-Motoc, et al., "Analyzing human body shadowing at 60 GHz: Systematic wideband MIMO measurements and modeling approaches," in *6th European Conference on Antennas and Propagation (EuCAP)*, Prague, Czech Republic, Mar. 2012, pp. 468–472.
[44] A. P. G. Ariza, W. Kotterman, R. Zetik, et al., "60 GHz-ultrawideband real-time multi-antenna channel sounding for multi giga-bit/s access," in *2010 IEEE 72nd Vehicular Technology Conference Fall (VTC 2010-Fall)*, Ottawa, Canada, Sept. 2010, pp. 1–6.
[45] A. P. Garcia, W. Kotterman, U. Trautwein, et al., "60 GHz time-variant shadowing characterization within an airbus 340," in *Proc. Fourth European Conference on Antennas and Propagation (EuCAP)*, Barcelona, Spain, Apr. 2010, pp. 1–5.
[46] 3GPP, "OTFS modulation waveform and reference signals for new RAT," Cohere, AT&T, CMCC, Deutsche Telekom, Telefonica, Telstra, Tech. Rep. 3GPP R1-163619, 2016.
[47] A. Monk, R. Hadani, M. Tsatsanis, and S. Rakib. (2016, Aug.). OTFS—Orthogonal Time Frequency Space. [Online]. Available: https://arxiv.org/abs/1608.02993v1

Manuscript received: 2016-06-15

Biographies

Thomas Kürner (Kuerner@ifn.ing.tu-bs.de) received his Dipl-Ing degree in electrical engineering in 1990, and his Dr-Ing degree in 1993, both from Universitaet Karlsruhe, Germany. From 1994 to 2003, he was with the Radio Network Planning Department at the headquarters of the GSM 1800 and UMTS operator E-Plus Mobilfunk GmbH & Co KG, Düsseldorf, where he was finally the team manager of radio network planning support. Since 2003 he is professor for mobile radio systems at the TU Braunschweig. His working areas are propagation, self-organization of cellular networks, as well as indoor channel characterization for high-speed short-range systems including future terahertz communication systems and accuracy of satellite navigation systems. Currently he chairs the IEEE 802.15 Task Group 3d and the WG Propagation of the European Association on Antennas and Propagation. Since 2008 he is an associate editor of *IEEE Transactions on Vehicular Technology*.

GUAN Ke (kguan@bjtu.edu.cn) received his BE and PhD degrees from Beijing Jiaotong University, China in 2006 and 2014, respectively. He is an associate professor in State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. In 2015, he was awarded a Humboldt Research Fellowship. He was the recipient of a 2014 International Union of Radio Science (URSI) Young Scientist Award. His paper received the honorable mention in the third International URSI student prize paper competition in 2014 URSI GASS. His current research interests are in the field of channel characterization for future millimeter wave and terahertz communication systems. He has authored or co-authored over 100 research papers in international journals and conferences.

Andreas F. Molisch (andreas.molisch@coheretechnologies.com) is a professor of electrical engineering at The University of Southern California, USA. Previously he was at TU Vienna, FTW, AT&T (Bell) Labs, Lund University, and Mitsubishi Electric Research Labs. His research interest is wireless communications, with emphasis on wireless propagation channels, multi-antenna systems, ultra-wideband signaling and localization, novel cellular architectures, and cooperative communications. He is the author of four books, 18 book chapters, more than 470 journal and conference papers, as well as 80 patents. He is a fellow of the National Academy of Inventors, IEEE, AAAS, and IET, as well as member of the Austrian Academy of Sciences and recipient of numerous awards.

AI Bo (boai@bjtu.edu.cn) received his Master and PhD degrees from Xidian University, China in 2002 and 2004 in China, respectively. He is now working with Beijing Jiaotong University as a professor and advisor of Ph.D. candidates. He is a deputy director of State Key Lab of Rail Traffic Control and Safety of the university. He has authored or co-authored 6 books, 140 scientific research papers and 26 invention patents in his research area till now. His current interests are the research and applications OFDM techniques, HPA linearization techniques, radio propagation and channel modeling, and LTE for railway systems. He is an IET fellow and an IEEE senior member.

HE Ruisi (ruisi.he@bjtu.edu.cn) received the BE and PhD degrees from Beijing Jiaotong University, China, in 2009 and 2015, respectively. He was a visiting scholar with Universidad Politécnica de Madrid, Spain, The University of Southern California, USA, and Université catholique de Louvain, Belgium, from 2010 to 2014. He has been an associate professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, since 2015. He has authored or co-authored over 80 research papers in international journals and conferences. His current research interests include measurement and modeling of wireless propagation channels and vehicular and high-speed railway communications.

LI Guangkai (gkli.rcs@bjtu.edu.cn) is a PhD student at State Key Laboratory of Rail Traffic Control and Safety in Beijing Jiaotong University, China. His research focuses on optimizing ray tracing algorithms based on measurements and deterministic modeling for mmWave and terahertz channel in dynamic scenarios.

TIAN Li (tian.li150@zte.com.cn) received the bachelor degree in communication engineering and the PhD degree in control science and control engineering from Tongji University, China, in 2009 and 2015, respectively. From 2013 to 2014, he was a visiting PhD student at the Department of Electronics and Information Systems (DEIS), University of Bologna, Italy working with Prof. Vittorio Degli-Esposti. He participated in the 5G project sponsored by National Natural Science Foundation of China. He is now a senior engineer at the Department of Algorithms, ZTE Corporation. His current research interests are in the field of 5G channel modeling and new air-interface. He has authored or co-authored over 30 scientific research papers.

DOU Jianwu (dou.jianwu@zte.com.cn) received the PhD degree in robotic mechanism from Beijing University of Technology, China in 2001. He is currently in charge of a National Science and Technology Major Project of the Ministry of Science and Technology of China and participated in the 5G project sponsored by the National Natural Science Foundation of China. He is the vice director of the Department of Algorithms in ZTE Corporation. His current research interests are in the field of 5G channel modeling, new air-interface, and high-layer design. Dr. Dou received the Science and Technology Awards (First Level) from the China Institute of Communications in 2014 and 2015. In 2011, he received the WIPO-SIPO Award for Chinese Outstanding Patented Invention.

ZHONG Zhangdui (zhdzhong@bjtu.edu.cn) is a professor and advisor of PhD candidates in Beijing Jiaotong University. He is a chief scientist of State Key Laboratory of Rail Traffic Control and Safety in Beijing Jiaotong University. He is also a director of the Innovative Research Team of Ministry of Education, China. His interests are wireless communications for railways. His research has been widely used in Qinghai-Xizang railway, Datong-Qinhuangdao Heavy Haul railway, and many high-speed railway lines of China. He has authored or co-authored 7 books, 5 invention patents, and over 200 scientific research papers in his research area. He received MaoYiSheng Scientific Award of China, ZhanTianYou Railway Honorary Award of China, and Top 10 Science/Technology Achievements Award of Chinese Universities.

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 003
http://www.cnki.net/kcms/detail/34.1294.TN.20161226.1558.002.html, published online December 26, 2016

State of the Art in Passive Bandpass Filter Solutions for 60 GHz Communications

XU Shanshan[1], MENG Fanyi[2], MA Kaixue[3], and YEO Kiat Seng[1]
(1. Singapore University of Technology and Design, Singapore 487372, Singapore;
2. Nanyang Technological University, Singapore 639798, Singaproe;
3. University of Electronic Science and Technology, Chengdu 610051, China)

Abstract

This paper reviews the state-of-the-art filter designs for 60 GHz applications. The most promising filter solutions at this frequency include filter-in-package where the filter itself is design in the packaging platform and filter-on-chip which is an on-chip filter co-design for miniaturized system size with low packaging cost. Design methodology, design technology, key performance parameters, similarities and differences, advantages and drawbacks, and future trends are explored and studied. Filters in the printed circuit board (PCB), low temperature co-fired ceramics (LTCC), organic material, and bipolar complementary metal oxide semiconductor (BiCMOS) chips are summarized and compared in details. Future design trends and challenges are also given after the review.

Keywords

60 GHz; bandpass filters; passive; review

1 Introduction

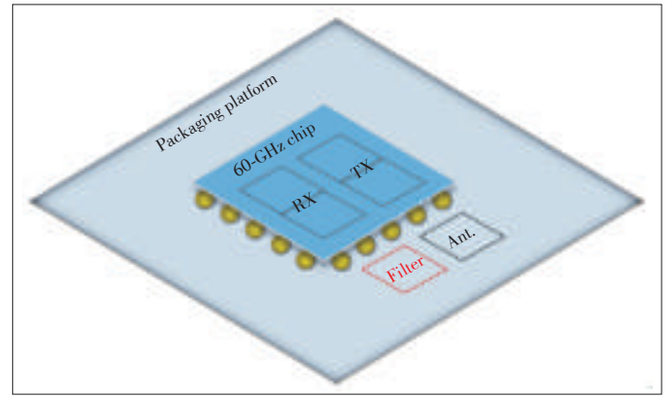
The unlicensed 60 GHz band is a promising solution to future Gb/s wireless communication. It has been a hot research topic in the past decade. To achieve high performance systems, the passive bandpass filter (BPF) plays a critical role in rejecting unwanted interferences or harmonics. For 60 GHz to enter commercial market where the cost and size are mostly concerned, the filters are required to achieve low insertion loss, feasibility of integration, compact size, small form-factor, and mechanical flexibility for bendable electronics.

Researchers have demonstrated many successful implementations of 60 GHz BPFs [1]–[12]. Many technologies have been explored, e.g. the printed circuit board (PCB) technology and low temperature co-fired ceramics (LTCC) technology [1]–[3], targeting for different application scenarios. Some filters have been developed using organic material, which can be bended without sacrificing much of the performance [4]. These filters belong to in-package filters that sit inside the packaging platform. Unlike in-package ones, filters on chip offer a higher level of integration [6]–[12]. However, their performance has to be degraded due to the thin oxide layer in bipolar complementary metal oxide semiconductor (BiCMOS) technologies.

In this paper, we provide a comprehensive overview of filter design progress for 60 GHz applications. With the compiled data and design method, researchers could quickly understand this research area and possibly identify research directions. The results are also tabulated for a clear comparison that serves as benchmark data.

2 Filter-in-Package Solutions

Fig. 1 shows an example of a flip-chip packaged 60 GHz chip with filter-in-package solution. Like other conventional approaches for low RF or a few GHz systems, the in-package filter has been developed for nearly a decade for 60 GHz com-



▲**Figure 1. An example of a flip-chip packaged 60 GHz chip with filter-in-package solution.**

munications. With the increase in operation frequency, the dimensions of designed filters shrink proportionally. However, the minimum feature size becomes a fundamental limit for these filters.

In the literature, it is rare to find a transmission line based filter in PCB technology, where the minimum dimension is around 4–8 mil that is difficult for high performance filter design. In LTCC technology, the feature size is normally 2–4 mil, but its multi-layer structure is more suitable for cavity filter designs. Such a filter has low insertion loss with compact size suitable for system integration. Besides PCB and LTCC, organic material is often used for filter designs to meet the flexible or bendable circuit requirement. In addition, a new technology named wafer transfer technology (WTT) uses conventional Rogers substrate with metal and dielectric layers similar to monolithic technology. It allows further reduction in filter size with moderate fabrication cost between complementary metal oxide semiconductor (CMOS) and PCB.

Table 1 gives the summary of filter-in-package solutions. In the following, filter-in-package solutions are discussed according to the adopted technologies.

2.1 PCB Filters

In [1], a fourth-order 60 GHz BPF based on inverted microstrip gap waveguide was designed and fabricated using PCB technology and Rogers 3003 substrate (**Fig. 2**). The minimum insertion loss of the filter is less than 1.6 dB, and 2 dB in average in the passband. The filter achieves 2 GHz bandwidth at 62 GHz. The 11 mm length filter is embedded within a 10 cm inverted microstrip gap waveguide transition prototype (transition has 1.6 dB loss).

2.2 LTCC Filters

In [2], a narrow-band 60 GHz 3-pole BPF based on 3D integrated cavity resonators was proposed (**Fig. 3**). The filter is fabricated using LTCC technology with a dielectric constant of 5.5. The $\lambda_g/4$ slot lines are used for the coupling between resonators. The filter achieves a minimum insertion loss of 2.48 dB at center frequency of 58.7 GHz, a return loss of 16.4 dB, a 3-dB bandwidth of 1.38% (~0.9 GHz) and a size of around 6.5×3.92 mm^2.

In [3], a quasi-elliptic 60 GHz filter is designed by stacking dual-mode substrate integrated waveguide (SIW) cavities with a shielding cavity (**Fig. 4**). Two dual-mode SIW cavities were stacked together with a coupling cavity in between results in the four-layer structure. Two Jerusalem cross apertures were defected on the top and bottom layers, while two circle apertures were defected on the middle two layers for coupling purpose. The 60 GHz filter was fabricated using the multilayer LTCC technology with dielectric constant of 7.38 and loss tangent of 0.01. The designed BPF has a center frequency of 60 GHz, a bandwidth of 4.5%, minimum insertion loss of 4.9 (if the loss tangent is 0.0015, insertion loss can be reduced to 1.1 dB), a return loss of 21 dB, and a size of 2.36×2.36×1.1 mm^3.

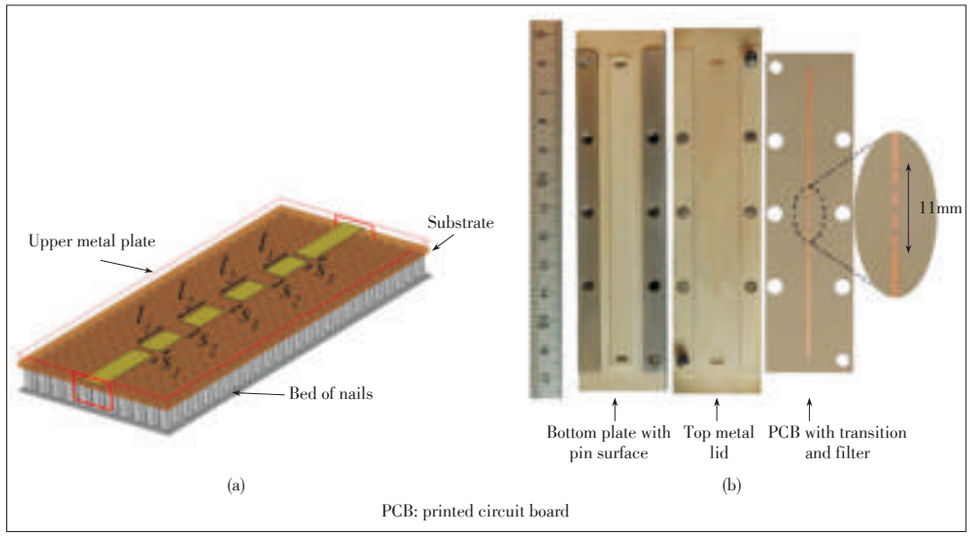
2.3 Flexible Material Based Filter

In [4], a wideband millimeter-wave filter is fabricated on a two-layered (or three layered including cover) flexible PerMX polymer (**Fig. 5**). The PerMX substrate is 50 μm thick and has a dielectric constant of 3. The three-pole filter is designed based on three half wavelength resonators. The filter is fabricated using PerMX 3050 polymer substrate and 14-μm thick Per-

▼**Table 1. Summary of state-of-the-art filter-in-package solutions**

Ref.	Technology	3-dB passband (GHz)/FBW (%)	Average insertion loss (dB)	Worst return loss (dB)	Lower stopband attenuation (dB/10 GHz)	Upper stopband attenuation (dB/10 GHz)	Size (mm^2)
[1]	PCB	61.5–63.5/3.2	2	12	~140	~93	~0.5
[2]	LTCC	58.3–59.2/1.38	3	12	~250	~300	25.5
[3]	LTCC	58.7–61.4/4.5	5	15	~57	~32	5.6
[4]	PerMX polymer	50.7–67.3/28	4.2	13	~17	~14	22.7
[5]	WTT	52–64/20.7	2.2	7	~1	~0.8	~2.5

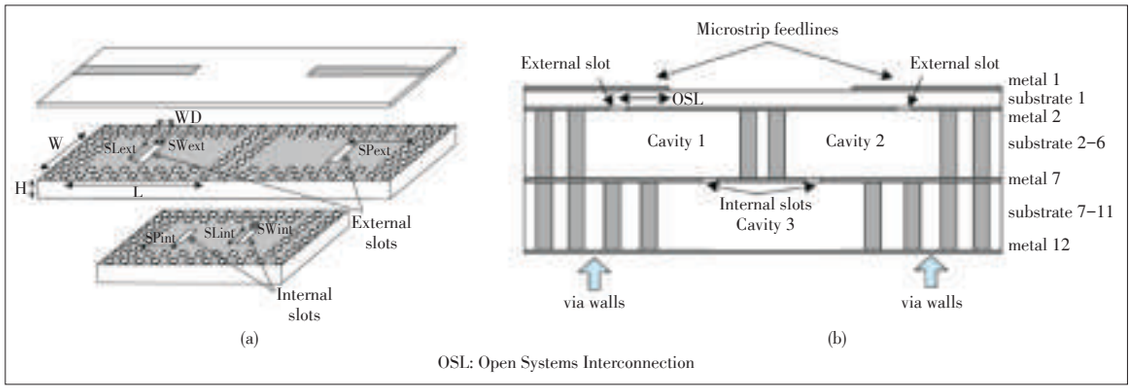
FBW: frequency widened bandwidth
LTCC: technology and low temperature co-fired ceramics
PCB: printed circuit board
WTT: wafer transfer technology



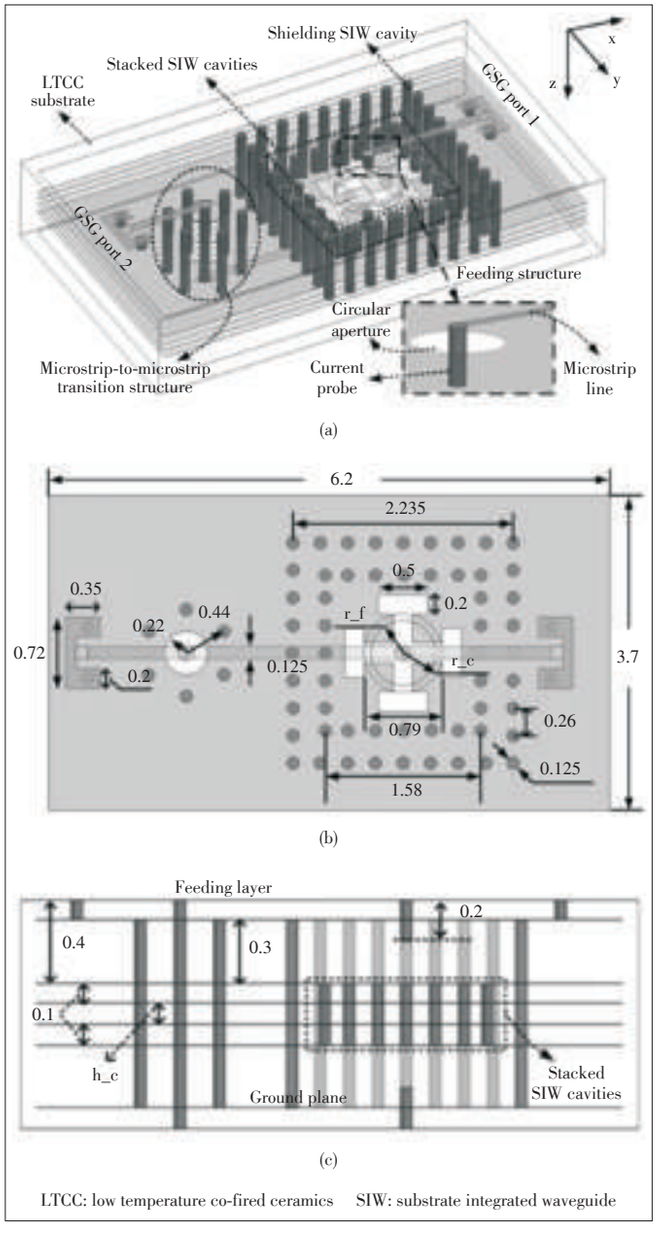
▲**Figure 2. Filter in [1]: (a) fourth-order inverted microstrip gap waveguide BPF and (b) a manufactured filter in 10 cm inverted microstrip gap waveguide transition prototype.**

State of the Art in Passive Bandpass Filter Solutions for 60 GHz Communications

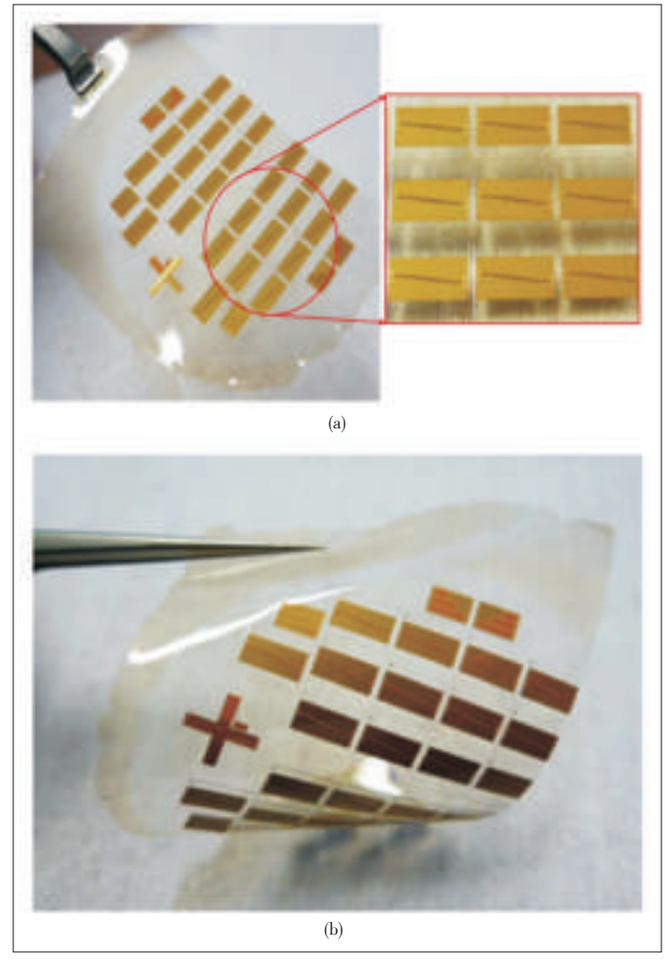
XU Shanshan, MENG Fanyi, MA Kaixue, and YEO Kiat Seng



◀**Figure 3. LTCC three-pole cavity BPF in [2] employing slot excitation with an open stub: (a) 3D overview and (b) side view of the proposed filter.**



▲**Figure 4. Filter in [3]: (a) Perspective view, (b) top view, and (c) side view.**



▲**Figure 5. Fabricated flexible PerMX substrate embedding filters: (a) PerMX substrate after the separation of Si support wafer and (b) bended PerMX flexible substrate [4].**

MX 3014 for cover. The designed filter achieves the minimum insertion loss of 3.8 dB at center frequency 59 GHz, a return loss of better than 13 dB, a 3-dB FBW of 28% and a size of 5.4×4.2 mm^2. PerMX has the advantages of low temperature process (<150° C), mechanical flexibility, bendable, low cost and low dielectric constant, and capability of implementing

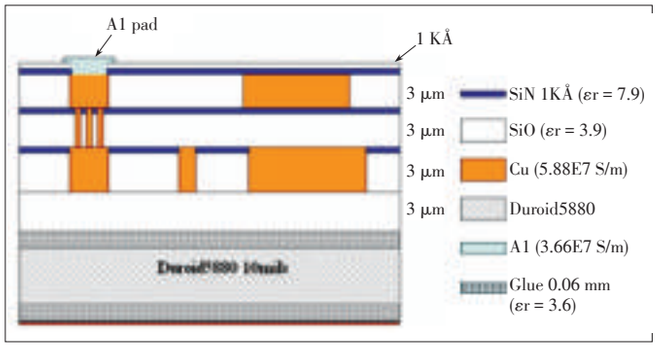
passive circuits to miniaturized RF SIP.

2.4 WTT Technology Based Filter

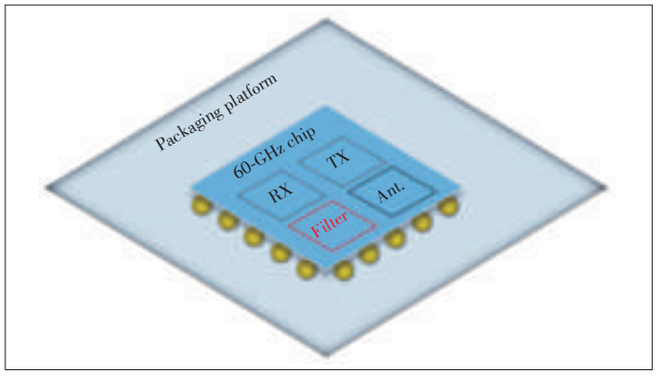
In [5], with the unique feature of small dimensions of about a few micrometers, WTT (**Fig. 6**) shows a promising future in solving the poor precision control of the dimension and the roughness of the pattern layer using PCB or thick-film process. Based on WTT, a low-insertion-loss bandpass filter is designed and fabricated at 60 GHz with 3 dB bandwidth of 12 GHz and insertion loss of 1.8 dB. The filter is prefabricated on a Si wafer with 3-μm-thick Cu and 1KÅ SiN dielectric layer and then is transferred to substrate Rogers RT/Duroid 5880.

3 Filter-on-Chip Solutions

Merging with 60 GHz RF front-end ICs, the on-chip filter is a promising solution for higher level of system integration compared to its in-package counterpart. **Fig. 7** shows an example of a 60 GHz chip using flip-chip packaging solution, with its 60 GHz filter designed and integrated on the same die. Generally, this filter solution has advantages of small packaged size, low form-factor, high level of integration, and possibility for antenna-on-chip integration. Further, the minimum achievable dimension of metal layers in CMOS technology is ~0.1 μm, which is far smaller than PCB or LTCC technology. Thus, the impedance control and design uniformity are much enhanced and reliable. However, due to the lossy silicon substrate and thin dielectric layer in (Bi)CMOS compared to other PCB or LTCC technologies, these filters have poor performance in terms of insertion loss, return loss, and stop-band rejection ratio. Furthermore, the overhead size occupied by the filter potentially increases the fabrication cost of chips.



▲**Figure 6. WTT layer definition of WTT in [5].**



▲**Figure 7. An example of a flip-chip packaged 60 GHz chip with filter-on-chip solution.**

Table 2 shows the performance of state-of-the-art filters on chips. In general, these filters meet the bandwidth requirement of IEEE 802.11ad standard. The insertion loss is around 3 dB–5 dB in average, while maintaining the return loss better than 8 dB. The stopband attenuation varies much across designs, and is usually worse than in-package filters. However, due to the on-chip technique, the filter size is very compact, resulting in a low form-factor overall packaged dimension of less than 0.2 mm^2.

It is noteworthy to highlight the following three works:

In [6], the authors proposed the second order 60 GHz BPF based on the folded open loop resonator (OLR) with H-shaped DGS (**Fig. 8**). The OLRs use metal 6 and the bridges use metal 5 to avoid inter-cross of the two OLRs. The filter was fabricated using 0.18 μm CMOS technology. The measured IL, return loss, centre frequency, and bandwidth are 2.85 dB, 18 dB, 59 GHz and 15.5 GHz with a chip size of 368 × 262 $μm^2$ including pads.

In [7], a low loss and compact filter was proposed using rectangular open-loop resonators with a proposed slotted ground, based on IBM 8RF-DM standard 0.13 μm CMOS process (**Fig. 9**). The open-loop resonators are designed on the top aluminum metal layer (MA) and the ground plane uses the bottom copper metal layer (M1). The thicknesses for MA and M1 are 4 μm and 1 μm, respectively, while the thickness between them is 16.6 μm. By optimizing the slot on M1 to prevent much more leakage to the lossy substrate, the filter achieves 1 dB bandwidth of 9 GHz with center frequency of 61.5 GHz, and an insertion loss of 1.5 dB, return loss better than 9.2 dB.

In [10], a 60 GHz multi-mode bandpass filter using a sandwich capacitor was implemented in 0.18 μm BiCMOS technology (**Fig. 10**). The capacitor functions as multi-mode perturbation that controls the stopband. Source/load coupling is used to generate two additional transmission zero point to further improve the stopband rejection. The measured filter features a minimum insertion loss of 4 dB and compact size of 0.16 mm^2.

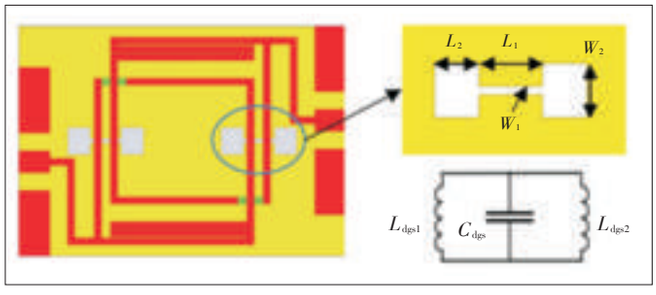
4 Future Trends and Design Challenges

The development of millimeter-wave filters is moving towards full on-chip solutions. SOC beam-forming systems better implement non-line-of-sight (NLOS) communications than the conventional point-to-point communication at millimeter-wave frequencies, Considering the huge demand of full SOC beam-forming systems [13]–[16], filters with high-level integration tend to be more attractive. In large-scale beam-forming systems especially, the packaging cost of in-packaging filters will

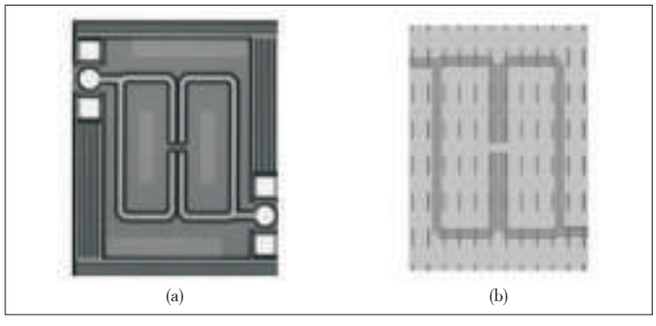
▼Table 2. Summary of state-of-the-art filter-on-chip solutions

Ref.	Technology	3 dB Passband (GHz)/FBW (%)	Average insertion loss (dB)	Worst return loss (dB)	Lower stopband attenuation (dB/10 GHz)	Upper stopband attenuation (dB/10 GHz)	Size (mm^2)
[6]	0.18 μm CMOS	51.25–66.75/26	3.2	10	19.5	7.7	0.10
[7]	0.13 μm CMOS	57–66/14.6	3.8	10	41.9	10	0.02
[8]	0.18 μm CMOS	49–71/36.7	5.4	8	15	25	0.10
[9]	0.18 μm CMOS	43–77/58	4.6	10	5	11	0.18
[10]	0.18 μm BiCMOS	50.5–66/27	5	15	15	8	0.18
[11]	0.13 μm CMOS	58.2–66.2/13	6.9	9.8	35	30	0.35
[12]	0.13 μm CMOS	56.2–67.4/18.1	3.9	10	7	11	0.29

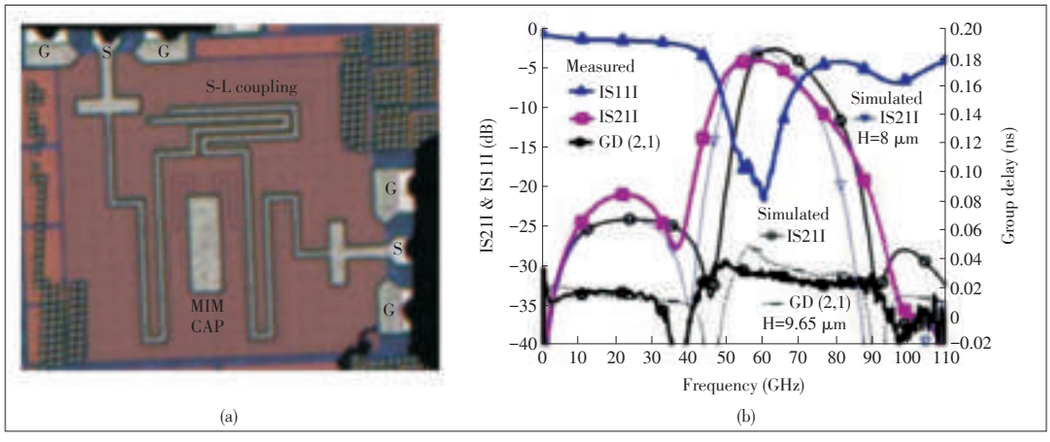
BiCMOS: bipolar complementary metal oxide semiconductor FBW: frequency widened bandwidth
CMOS: complementary metal oxide semiconductor



▲Figure 8. Proposed filter and its DGS geometry in [6].



▲Figure 9. Proposed filter in [7]: (a) top view and (b) ground plane.



▲Figure 10. Proposed filter in [10]: (a) micrograph of fabricated chip and (b) simulated and measured results.

become a burden as each of the beam-forming elements requires at least one filter.

However, the existing published works lack of good stopband rejection property, which is essential to at least reject interferences from radar channels at ~40 GHz–50 GHz and 77 GHz. At these radar frequencies, the signal strength is usually much higher than the 60 GHz communication carrier. Thus, poor rejection leads to possible saturation of input amplifier and malfunction of phase locked-loop due to cross-coupling. The fundamental reason is the low quality-factor of on-chip devices, typically around 10–20.

One possible solution is to utilize integrated passive devices (IPD) [17]–[20]. It is a form of separately fabricated devices but integrated with CMOS chip by flip-chip to the CMOS pad opening. The quality factor of such devices is around 100. Therefore, the filter developed in such process is expected to have excellent performance. The current challenges lie in the cost of IPD and flip-chip packaging cost by integrating IPD onto CMOS chips. The recent work in [21] has demonstrated the concept of using IPD for millimeter-wave applications. With the ever-evolving technologies, there is a great vision for IPD.

5 Conclusions

In the paper, we reviewed several recently published 60 GHz BPFs. The technology comparision shows that all filter designs are towards a lower fabrication cost and compact size. PCB or LTCC based filters have good insertion loss and return loss. However, the 60 GHz systems using PCB or LTCC filters have the size limited by filter sizes due to the lack of integration capability. In addition, most of the filters using PCB or LTCC are narrow band filters with FBW less than 5% . Technologies like WTT and new material like PerMX polymer are adopted to widen the bandwidth. This however sacrifice the filter selectivity, which is an important factor for filters especially when applied to the OOK transceiver. Due to the advancement of CMOS technologies, the performance of CMOS filters have been improved tremendously over the

years. Its insertion loss, however, is still too poor to be used in systems. To achieve both high-level integration and high performance, IPD is a potential solution in the future.

References

[1] A. Vosoogh, A. A. Brazález, and P.-S. Kildal, "A V-band inverted microstrip gap waveguide end-coupled bandpass filter," *IEEE Microwave and Wireless Components Letters*, vol. 26, no. 4, pp. 261–263, Apr. 2016. doi: 10.1109/LMWC.2016.2538598.

[2] J. H. Lee, S. Pinel, J. Papapolymerou, J. Laskar, and M. M. Tentzeris, "Low-loss LTCC cavity filters using system-on-package technology at 60 GHz," *IEEE Transaction on Microwave Theory and Techniques*, vol. 53, no. 12, pp. 3817–3824, Dec. 2005. doi: 10.1109/TMTT.2005.859864.

[3] D. Wang, K.-S. Chin, W. Che, Y. Wu, and C.-C. Chang, "Compact 60 GHz low-temperature cofired ceramic filter with quasi-elliptic bandpass response," *IET Microwaves, Antennas & Propagation*, vol. 10, no. 6, pp. 664–669, Apr. 2016. doi: 10.1049/iet-map.2015.0694.

[4] S. Seok and J. Kim, "Design, fabrication, and characterization of a wideband 60 GHz bandpass filter based on a flexible PerMX polymer substrate," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 3, pp. 1384–1389, 2013.

[5] Y.-X. Guo, J. Wang, B. Luo, and E. Liao, "Development of a 60-GHz bandpass filter and a dipole antenna using wafer transfer technology," *IEEE Electron Device Letters*, vol. 30, no. 7, pp. 784–786, Jul. 2009.

[6] A. Barakat, et al., "60 GHz on-chip mixed coupled BPF with H-shaped defected ground structures," *Electronics Letters*, vol. 52, pp. 533–535, 2016.

[7] B. Yang, E. Skafidas, and R. J. Evans, "Design of 60 GHz millimetre-wave bandpass filter on bulk CMOS," *IET Microwaves, Antennas & Propagation*, vol. 3, no. 6, pp. 943–949, Sept. 2009. doi: 10.1049/iet-map.2008.0222.

[8] J. Brinkhoff and F. Lin, "Integrated filters for 60 GHz systems on CMOS," in *IEEE International Workshop on Radio-Frequency Integration Technology (RFIT)*, Singapore, Dec. 2007, pp. 154–157. doi: 10.1109/RFIT.2007.4443940.

[9] S. Sun, J. Shi, L. Zhu, S. C. Rustagi, and K. Mouthaan, "Millimeter-wave bandpass filters by standard 0.18-μm CMOS technology," *IEEE Electron Device Letters*, vol. 28, no. 3, pp. 220–222, Mar. 2007.

[10] K. Ma, S. Mou, and K. S. Yeo, "Miniaturized 60-GHz on-chip multimode quasi-elliptical bandpass filter," *IEEE Electron Device Letters*, vol. 34, no. 8, pp. 945–947, Aug. 2013.

[11] B. Yang, E. Skafidas, and R. J. Evans, "60 GHz compact integrated cross-coupled SIR-MH bandpass filter on bulk CMOS," *Electronics Letters.*, vol. 44, no. 12, pp. 738–740, Jun. 2008.

[12] A.-L. Franc, E. Pistono, D. Gloria, and P. Ferrari, "High-performance shielded coplanar waveguides for the design of CMOS 60-GHz bandpass filters," *IEEE Transactions on Electron Devices*, vol. 59, no. 5, pp. 1219–1226, May 2012. doi: 10.1109/TED.2012.2186301.

[13] L. Kuang, X. Yu, H. Jia, et al., "A fully integrated 60-GHz 5-Gb/s QPSK transceiver with T/R switch in 65-nm CMOS," *IEEE Transaction on Microwave Theory and Techniques*, vol. 62, no. 12, pp. 3131–3145, Dec. 2014. doi: 10.1109/TMTT.2014.2364589.

[14] M. Boers, B. Afshar, I. Vassiliou, et al., "A 16TX/16RX 60GHz 802.11ad chipset with single coaxial interface and polarization diversity," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 12, pp. 3031–3045, Dec. 2014. doi: 10.1109/JSSC.2014.2356462.

[15] V. Vidojkovic, V. Szortyka, K. Khalaf, et al., "A low-power radio chipset in 40nm LP CMOS with beamforming for 60GHz high-data-rate wireless communication," in *IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, USA, Feb. 2013, pp. 236–237. doi: 10.1109/ISSCC.2013.6487715.

[16] E. Topak, J. Hasch, C. Wagner, and T. Zwick, "A novel millimeter-wave dual-fed phased array for beam steering," *IEEE Transaction on Microwave Theory and Techniques*, vol. 61, no. 8, pp. 3140–3147, Aug. 2013. doi: 10.1109/TMTT.2013.2267935.

[17] L. C. Hsu, Y. L. Wu, J. Y. Zou, H. N. Chu, and T. G. Ma, "Periodic synthesized transmission lines with 2-D routing capability and its applications to power divider and couplers using integrated passive device process," *IEEE Transaction on Microwave Theory and Techniques*, vol. 64, no. 2, pp. 493–501, Feb. 2016. doi: 10.1109/TMTT.2015.2513039.

[18] H. T. Kim, K. Liu, R. C. Frye, et al., "Design of compact power divider using integrated passive device (IPD) technology," in *59th Electronic Components and Technology Conference*, San Diego, USA, May 2009, pp. 1894–1899. doi: 10.1109/ECTC.2009.5074278.

[19] Y. C. Tseng and T. G. Ma, "On-chip miniaturized 3-dB directional coupler using coupled synthesized CPWs on integrated passive device (IPD) process," in *44th European Microwave Conference (EuMC)*, Rome, Italy, Oct. 2014, pp. 81–84. doi: 10.1109/EuMC.2014.6986374.

[20] Y. T. Lee, K. Liu, R. Frye, et al., "Ultra-wide-band (UWB) band-pass-filter using integrated passive device (IPD) technology for wireless applications," in *59th Electronic Components and Technology Conference*, San Diego, USA, May 2009, pp. 1994–1999. doi: 10.1109/ECTC.2009.5074295.

[21] A. Bisognin, D. Titz, F. Ferrero, et al., "IPD technology for passive circuits and antennas at millimeter-wave frequencies," in *7th European Conference on Antennas and Propagation (EuCAP)*, Gothenburg, Sweden, Apr. 2013, pp. 326–329.

Manuscript received: 2016-06-15

Biographies

XU Shanshan (shans.xu@gmail.com) received the BEng degree in information science and engineering from Southeast University (SEU), China in 2010, and is currently working towards the PhD degree at Nanyang Technological University (NTU), Singapore. Her research interests include microwave filter designs, substrate-integrated waveguide, defected ground structures, and millimeter-wave integrated circuits in CMOS technology.

MENG Fanyi (meng.fanyi@gmail.com) received the BEng (Hons) and PhD degrees from NTU in 2011 and 2016. His research interests include microwave, millimeter-wave and terahertz integrated circuits and phased-arrays in CMOS technology. Dr. Meng was the recipient of the 2013 Infineon-NTU Design Competition Bronze Award and 2015 Student Travel Grant from IEEE Solid-State Circuits Society.

MA Kaixue (makaixue@uestc.edu.cn) received his BE and ME from Northwestern Polytechnolgical University, China, and PhD degree from NTU. From Aug. 1997 to Dec. 2002, he was with China Academy of Space Technology (Xi'an), where he became the leader of the millimeter-wave group for space-borne microwave & mm-wave components and subsystem for satellite payload and VSAT ground station. From Sept. 2005 to Sept. 2007, he was with MEDs Technologies as a R&D manager. From Sept. 2007 to Mar. 2010, he was with ST Electronics (Satcom & Sensor Systems) as a R&D manager, project leader and member of Technique Management Committee of ST Electronics. He joined NTU as a senior research fellow and millimeter-wave RFIC team leader for the 60 GHz Flagship Chipset project in March 2010. As a PI/Technique leader, He did projects with fund more than S$12 Million (excluding projects done in China). In Feb. 2014, he joined the University of Electric Science and Technology of China as a full professor. His research interests include satellite communication, software defined radio, Microwave/MM-wave circuits and system using CMOS, MEMS, MMIC and LTCC. He has eight patents, two patents in pending and authored/co-authored over 80 referable international journal and conference papers in the related area. He received Best Paper Award from IEEE SOCC2011, IEEK SOC Design Group Award, Excellent Paper Award from International Conference on HSCD2010, Chip Design Competition Bronze Award of ISIC2011. He is a senior member of IEEE.

YEO Kiat Seng (kiatseng_yeo@sutd.edu.sg) received the BEng (EE) in 1993, and PhD (EE) in 1996 both from NTU. He is now Associate Provost (International Relations and Graduate Studies) at Singapore University of Technology and Design (SUTD) and a member of Board of Advisors of the Singapore Semiconductor Industry Association. Dr. YEO is Fellow of IEEE and a widely known authority in low-power RF/mm-wave IC design and a recognized expert in CMOS technology. He has secured over S$30M of research funding from various funding agencies and the industry in the last 3 years. Before his new appointment at SUTD, he was Associate Chair (Research), Head of Division of Circuits and Systems and Founding Director of VIRTUS of the School of Electrical and Electronic Engineering at NTU. He has published six books, five book chapters, over 400 international top-tier refereed journal and conference papers and holds 35 patents. Dr. Yeo served in the editorial board of *IEEE Transactions on Microwave Theory & Techniques* and hold/held key positions in many international conferences as Advisor, General Chair, Co-General Chair and Technical Chair. He was awarded the Public Administration Medal (Bronze) on National Day 2009 by the President of the Republic of Singapore and was also awarded the distinguished Nanyang Alumni Award in 2009 for his outstanding contributions to the university and society.

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 004
http://www.cnki.net/kcms/detail/34.1294.TN.20161221.1058.002.html, published online December 21, 2016

Low-Power High-Efficiency Multi-Gigabit 60 GHz Transceiver Systems Routing in Vehicular Environments

Chul Woo Byeon[1] and Chul Soon Park[2]
(1. Department of Electronic Engineering, Wonkwang University, Iksan 54538, Korea;
2. Department of Electrical Engineering, KAIST, Daejeon 305-701, Korea)

Abstract

This paper proposes low-power high-efficiency multi-gigabit 60 GHz transceiver systems for short-range communications. 60 GHz multi-gigabit on-off keying (OOK) receiver system-in-package (SiP) module is developed using a low temperature co-fired ceramic (LTCC) technology. Integrated with a low-power complementary metal oxide semiconductor (CMOS) OOK demodulator, the LTCC receiver module demonstrates Full-HD uncompressed video streaming at a distance of 1 m. A low-power and high-efficiency fully integrated OOK transceiver is also developed to be integrated in a handheld device. The transceiver consumes 67 mW at 10.7 Gb/s and occupies an active footprint of 0.44 mm^2. With an on-board Yagi-Uda antenna, the transceiver achieves 10.7 Gb/s of data transmission, resulting in a high energy efficiency of 6.26 pJ/bit. The antenna-in-package module with the transceiver demonstrates mobile-to-display 1080p Full-HD video transmission over a distance of 60 cm.

Keywords

60 GHz; Transceiver; on-off keying (OOK); Multi-gigabit

1 Introduction

In response to the demand for higher data rates, the millimeter-wave (mm-wave) band has been studied. Among the mm-wave bands, the 60 GHz band has been widely researched to achieve multi-gigabit data transmission [1]–[12]. The studies on a silicon-based 60 GHz band have demonstrated multi-gigabit wireless data transmission, which confirms the possibility of removing all high speed wire lines. With the 60 GHz band, we can surely remove wires of the signal lines and wirelessly connect all devices such as smart phones, laptops, displays, and tablet PCs. However, the transceivers in [1]–[3] consume a huge amount of DC power, which limits the use of mobile devices because of their limited battery capacity. The transceivers in [4]–[6] consume a low amount of DC power, but they support a limited communication distance of less than 20 mm.

This paper presents an overview of the low-power high-efficiency multi-gigabit 60 GHz transceiver systems. The proposed work simultaneously meets the requirements of the power and transmission distance budgets. This paper is organized as follows. Section 2 describes the low-power, high-efficiency 60 GHz multi-gigabit complementary metal oxide semiconductor (CMOS) solutions. Section 3 describes a 60 GHz multi-gigabit on-off keying (OOK) transmit/receive module integrated with a low-power CMOS OOK demodulator and GaAs chips. Section 4 presents a low-power high-efficiency transceiver module with an on-board antenna. Finally, this paper concludes with a summary in Section 5.

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MEST) (No. R11-2005-029-05001-0), the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (13-911-04-007), and the National Research Foundation of Korea (NRF-2014R1A2A1A11054107).

2 60 GHz Multi-Gigabit CMOS Solutions

OOK modulation is one of the promising candidates for the short-range wireless communications due to its high integrity, simplicity, and low-power consumption. Several studies investigated the OOK modulation using silicon-based technologies [4]–[10]. The OOK transmitter simultaneously sends a data signal and a local carrier signal, and the OOK receiver detects the data using the local carrier signal from the transmitter. With this principle, the transmitter and receiver require no phase-lock alignment and the architecture and design are simplified.

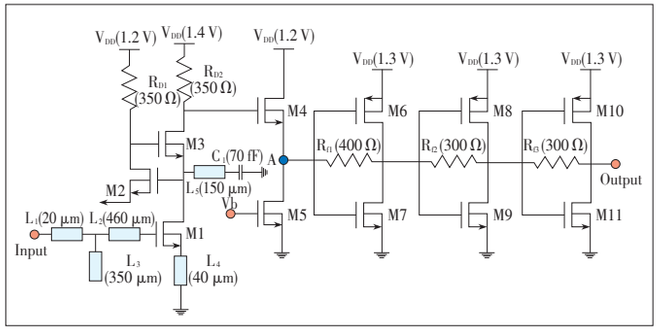
The transceiver in [4] presents the basic architecture of the OOK system. The amplitude modulation was implemented by

an up-conversion mixer, and the modulated signal was amplified by a power amplifier. Similarly, the transmitters in [5], [6], [13], [14], and [15] modulated the signal using an up-conversion mixer and amplified the modulated signal using a power amplifier. The up-conversion mixer as an OOK modulator has a high conversion loss, which causes high power dissipation to obtain a target transmit power at the transmitter. Therefore, the OOK transceiver systems are limited by the transmission distance or DC power dissipation. In the receiver, a detector without a phased-locked oscillator also has a high conversion loss [4], [14]. To overcome the high conversion loss of the detector, the receivers described in [5] and [6] added an injection path for the injection-locked oscillator. The injection-locked oscillator improved conversion gain, but the additional gain stage and oscillator consumed huge DC power, and the sensitivity of the injection-locked oscillator limited the receiver sensitivity. Alternatively, the receiver in [13] recovered the signal at an intermediate frequency but enhanced the gain and sensitivity of the detector at the cost of high DC power.

To overcome these drawbacks, we introduce a gain-boosting demodulator [9], a switching amplifier type of modulator, and a high-efficiency transmitter [8].

2.1 Highly Efficient Gain-Boosting Demodulator

To substitute the injection-locked oscillator [5], [6], the detector should have a high conversion gain for lower DC power and better sensitivity and wider bandwidth for high data rate. Therefore, the gain-boosting demodulator was introduced. A schematic of the 60 GHz gain-boosting OOK demodulator in 0.13 μm CMOS is shown in **Fig. 1**. The gain-boosting stage, M2-M3, contributes higher conversion gain while the conventional detector only uses M1 and has limited gain and bandwidth characteristics. The voltage conversion gain of the gain-boosting detector is $1+A$ times greater than the voltage conversion gain of a conventional detector, where A is the gain from the common source of M2. The dominant pole frequency of the gain-boosting detector is $1/((C_{D3}+C_L)R_{03})$, where R_{03} is the output resistance at the drain of M3, while dominant pole frequency of the conventional detector is $1/((C_{D1}+C_L)R_{03})$. C_{D1} is much larger than C_{D3} for larger conversion gain (higher gm2); thus, the pole frequency of the gain-boosting detector is larger than one of the conventional detectors. A resistive feedback inverter type amplifier is added after the 60 GHz OOK detector for high output voltage.

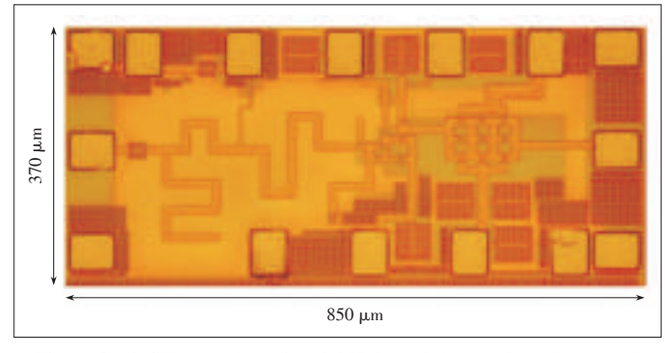


▲Figure 1. 60 GHz gain-boosting OOK demodulator schematic.

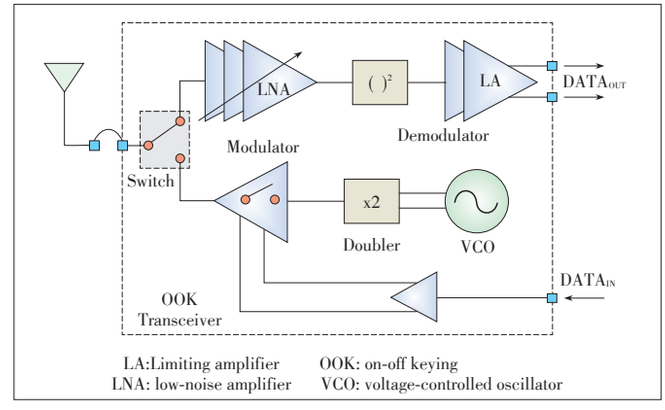
A microphotograph of the fabricated demodulator is shown in **Fig. 2** The demodulator consumes a small chip area of 0.32 mm^2 and a low DC power of 14 mW. With the gain-boosting technique, the detector has a voltage responsivity of 2434 mV/mW and the demodulator has a conversion gain of 13.6 dB at an input power of -16 dBm. The demodulator possesses a data capability of 5 Gb/s. Thus, an energy efficiency of 2.94 pJ/bit is achieved.

2.2 Low Power and Highly Efficient OOK Transceiver

Future wireless connectivity will require low power consumption, a high data rate, and short-range data transmission. To meet those requirements, a gain-boosting demodulator, a switching amplifier type of modulator, and a high-efficiency transmitter were introduced in [8]. We fabricated and demonstrated a low power and highly efficient OOK transceiver in 90 nm CMOS. **Fig. 3** shows a block diagram of the transceiver. The required signal-to-noise ratio of 10^{-12} bit error rate (BER) for OOK modulation is approximately 17 dB. A noise figure of 9 dB in the Rx is assumed. The required output power of the Tx is 5 dBm with antenna gain of 7 dBi over a distance of 30 cm at a data rate of 3.5 Gb/s. And the Rx sensitivity is -49 dBm. The transmitter consists of an OOK modulator, a dou-



▲Figure 2. 60 GHz gain-boosting OOK demodulator microphotograph.



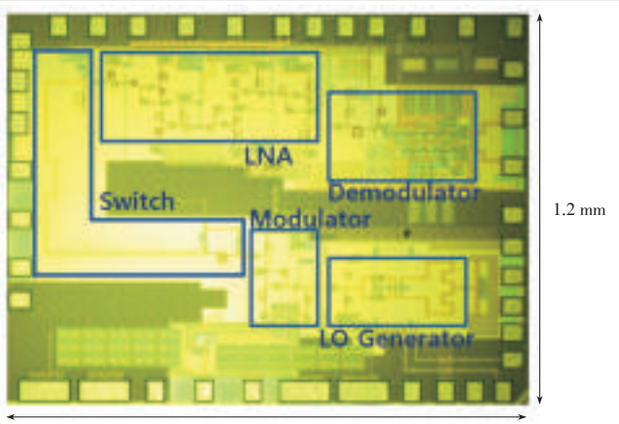
▲Figure 3. 60 GHz transceiver block diagram.

bler, and a 30 GHz voltage-controlled oscillator (VCO). The modulator provides high output power and a modulation function. Futhermore, the modulator consumes DC power only in the data-on state, which results in a huge reduction in DC power consumption. The modulator has a gain of 9.1 dB in the on-state and a gain of -15.2 dB in the off-state; hence, it has an on/off isolation of 24.3 dB at 60 GHz. The output 1 dB compression point is 5.1 dBm with power-added efficiency of 12.7% at 60 GHz. To further reduce the DC power consumption and enhance the efficiency, the frequency doubler is directly attached to the 30 GHz VCO. The local oscillator (30 GHz VCO + frequency doubler) generates an output power of -0.9 dBm at 60 GHz. The on-state transmitter has an output power of 4.4 dBm with 8.9% power efficiency, including the switch loss.

The receiver consists of a low-noise amplifier (LNA) and a demodulator. To achieve a high data rate and high efficiency, we designed the LNA to have a wide bandwidth, low-noise figure, and high gain. The LNA has gains of 24.5 dB (high gain mode) and 15.5 dB (low gain mode) at 60 GHz with a 3 dB bandwidth of 14 GHz. A noise figure of 4.9 dB–6 dB is attained for the high gain mode. The demodulator adopts a gain-boosting technique and consists of a limiting amplifier and a detector. The detector has a voltage responsivity of 5892 mV/mW at 5 Gb/s and the limiting amplifier has a gain of 21.7 dB with a 3 dB bandwidth of 3.5 GHz. A die photograph of the transceiver is shown in **Fig. 4**. The transmit/receive switch is integrated for a small package size.

3 60 GHz Multi-Gigabit Transmit/Receive Modules

Multi-layer low-temperature co-fired ceramic (LTCC) technology is widely used due to its low-attenuation, low cost, similar value of temperature coefficient of expansion, and high integration capability to multi-/single-chip solutions [13]. **Fig. 5** shows the LTCC system-in-package (SiP) receiver and transmitter modules. The LTCC has a relative permittivity and a loss tangent of 5.8 and 0.0035, respectively. The size of the entire SiP with five ceramic layers is 37mm x 11mm x 0.5mm and 25.5mm x 10.5mm x 0.5mm for Tx and Rx, respectively. The radiating patches are placed on the top layer, and feeding network is located on the second layer while the ground plane is on the back side of the third layer. The antenna, fed by the bondwire and with the T-network for bondwire compensation, radiates through the coplanar waveguide with the ground to the embedded micro-strip line. The T-network consists of inductor, wide width transmission line (capacitor) and narrow width transmission line (inductor). The receiver and transmitter modules have a 1 by 2 patch antenna and a single patch antenna, respectively. The antennas have a simulated gain of 9.8 dBi and 4.8 dBi and a half power beamwidth of 100° and 60°, respectively. The transmitter module is integrated with a 30 GHz VCO, a frequency doubler, an up-conversion mixer, a drive amplifier, and a power amplifier. The receiver module is integrated with an LNA and the gain boosting demodulator in Fig. 2.



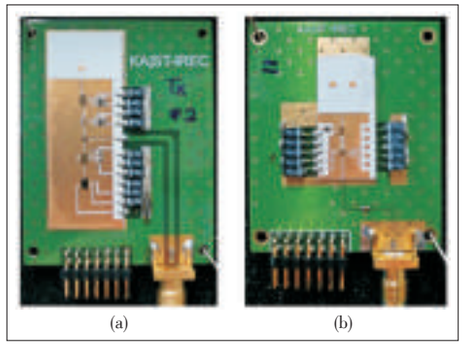
LNA: low-noise amplifier LO: local oscillator

▲Figure 4. 60 GHz transceiver chip microphotograph.

For multi-gigabit data transmission, we tested the fabricated modules. **Fig. 6** shows a demonstration of the multi-gigabit video signal transmission. Video data of 1.485 Gb/s is transmitted at a distance of 1 m.

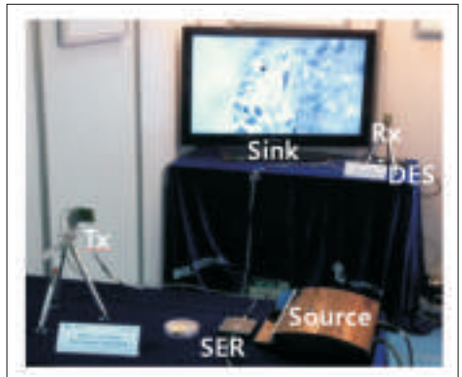
4 Highly Efficient 60 GHz Transceiver Module

An on-board antenna is one of the best candidates consider-



(a) (b)

Figure 5. ▶ 60 GHz system-in-package (SiP): (a) transmitter and (b) receiver.



DES: Deserializer SER: Serializer

Figure 6. ▶ Demonstration of multi-gigabit video signal transmission.

ing yield and cost. In this design, we use a Taconic substrate for the antenna design. The substrate has a dielectric constant of 3 and a thickness of 0.13 mm. A Yagi-Uda antenna [16] is used for high gain and end-fire radiation. The fabricated antenna size is 4 mm X 9 mm. **Fig. 7** shows the simulated radiation pattern of the antenna and the 60 GHz transceiver module with the high-efficiency CMOS transceiver and external bias circuitry in Fig. 4. The fabricated antenna is directly attached next to the CMOS transceiver chip on the PCB. The 60 GHz signal is connected to the antenna through two parallel gold wires with an approximate length of 300 μm, and their parasitic inductance is resonated out using a matching network on the antenna. The antenna has a simulated gain of 9.8 dBi. The beamwidth for E-plane and H-plane is 48° and 60°, respectively.

The measurement performance and wireless video signal transmission of the transceiver module are depicted in **Fig. 8**. A data rate of 10.7 Gb/s was achieved over a distance of 6 cm with a BER less than 10^{-12} for the 2^7-1 pseudorandom binary sequence (PRBS). Also, data rates of 5 Gb/s were observed over distances of 40 cm with a BER of less than 10^{-12}. **Table 1** compares this work with previously published works. This work shows the lowest energy efficiency of 6.26 pJ/bit with a high data rate of 10.7 Gb/s and a low power consumption of 67 mW. The 60 GHz transmitter module sends Full-HD data from a mobile phone to a display through the 60 GHz receiver module. Video data of 2.97 Gb/s is transmitted a distance of 60 cm.

5 Conclusions

This paper presents low-power high-efficiency multi-gigabit 60 GHz transceiver systems. For a highly efficient design, we develop a gain-boosting demodulator and a highly integrated low-power transceiver in a CMOS technology, which can handle multi-gigabit data transmission faster than 5 Gb/s. We develop the 60 GHz multi-gigabit LTCC receiver module with the low-power CMOS OOK demodulator and demonstrate multi-gigabit (1.485 Gb/s) video data transmission. Futhermore, to be integrated in a handheld device, the antenna-in-package module with the transceiver demonstrates mobile-to-display 1080p (2.97 Gb/s) wireless Full-HD video transmission over a distance of 60 cm.

References

[1] M. Boers "A 16TX/16RX 60 GHz 802.11ad chipset with single coaxial interface and polarization diversity," in *IEEE International Solid-State Circuits Conference*, San Francisco, USA, Feb. 2014, pp. 344– 345. doi: 10.1109/ISSCC.2014.6757462.

[2] A. Tomkins, et al., "A 60 GHz, 802.11ad/WiGig-Compliant Transceiver for Infrastructure and Mobile Applications in 130 nm SiGe BiCMOS," *IEEE Journal of*

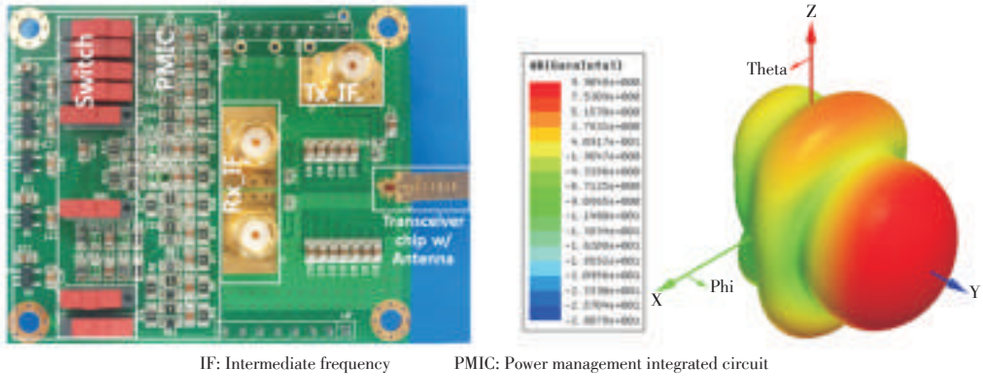


Figure 7. Photograph of 60 GHz transceiver module and simulated radiation pattern of the antenna.

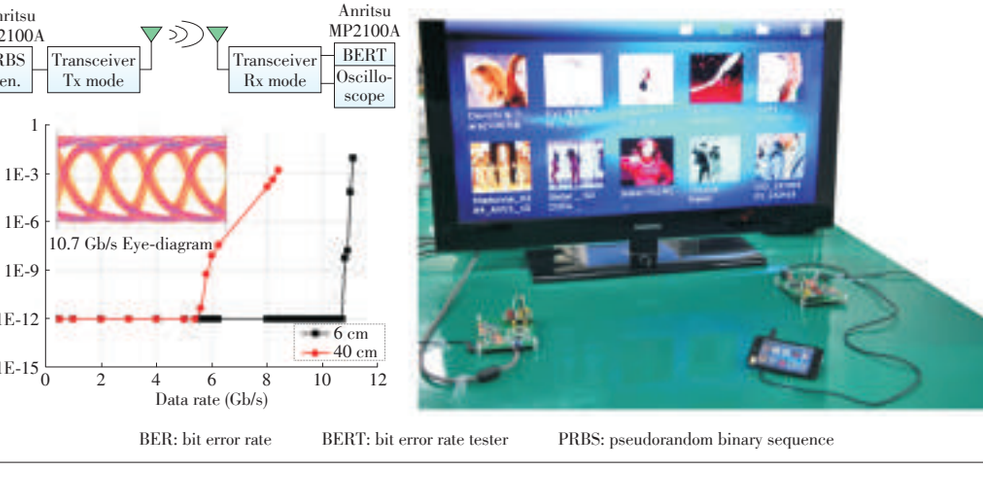


Figure 8. Measured transceiver performance and demonstration of multi-gigabit (Full-HD) video signal transmission.

▼Table 1. Performance comparison

References	[14]	[5]	[6]	[17]	[18]	This work
Technology	90 nm CMOS	40 nm CMOS	40 nm CMOS	45 nm SOI CMOS	65 nm CMOS	90 nm CMOS
Carrier frequency	60 GHz	57 GHz	57/80 GHz	60 GHz	60 GHz	60 GHz
Modulation	OOK	ASK	ASK	BPSK	16QAM	OOK
Power dissipation	286 mW	70 mW	137 mW	317 mW	542 mW	67 mW
Switch	No	No	No	No	No	Yes
Tx/PA P_{1dB}	5 dBm (PA)	-2 dBm (Tx)	-15.2/-9.5 dBm (Tx)	15 dBm (Tx)	6 dBm (Tx)	5.1 dBm (PA) 4.4 dBm (Tx)*
Tx efficiency (P_{OUT}/P_{DC})	1.73 %	2.18 %	0.27 %	15.3 %	1.3 %	8.9 %**
Rx sensitivity	-30 dBm***	-33 dBm***	-	-	-35 dBm***	-32.5 dBm
Data rate	2.5 Gb/s	11 Gb/s	20 Gb/s	5 Gb/s	7 Gb/s	10.7 Gb/s
Distance	40 mm	14 mm	5 mm	loopback	300 mm	100 mm
EE	114 pJ/bit	6.4 pJ/bit	6.85 pJ/bit	63.4 pJ/bit	77.4 pJ/bit	6.26 pJ/bit
Chip area	Tx: 0.43 mm^2 Rx: 0.68 mm^2	Tx: 0.06 mm^2 Rx: 0.07 mm^2	Tx: 0.16 mm^2 Rx: 0.26 mm^2	4.42 mm^2	5.48 mm^2	Tx: 0.15 mm^2 Rx: 0.29 mm^2

*: Tx output power including switch
**: Including switch
***: Calculated from table
ASK: amplitude-shift keying
BPSK: binary phase shift keying
CMOS: complementary metal oxide semiconductor
EE: energy efficiency
OOK: on-off keying
PA: Power Amplifier
QAM: quadrature amplitude modulator
SOI: silicon on insulator

Solid-State Cicuits, vol. 50, no. 10, pp. 2239–2255, Oct. 2015. doi: 10.1109/JSSC.2015.2436900.
[3] K. Okada "A 64-QAM 60 GHz CMOS transceiver with 4-channel bonding," in *IEEE International Solid-State Circuits Conference*, San Francisco, USA, Feb. 2014, pp. 346–347. doi: 10.1109/ISSCC.2014.6757463.
[4] W. Chen, S. Joo, S. Sayilir, et al., "A 6-Gb/s wireless inter-chip data link using 43-GHz transceivers and bond-wire antennas," *IEEE Journal of Solid-State Cicuits*, vol. 44, no. 10, pp. 2711– 2721, Oct. 2009. doi: 10.1109/JSSC.2009.2027932.
[5] K. Kawasaki, Y. Akiyama, K. Komori, et al., "A millimeterwave intra-connect solution," *IEEE ISSCC Digest of Technical Papers*, Feb. 2010, pp. 414–415. doi: 10.1109/ISSCC.2010.5433831.
[6] Y. Tanaka, Y. Hino, Y. Okada, et al., "A Versatile Multi-Modality Serial Link," *IEEE ISSCC Digest of Technical Papers*, Feb. 2012, pp. 332–333. doi: 10.1109/ISSCC.2012.6177034.
[7] C. Byeon, et al., "A 60-GHz transceiver system with low-power CMOS OOK modulator and demodulator," in *IEEE International Microwave Workshop Series on Intelligent Radio for Future Personal Terminals (IMWS-IRFPT)*, Daejeon, Korea (South), Aug. 2011, pp. 100–101. doi: 10.1109/IMWS2.2011.6027205.
[8] C. Byeon, C. Yoon, and C. Park, "A 67-mW 10.7-Gb/s 60-GHz OOK CMOS transceiver for short-range wireless communications," *IEEE Transaction on Microwave Theory and Techniques*, vol. 61, no. 9, pp.3391–3401, Sept. 2013. doi: 10.1109/TMTT.2013.2274962.
[9] C. Byeon, J. J. Lee, K. Eun, and C. S. Park, "A 60 GHz 5 Gb/s Gain-boosting OOK Demodulator in 0.13μm CMOS," *IEEE Microwave and Wireless Components Letters*, vol.21, no.2, pp. 101– 103, Feb. 2011. doi: 10.1109/LMWC.2010.2096411.
[10] J. J. Lee and C. Park, "A 60 GHz Gbps OOK Modulator with High Output Power in 90nm CMOS," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 58, no. 5, pp. 249– 253, May 2011. doi: 10.1109/TCSII.2011.2124890.
[11] C. Byeon and C. S. Park, "Design and Analysis of the Millimeter-wave SPDT Switch for TDD Applications," *IEEE Transaction on Microwave Theory and Techniques* , vol. 61, no. 8, pp. 2258–2864, Aug. 2013. doi: 10.1109/TMTT.2013.2271613.
[12] C. Byeon, J. J. Lee, I. Song, and C. S. Park, "A 60 GHz Current-Reuse LO-Boosting Mixer in 90 nm CMOS," *IEEE Microwave and Wireless Components Letters*, vol.22, no.3, pp. 135– 137, Mar. 2012. doi: 10.1109/LMWC.2011.2182636.
[13] I. S. Song, et al., "A low loss 60GHz radio integrating CMOS circuits with LTCC AiP," in *Proc. 2011 IEEE EDAPS*, Hangzhou, China, pp. 1–4. doi: 10.1109/EDAPS.2011.6213757.
[14] Jri Lee, Yenlin Huang, Yentso Chen, Hsinchia Lu, and Chiajung Chang, "A Low-Power Fully Integrated 60GHz Transceiver System with OOK Modulation and On-Board Antenna Assembly," *IEEE ISSCC Digest of Technical Papers*, Feb. 2009, pp. 316–317,317a. doi: 10.1109/ISSCC.2009.4977435.
[15] E. Juntunen, et al., "A 60-GHz 38-pJ/b 3.5-Gb/s 90 nm CMOS OOK digital radio," *IEEE Transaction on Microwave Theory and Techniques*, vol. 58, no. 2, pp. 348–355, Feb. 2010. doi: 10.1109/TMTT.2009.2037867.
[16] H. Yagi, "Beam transmission of the ultra short waves," *Proc. IRE*, vol. 16, pp. 715–741, Jun. 1928. doi: 10.1109/JRPROC.1928.221464.
[17] T. Dinc, A. Chakrabarti, and H. Krishnaswamy, "A 60 GHz CMOS Full-Duplex Transceiver and Link with Polarization-Based Antenna and RF Cancellation," *IEEE Journal of Solid-State Cicuits*, vol. 51, no. 5, pp. 1125–1140, May 2016. doi: 10.1109/JSSC.2015.2507367.
[18] K. Okada, et al., "Full four-channel 6.3-Gb/s 60-GHz CMOS transceiver with low-power analog and digital baseband circuitry," *IEEE Journal of Solid-State Cicuits*, vol. 48, no. 1, pp. 46–65, Jan. 2013. doi: 10.1109/JSSC.2012.2218066.

Manuscript received: 2016-06-15

Biographies

Chul Woo Byeon (cwbyeon@wku.ac.kr) received the PhD degree in electronic engineering from Korea Advanced Institute of Science and Technology (KAIST), Korea, in 2013. During 2013, he was a postdoctoral researcher with the Department of Electrical and Computer Engineering from University of California, San Diego, USA. From 2014 to August 2015, he was a senior engineer with Samsung DMC R&D Center, Korea, where he was involved in developing a millimeter-wave/RF transceiver for 5G cellular communications. In September 2015, he joined the Department of Electronic Engineering, Wonkwang University, Korea, where he is currently an assistant professor. His research interests include CMOS/SiGe millimeter-wave/RF integrated circuits, antennas, packages, and system design for wireless communications.

Chul Soon Park (c-spark@kaist.ac.kr) received the BS degree from Seoul National University, Korea in 1980, and the MS and PhD degrees in materials science and engineering from the KAIST in 1982 and 1985, respectively. From 1985 to 1999, he was with the Electronics and Telecommunication Research Institute (ETRI), Korea, where he contributed to the development of semiconductor devices and circuits. From 1987 to 1989, he studied the very initial growth of group IV semiconductors for laser applications during a visit to AT&T Bell Laboratories, USA. Since 1999, he has been with the Information and Communications University (which merged with KAIST in 2009), where he is a full professor with the Engineering School. His research interests include reconfigurable RF integrated circuits (RFICs), millimeter-wave integrated circuits (ICs), and their system-on-chip (SoC)/system-on-package (SoP) integration.

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 005
http://www.cnki.net/kcms/detail/34.1294.tn.20170105.1823.004.html, published online January 5, 2017

Current Situation and Development of Intelligence Robots

REN Fuji[1, 2] and SUN Xiao[1]
(1. Anhui Key Lab of Affective Computing and Advanced Intelligence Machine, Hefei University of Technology, Hefei 230009, China;
2. Tokushima University, Tokushima 770800, Japan)

Abstract

Industrial intelligent robots are treated as a measure of national scientific level and technology innovation, and also the important symbol of high-level manufacturing, while service intelligent robots can directly affect people's daily lives. The development of artificial robots in different areas is attracting much attention around the world. This article reviews the current situation and development of Chinese and international intelligent robot markets including industrial robots and service robots. The intelligent robot technology and the classification of robots are also discussed. Finally, applications of intelligent robots in various fields are concluded and the development trends and outlook of intelligent robots are explored.

Keywords

intelligent robot; artificial intelligence; development trends

1 Introduction

In 1997, IBM's Deep Blue computer defeated the world's chess champion, Garry Kasparov. The match was a momentous occasion for artificial intelligence, showing how far a clever algorithm could go in a game prized for its intellectual difficulty. In 2015, AlphaGo became the first program to beat a professional Go player in an even game after it won 5−0 in a formal match against the reigning 3-times European Champion, FAN Hui. Alpha Go then won 4−1 against the top Go player in the world, LEE Sedol in 2016. After that, an era of artificial intelligence opened.

Since the 21st century, the development of robot technology has been paid more attention at home and abroad. Robotics is considered to be one of the high-level technologies of great significance for the development of emerging industries. UK Royal Academy of Engineering predicted that 2019 would usher in the robot revolution according to the Division Studies Report on Autonomous Systems in 2009 [1]. In 2014, Chinese President XI Jinping emphasized at the Seventeenth Academician Conference at Chinese Academy of Science: "Development, manufacture and application of robots are the measure of a national scientific and technological innovation and an important symbol of high-end manufacturing level; we need not only improve the level of Chinese robots, but also occupy the market as much as possible [2]." Robot revolution is expected to become an entry point and an important growth point of the third industrial revolution, and to affect the pattern of global manufacturing. China is expected to become the world's largest robot market.

Robots broadly include any machines that can simulate behaviors or thoughts of humans and animals, such as robotic dogs, robotic cats, and robotic fish. In a narrow sense, robots have many different definitions and classifications. Some computer programs can also be treated as a robot (such as the crawling robot for search engine). The United Nations Organization for Standardization adopts the definition from the US Robotics Association and defines the Robot as "a programmable and multifunctional manipulator; or a specialized system that performs different tasks can be changed by computer and is programmable. Generally, it's comprised of implementing agencies, driving device, detection device, control system, complex mechanism and other components [3]." A robot is a complex intelligent machine composed by machinery, electron, computers, sensors, control technology, artificial intelligence, bionics and other disciplines. Currently, the intelligent robot has become one of the research hotspots in the world, and an important symbol to measure a country's level of industrialization. Robots are automated robotic devices to perform some specific work, so that it can accept human commands, and also run pre-programmed programs. Moreover, it can perform some human tasks based on the principles and program developed by the artificial intelligence technology. In contemporary industry, the robot means an artificial robotic device that can perform tasks automatically, used to replace or assist human work [4]; usually it is an electromechanical device, controlled by a computer program or electronic circuit. Robots can be autonomous or semi-autonomous and range from humanoid such as Honda's Advanced Step in Innovative Mobility (ASIMO) and TOSY's TOSY Ping Pong Playing Robot (TOPIO) to industrial robots. It also includes collectively programmed swarm robots, and even Nano robots. By mimicking a lifelike appearance or automating movements, the ideal high simulation robot is an advanced product that integrates control theory, machinery,

electronics, computer and artificial intelligence, materials science, and bionics. Robots can perform repetitive and dangerous tasks that humans prefer not to do or are unable to do because of size limitations or extreme environments such as outer space or the bottom of the sea. Nowadays, robots resemble humans in more and more fields or aspects, for example, robots can resemble humans in appearance, behavior cognition, and even emotion. Robotics technology was first used in industry. In recent years, driven by the computer technology, network technology, micro-electro-mechanical system (MEMS) technology, development of new technologies, the use of robot technology has been expanding rapidly from the traditional industrial manufacturing to medical service, education and entertainment, exploration and surveying, bio-engineering, disaster relief and many other areas. Robot systems adapted to the needs of different areas are intensively researched and developed. Over the past decades, research and application of robotics technology have greatly promoted the industrialization and modernization of the human society, and gradually formed a robotics industry chain, which is further expanding the scope of robot applications.

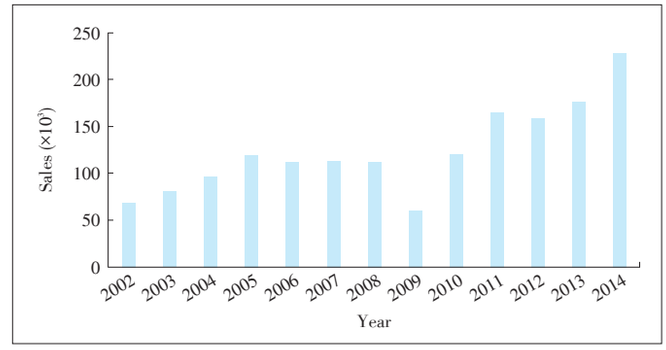
In the following sections, the current development of robots is introduced in section 2. Intelligent robot-related technologies and trends are presented in section 3. The emotional robot is also presented in this section. Section 4 is the conclusions and suggestions for future development of robotics in China.

2 Development Situation of Robots

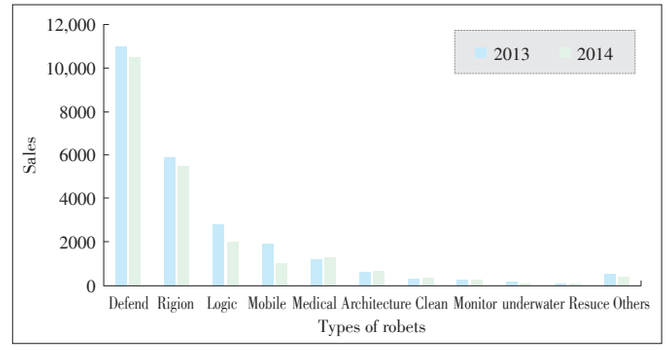
2.1 Global Development

As a measure of a country's scientific and technological innovation and an important indicator of the level of high-end manufacturing, the robot industry gets more and more attention around the world. The major economies have treated the development of this industry as a national strategy and an important means to maintain or regain their competence in manufacturing. The study of robotics started earlier abroad and is comparatively mature there. As the representatives, the United States, Japan and Europe have developed a variety of robots based on their needs. According to the report of the International Federation of Robots (IFR) [5], 229,261 industrial robots [6] were sold in 2014, with an increase of 29% compared to 2013 (**Fig. 1**). All industrial robot companies have achieved growth. In China, the robots sold reached 37,000 units in 2014, with an increase of 60%, ranking the world's top one [7].

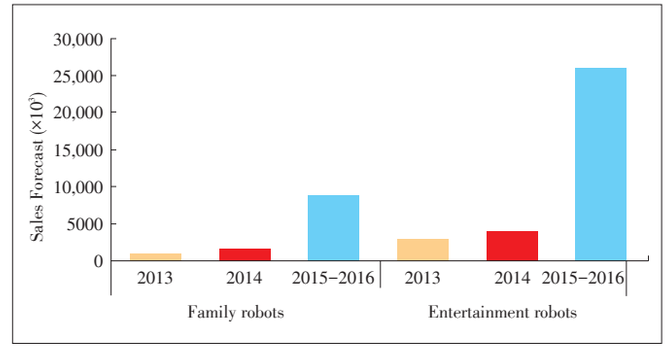
The sales of service robots grew by 11.5% in 2014 [8] (**Figs. 2** and **3**), and the sales amount increased 3% to $377 million [9]. It was predicted that various types of service robots could reach 25.9 million sets during 2015 to 2018, with an estimated value of $1.22 billion [10]. According to a report by Allied Market Research [11], the global industrial robot market would grow at a compound growth rate of 5.4% from 2014 to 2020



▲**Figure 1. Annual supply of industrial robots (IFR) [6].**



▲**Figure 2. Comparison of service robots (IFR) [8].**



▲**Figure 3. Sales prediction of service robots (IFR) [8].**

and its sales would reach $41.17 billion in 2020. McKinsey Global Institute published a report named "12 Disruptive Technologies Leading Global Economic Changes" in 2013 [12], and listed the advanced robotics, cloud computing, the next generation of gene technology, 3D printing, new materials, renewable energy and other 6 technologies. In 2025, robots will bring economies of scale from $1.7 to $4.5 trillion per year worldwide according to McKinsey's prediction [12].

2.1.1 Development in US

In 2010, the US launched its "Advanced Manufacturing Partners' Program", with an explicit goal of revitalizing the manufacturing through the development of industrial robots and developing a new generation of intelligent robots by using

information network technology. In 2013, it released a robot development route report with a subtitle of "From Internet to Robotics", in which the intelligent robot was put in the equally important position with the Internet in 21th Century [13]. Robots will influence human life, economic and social development in all respects, and robots technology has been listed as the core technology to achieve manufacturing improvement and economic development in the US.

2.1.2 Development in EU

The EU has launched the world's largest civilian robot research and development program—SPARC, planning to vote 2.8 billion Euros and create 240,000 jobs in 2020. The program includes more than 200 companies and 12,000 R&D persons are involved. Robot applications in various fields such as manufacturing, agriculture, health, transport, security and the family are included in the plan.

2.1.3 Development in Germany

Germany has proposed its Industry 4.0 plan to maintain its leading position in manufacturing. The plan considers intelligent robotics and intelligent manufacturing technology as an entry point to greet the new industrial revolution.

2.1.4 Development in Japan

Japan also has a long-term development strategy of robot technology. The robot industry is one of the priorities in the seven major industries supporting its new industrial development strategy. The Japanese government is planning to use robots as an important pillar of economic growth.

2.1.5 Development in South Korea

South Korea has developed a "Smart Robot Basic Plan", and issued a "Robot Future Strategic Vision 2022" in October 2012. The policy focuses on expanding its robot industry and supporting Chinese robot enterprises to expand overseas markets.

2.2 Development in China

Robot manufacturing in China began in the 1970s, with a focus on researching and manufacturing industrial robots. Although robot technology research and development in China started late, the development speed is very fast. The development can be roughly divided into three stages. In 1980s, the main emphasis was researching and manufacturing industrial robots, with the support of national "863" program and other plans. In the early 1990s, the techniques with independent intellectual property rights emerged, including spot welding, arc welding, assembling, painting, cutting, handling, packaging palletizing and more, and China took an important step in industrial robots in practice. The Chinese government introduced robot development plans covering different aspects of robotics during this period. After the prototype and demonstration phase of the 1990s, China started to enter the industrial stage in 2000. In 2006, China included intelligent service robots in the national long-term technology development program, and in 2012, it published the "Twelfth Five-year Special Plan" for service robot technology development. The "Ministry of Industry Guidance on Promoting the Development of Industrial Robots" was released in 2013. After 2010, the capacity of Chinese installed robots has increased annually, with the development of the robot industry chain.

In recent years, China has made a lot of robot technology achievements and the market prospect of robot industry is bright. Many Chinese research institutes, such as Harbin Institute of Technology, Beijing University of Aeronautics and Astronautics, Tsinghua University and Hefei University of Technology, are carrying out related research of intelligent robots. Hefei University of Technology has built an emotion robot platform which will be presented in the following sections.

The world's four giant robot manufacturers including Sweden ABB, German KU-KA, Japanese FANUC) and Japanese YASKAWA have set up branches in China and are positive to the prospect of Chinese robot market. On the other hand, the Chinese robot brands have also grown and begun to take shape, such as Shenyang XINSONG Robot Automation, Anhui Efort Intelligent Equipment, and Guangzhou Dongguan STS robotics. Until October 2014, there were more than 430 robot-related companies in China, with an average of two new companies appearing per week. There are more than 4000 companies involved in industrial robots and the number is still growing with an annual increase of more than 300 enterprises. According to Great Wall Securities' latest research, Chinese industrial robot industry will have a explosive growth and the market is expected to reach 100 billion in 2020. In the next six years, the total installed capacity is expected to reach the range of 638,000–1,760,000, with a conservative estimate of 850,000 units. Chinese service robotic industry is also expected to an unparalleled growth in 5 to 10 years. With a conservative forecast, in the next six years, the total Chinese market size of industrial robots will be ¥127.5 billion, that of service robots is ¥144.3 billion, and that of the system integration market is ¥382.5 billion. Specifically, the market size of military ground robots will be ¥34 billion; that of unmanned aerial vehicles will be about ¥46 billion; that of service robots for aged people will be about ¥39 billion; that of assistive robots will be ¥24.3 billion; and that of public service robots will be around ¥10 billion.

The Chinese Robot Industry Alliance led by China Machinery Industry Federation was established in Beijing on April 21, 2013. The alliance aims to vigorously promote the production, science and research of Chinese robots, and to accelerate the universal application of robotics technology and products in various sectors. The alliance has more than one hundred members and try to achieve the healthy and orderly development of the robot service platform based on the optimization of the in-

dustrial chain, innovative integration of resources, complementary advantages, cooperative development, and win-win cooperation models. China has exported mobile robots in batches, and Chinese service robots and special robots began to form a strong competitive power. The total sales of Chinese industrial robots are over 9500 units in 2013. Based on this, China supports the construction of robot industry bases, such as Shenyang XINSONG Robot Automation and Harbin Boshi Automation, which are leading the development of Chinese robotics industry.

However, Chinese robot brands still have a low market share in China. International robot brands account for more than 90% of the Chinese market share. Six large Japanese robot companies occupy 50% of the Chinese industrial robot market, while Chinese four robotic equipment manufacturers only account for 5% of the Chinese market. Chinese industrial chain of robots, including production, manufacture, sale, integration and ordered service has not yet formed. Compared with the major developed countries, Chinese robot industry develops slowly, the core technology is weak, and the market share and additional values are low. However, with the wide application and expanding of robotics technology and products, robots will maintain rapid development in China.

3 Intelligent Robot-Related Technologies and Trends

3.1 Intelligent Robot-Related Technologies

There are certain technical indicators and criteria to measure intelligent robot technology level, and the robotics capability evaluation includes:

- Degree of intelligence. It mainly refers to a robot's feel and perception of the outside world, including memory, calculation, comparison, identification, judgment, decision making, learning ability, logical reasoning ability, etc.
- Performance characteristics, including task flexibility, common areas or space possessory, etc.
- Physical energy index. It generally refers to a robot's power, speed, reliability, combinability and service life. For some special robots, it also includes actuators, drives, sensors and control systems, and complex mechanical and other components.

3.1.1 The Robot Actuator

The robot actuator is the body of a robot and the robot arm (if any) and generally uses the space open chain bar linkage, where the motion pairs (rotation pair or revolute pair) are often called joints. The number of joints is typically the robot degrees of freedom. According to the different joint configuration types and movement coordinate forms, robot actuators can be divided into rectangular coordinate type, cylindrical coordinate type, polar coordinate type, and joint type. In some certain application scenarios, the relevant parts of the robot body are often called base, waist, arms, wrists, hands (gripper or end effecter), walking part (for mobile robots), etc. for anthropomorphic considerations.

3.1.2 The Drive Device

The drive device drives the movement of the actuator, in accordance with instructions signals issued by the control system, and by means of the dynamic element to make robot perform the related action. The input of a drive device is an electric signal and the output is a linear and angular displacement. Driving devices used for a robot are mainly electric drives, such as stepper motors and servomotors. In addition, for the specific needs of a particular scene, the hydraulic and pneumatic drives are also used.

3.1.3 The Sensing Device

Robots generally get external information through a variety of sensors. The sensors are used for real-time detection of the robot's internal movement and work and of the external operating environment information as well. A sensor also feeds back to the control system as needed, after comparison with the set of information, adjusting the actuator to ensure that the operation of the robot meets predetermined requirements. Sensor devices used for detection can be divided into two categories. One is the internal information sensor for detecting the internal status of the various parts of the robot, such as the position of each joint, velocity and acceleration, and the measured information as a feedback signal will be sent to the controller, forming a closed loop control. The other is the external information sensor for acquiring the work object, external environment and other relevant information of the robot, to make movement of the robot adapt to changes in the external conditions. In this way, the robot can achieve a higher level of automation and even lead the robot to have some human-like feels. The robot also gets the development of intelligence, such as using the visual, sound and other external sensors to get the work objects, work environment and relevant information, and then use the information to form a large feedback loop, which will greatly improve the working precision of the robot.

3.1.4 Control System

One is the centralized control, that is to say, the entire control of the robot is completed by a micro computer. The other is the dispersed (level) control, which uses more than one computer (at upper and lower levels) to complete control of the robot. The host computer is often responsible for system management, communications, kinematics and dynamics calculations, and for sending command information to the lower level computer. As a subordinate slave, each joint corresponds to a CPU which could do interpolation operation and servo control process with real specific movement, and then passes the feedback to the host. Depending on the different mission require-

ments of tasks, the control modes of the robot can be divided into the position control, continuous path control and force (torque) control.

3.1.5 Intelligent System

Intelligent system refers to the computer system that produces human-like intelligence or behavior. This system can own the self-organization and adaptability to run on a conventional Von Neumann machine, and also run on a new generation computer of non-Neumann architecture. The concept of "intelligence" covers a wide range and is also constantly in an evolving process. Its essential needs further exploration, thus, it is difficult to give a complete and precise definition for the word "intelligence". It is generally stated like: intelligence is reflection of higher activity of the human brain, and it should at least have the ability to automatically acquire and apply knowledge, think and reason, solve problem and learn automatically [14]. Smart robots have the function of completing similar intelligence capabilities with humans. For an intelligent system, its processed objects are not only data, but also knowledge. The ability to represent, acquire, access and process is one of the main differences between the intelligent system and the mechanical system. Therefore, an intelligent system is a knowledge-based processing system. It requires the following facilities: knowledge representation language; knowledge organization tools; the method and environment to establish, maintain and query knowledge base; and supporting the reuse of existing knowledge. Intelligent systems often use the artificial intelligence problem solving mode to get the results. Compare with the traditional system's problem solving mode, it has three distinct features: 1) its problem solving algorithms are often non-deterministic or heuristic; 2) problem solving relies heavily on knowledge; and 3) intelligent system problems tend to have exponential computational complexity. The typical problem solving methods an intelligent system uses are roughly divided into three categories of search, reasoning, and planning. Another important difference between the intelligence robot system and traditional system is that the intelligent system has field perception (acclimatization) capabilities. Scene perception helps robots to interact with the abstract of real world and adapt to their scene. Such exchanges include perception, learning, reasoning, judgment and appropriate action making. This is well known as automatic organizing and automatic adapting. Nowadays, the chatbot is another hot spot in robot technology. Many chatbot systems and methods have been developed and studied such as Siri and Xiaobing.

3.1.6 Intelligent Human-Machine Interface System

Intelligent robots cannot be in full autonomy yet, and it still needs to interact with people. Even a fully autonomous robot also needs to feedback real-time implementation of the mandate to people. The intelligent human-machine interface system enables a robot to provide users with a friendly, natural and good adaptive human-computer interaction system. With the support of intelligent interface hardware, the intelligent human-machine interface system generally has the following features:

- Using natural language directly to lead a human-machine dialogue
- Allowing multimedia such as sounds, text, graphics and images to lead the human-machine interaction
- Interacting with human by brain waves and other physiological signals
- Self-adapting to different user types
- Self-adapting to the needs of different users
- Self-adapting to different computer systems support.

3.2 Development and Trends of Intelligent Robot Technology

Intelligent robots are the third generation robots with a variety of sensors. It can fuse information obtained by multiple sensors and effectively adapt to the changing environments, with strong adaptive ability, learning ability and autonomic functions. Multiple key technologies decide the intelligence level of an intelligent robot, which are introduced as follows.

3.2.1 Multi-Sensor Information Fusion Technology

The multi-sensor information fusion technology integrates sensory data from multiple sensors to produce more reliable, accurate and comprehensive information. The multi-sensor fusion system can more accurately reflect the characteristics of the detection target, eliminate uncertainty information, and improve the reliability of information. For example, the emotional robot can adopt multimodal information from human to get accurate emotion of human, by combing text, sounds, facial expressions and information from other channels or models.

3.2.2 Navigation and Location Technology

For autonomous mobile robot navigation, whether it is local real-time obstacle avoidance or global planning, the technology is used to precisely tell the current state and position of the robot and obstacles for completing tasks like navigation, obstacle avoidance and path planning.

3.2.3 Path Planning Technology

Based on one or multiple optimization criterions, the optimal path planning technology finds an optimal path from the initial state to target state and avoids obstacles in the robot workspace. Nowadays, almost all the moving robots use this technology to increase the covering rate.

3.2.4 Robot Vision Technology

The robot vision system implements image acquisition, image processing and analysis, output and display. Its core task is feature extraction, image segmentation and image recognition. Nowadays, deep learning and related technologies have been widely adopted in this field and great progress has been

made.

3.2.5 Intelligent Control Technology

Intelligent control methods improve the speed and precision of the robot. The human-machine interface technology makes it possible for people to naturally and conveniently communicate with the robot.

3.3 Wide Use of Intelligence Robots

Modern intelligent robots can basically complete various complex tasks according to people's instructions, such as deep sea exploration, combat, reconnaissance, intelligence gathering, and rescue. Robots can complete the tasks humans are unable or unwilling to complete independently, and also collaborate with people to complete tasks under the guidance of people. Therefore, they have been widely used in different fields.

According to different workplaces, intelligent robots can be divided into pipes, water, air and ground robots. Pipe robots can be used to detect the rupture, corrosion and weld quality in the course of using pipeline, doing pipe cleaning, painting, welding, internal polishing and other maintenance work in the harsh environments, and repairing underground pipeline. Underwater robots can be used for scientific marine research, offshore oil development, seabed mineral exploration, undersea salvage lifesaving, etc., while air robots can be used in terms of communications, meteorology, disaster monitoring, agriculture, geology, transport, radio and television and other aspects. Service robots work semi-autonomously or autonomously to provide services for human, for example, the robots used in the traditional Chinese medicine field have a good application prospect. With a human-like shape, humanoid robots have a mobile function, operation function, perceptive function, memorizing and self-government ability, to achieve friendly man-machine interaction. Micro-robot based on nanotechnology has broad application prospects in bio-engineering, medical engineering, micro-electromechanical systems, optical, ultra-precision machining and measurement (such as scanning tunneling microscope) and more.

3.3.1 Robots for National Defense

In the field of national defense, military intelligent robots have got unprecedented attention and development. In recent years, the US and UK have developed the second generation of military intelligent robots, such as the United States' Navlab autonomous navigation vehicles, SSV autonomous ground combat vehicles, and Big Dog robots. The robots use independent control to complete the reconnaissance, combat and logistical support and other tasks. They are able to see and smell on the battlefield, to automatically track the terrain and choose the path, and to automatically search, identify and eliminate enemy targets. In the future, military intelligent robots will include intelligent fighting robots, intelligent reconnaissance robots, intelligent alert robots, intelligent engineering robot, intelligent transportation robot, and more, and they will become a new bright spot in defense equipment. The unmanned combat air vehicle (UCAV) is an upgraded form of the unmanned aerial vehicle (UAV) and can do a variety of tasks, including combat. Unmanned aircrafts, such as BAE Systems Mantis, have the ability to fly autonomously, choose routes and goals independently, and make decisions independently. BAE Raytheon is an unmanned combat aircraft researched and developed by the UK, and it cannot fly across the continent without pilots and new methods to avoid detection.

3.3.2 Service Robots

The whole world, especially western countries, is committed to the research and development of the wide application of intelligent service robots, although the definition of "service robots" is not clear. The International Federation of Robotics gives a preliminary definition: "service robot refers to a class of robots, by its semi-autonomous or fully autonomous operation, providing helpful service for the guardian of human health or monitoring equipment operation status, but does not include industrial operations [15]." Take the cleaning robots as an example. With the scientific and technological progress and social development, people want to be free more from the tedious daily affairs, which make cleaning robots enter into families possible. The floor cleaning robot developed by Japanese companies can start from any one location along the wall automatically using constantly rotating brush to sweep the waste into its own container. The station floor cleaning robot sprays the cleaning fluid onto the ground, and simultaneously uses constantly rotating brush scrubbing the floor and sucked dirty water to its container. The factory automatic cleaning robot can be used for a variety of clean-up work. A cleaning robot named "Roomba" of America, with high autonomy, can walk in the gaps of furniture in each room, deftly completing clean-up work. A Sweden robot "trilobite", with smooth surface, round shape and built-in search radar, can quickly detect and avoid the leg, glassware, pets or any other obstacles. Once its microprocessor recognizes these obstacles, it will re-select the route, and make judgment and re-calculated for the entire room to ensure that every corner of the room is to be cleaned.

Service robots can recognize people or objects, talk, provide companionship, monitor environment quality, respond to alarms, pick up supplies and perform other useful tasks. They perform multiple functions at the same time, or play different roles at different times of the day. Some robots try to imitate humans; this type of robot is called a humanoid robot. The humanoid robot is still at a very limited stage, and until now no humanoid robot can navigate in the room it had never been to. Thus, the functions and applications of humanoid robots are quite limited, despite their considerable intelligence behavior in familiar environment. Semi-autonomous robots, such as Friend (**Fig. 4**) and other various wheelchair robots can help the elderly and the disabled people to complete some common

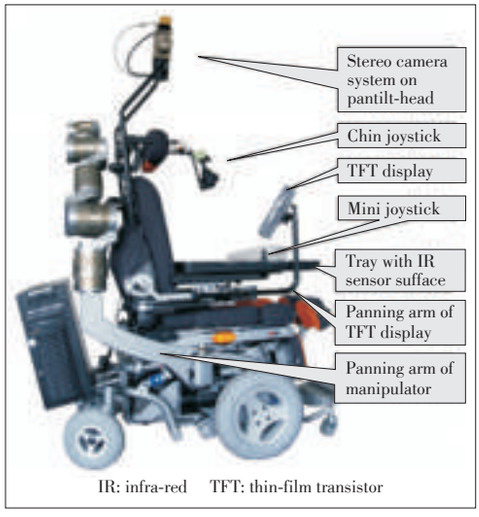


Figure 4. Wheelchair robot "Friend".

tasks. The increase of aging population in many countries, especially in Japan, means that there are more and more elderly people in need of care, but relatively fewer young people could care for them. Human beings are the best caregivers, but they are very busy. That is why robots are gradually being introduced. Friend (Fig. 4) is a semi-autonomous robot (wheelchair robot), helping the disabled and the elderly in daily life, such as preparing and serving meals. Friend makes it possible for paraplegic patients with muscle diseases or serious paralysis (due to stroke and other reasons) to complete some tasks without the help of a therapist or a nurse.

3.3.3 Education Robots

In the field of education, the robot has had a long involvement since the 1980s; the Turtle robot has been put into use in schools with Logo programming language [9], [16]. Moreover, the robot kits, such as Lego, robot teaching kit "BIOLOID", OLLO robot and BotBrain educational robot, can help children learn math, physics, program, electronics and other knowledge. FIRST company also introduced robots into primary and secondary students' life through robot competitions. FIRST company organized the first robot competition, the first LEGO League, the basic of Junior Lego Match and the first Technique Challenge. There were also some devices whose shapes like robots, such as teaching computer Leachim (1974) and 2-XL (1976) that is an 8-track tape player.

3.3.4 Sports Robots

In the field of sports, the intelligent robot has also been greatly developed. In recent years, high-tech combat activities of soccer robots have been carried out among nations, and the international community has set up relevant federations, such as Federation of International Robot-soccer Association (FIRA). Many regional associations have also been set up, reaching a more formal level and a considerable scale. In order to win a robot soccer match, the camera hanging in the air transfers the match status to the computer and the pre-installed software makes appropriate decision and measures and passes the command to the robot players by way of wireless communications. The robot in a soccer match fuses the computer vision, pattern recognition, decision countermeasure, wireless digital communications, automatic and optimal control, intelligent body design and electric drive technology into its system to realize intelligence.

3.3.5 Humanoid Robots

Robots with emotions have also been developed rapidly in recent years. The REN Research Lab in Tokushima University, Japan has developed a conversation robot with emotions [17] (**Fig. 5**). During a conversation with the human, the robot can detect the human emotions and make corresponding responses considering both content context and human emotions.

Anhui Key Lab of Affective Computing and Advanced Intelligence Machine, Hefei University of Technology (China), has achieved some progress in emotional assist robots. The researchers in the Lab study affective computing systems on the humanoid robot platform for mental health problems, funded by the National "863" Program of China. The humanoid robot is built in the Lab according to a real person's outlook and body structure (**Fig. 6**). In order to make a perfect approximation, the reverse method is used for the robot head design according to the real person's size; the robot's hair is also planted and cut according to the real person's style (**Fig. 7**). The researchers in the Lab think that a realistic appearance can improve the acceptability of the robot.

The robot was first made to perform six basic emotions, including surprise (**Fig. 8**) and sad (**Fig. 9**) . Then the researchers built a personalized fusion heart state transition network for the robot, and developed a multi-model emotion conversation model based on the transition network. In this way, the robot could make reactions to emotions (**Fig. 10**).

The researchers of the Lab have also built a female emotional robot (**Fig. 11**). This female emotional robot is designed as a platform to enhance mental health for the human. According to the user's micro-blog, blog, conversational language, voice, facial expressions and emotional interaction, the robot could perceive human's mental health status and calculate the heart en-

Figure 5. The conversation robot with emotions from REN Lab.

◀Figure 6. Body structure size collection for robot.

◀Figure 7. Planting and cutting robot's hair.

Figure 8. ▶ The robot simulates surprise.

Figure 9. ▶ The robot simulates sadness.

▲Figure 10. Robot-human emotional conversation.

▲Figure 11. Female emotional robot from HFUT.

richment degree. The system is able to appease and chat with persons. The researchers have also developed an emotional robot conversation cloud platform. Key functions of the platform include person identification and emotion cognition, gestures and voice interaction, intelligent emotional chat, and other emotional interaction. Emotional robots can be adopted at home and medical facilities for people of all ages (especially the elderly), providing recovery assistant service in specific conditions (autism and depression).

Modern intelligent robots have not been widely used, but will permeate all aspects of life. For example, the coal and mining industry has demands for intelligent robots because of the harsh environment in coal mining; for building construction, there are high-rise buildings plastering robots, mounting robots, indoor decoration robots, robots for wiping glass, floor polishing robots and so on. The nuclear industry requires smart, accurate, reliable, quick and lightweight robots. For emotional accompanying, service robots use cameras and microphone sensors to collect people's real-time information, obtain their real-time emotional states by machine learning and data mining algorithms, and then synthesize the corresponding human facial expressions and gestures to do the real-time emotional interaction. The application fields of intelligent robots are expanding day by day and people are using intelligent robots to replace human beings to complete more complex and more advanced work.

3.4 Robot Technology Trends

Various robotic technologies constantly emerge. One approach is to use evolutionary robotics. A parent robot constructs some different sub-robots and after testing, the sub-robots with best performance will be used as a standard model to create a new generation of the robot. Another approach is developing robots (developmental robotics) by tracking the inner changes of the robot in solving problems in specific functional areas, thereby improving the intelligence of the robot. A new one has just launched and is named new RoboHon robot, which can be seen as a smart phone and a robot as well [18]. Japan expects to achieve full commercialization of service robots by 2025 and many leading technology research institutions there are led by the Japanese government, particularly by its Ministry of Economy and Trade [19]. With the development

of robots, it is expected to establish a standard computer operating system for robot design. The robot operating system is a set of open source codes and is being developed by Stanford University, Massachusetts Institute of Technology, Technical University of Munich, and more. Robot Operating System (ROS) provides methods for robot navigation and program of its limbs. These methods do not need to consider the specific hardware involved. It also provides high-level commands, such as image recognition and even opening the door. When ROS starts on computer of a robot, it will get the robot's attribute data, such as the length of the limbs and robot motion data. Then it passes the data to a higher-level algorithm. Microsoft started its research and development of robots in 2007 [20], and is developing a "robot window" system. Caterpillar is making an automatic driving truck [21], without any manual operation.

Intelligent robots in various industries have bright application prospects and the intelligent robot research at home and abroad has made many achievements. However, there is still much room for growth of intelligence level [22], [23]. Future intelligent robots will be task-oriented [24], [25]. Because the current artificial intelligence cannot provide intelligent machines with a complete theory and methods for open tasks, most of the existing artificial intelligence technologies rely on domain knowledge and special robots are developed for special tasks.

Some artificial intelligence technologies in specific areas are used for promoting the development of intelligent robots:

- Sensor technology and integration technology: The sensor technology is used to develop better and more advanced processing methods based on the existing sensors, and find new sensors as well. The integration technology improves the integration information.
- Robot cloud interconnection technique: Using the cloud Internet network technology, robots are connected to a computer network and their knowledge base comes from cloud [26], [27]; on the other hand, the cloud can effectively control the robots through the computer network.
- Calculation methods of intelligent control system: Compared with traditional methods, the fuzzy logic, reasoning based on probability theory, neural networks, genetic algorithms and chaos calculation [28]–[30] have higher robustness, ease of use, low cost calculation and more advantages. When applied to robotics, these methods can help robots to improve the speed of problem solving and to handle multivariable and nonlinear system problems better.
- Machine learning in smart robots: The emergence of various machine learning algorithms promote the development of artificial intelligence, depth of learning, reinforcement learning, ant colony algorithm, immune algorithm and others for robotic systems, making robots possess the human-like learning ability, adapt to increasingly complex, uncertain and unstructured environments.
- Human interface of intelligent optimization: The demand of human-computer interaction is developed toward simplicity, diversity, intelligence and humanization. Therefore we need study and design a variety of intelligent human-machine interfaces [30]–[32] such as multi-lingual voice, natural language understanding, images, handwriting recognition, and even physiological information, in order to better adapt to different users and different application tasks, and eventually to improve human-robot interaction harmony.
- Multi-robot coordination schemes: They organize and control multiple robots to collaborate on complex tasks that cannot be completed by a single robot, realizing real-time reasoning reaction, group decision making and operation of interaction in a complex and unknown environment.

4 Conclusions

About 50% of the world's robots are in Asia, 32% in Europe, 16% in the northern United States, 1% in Australia, and 1% in Africa [33]. Japan has 40% of the robots in the world [34]. Currently, Japan has the largest number of robots. As robots become more advanced and complex, more and more experts and scholars have even begun to explore the need to establish what kind of ethics to manage the robots' behavior [35], and study whether the robot can have any type of social, cultural, moral or legal rights [36]. A scientific team has said that the robot's brain will probably exist by 2019 [37]. Others predict that 2050 will be a breakthrough of intelligent robots [38]. The recent developments have made the robots' behavior more complex [39]. The social impacts of intelligent robots are discussed in a documentary film called Plug & Pray in 2010.

For the development of intelligent robots in the future, it is very important to improve technology and comprehensive application in all directions, including improving the intelligence level of robots, and improving the autonomy and adaptability of intelligent robots.

Simultaneously, intelligent robots for multiple disciplines' cooperative work involve the technology base, and even the psychology, ethics and other social sciences. It is essential to lead the intelligent robots to complete the work beneficial to mankind, and liberate human from the heavy, repeated and dangerous work, as described in the science fiction writer Isaac Asimov's *Three Laws of Robotics*, and to make the intelligent robots truly serve the interests of humanity, not a tool against humanity. It is believed that in the near future, all walks of life will be filled with all kinds of intelligent robots, and scenarios in science fictions will become a reality under the scientists' efforts, which is expected to improve the quality of human life and the ability to explore the unknown.

Intelligent robots' development in China still lags behind the world's advanced level, while the intelligent robot technology is a concentrated reflection of high-tech and has an important development value. Therefore, in the field of intelligent ro-

bots, China should find a clear development target, take feasible development strategy in line with China's national conditions, strive to narrow the gap with the world advanced level, and make the intelligent robot serve roundly for the social development. We believe that the development level of China's intelligent robots can reach a new height through the government's attention and investment, as well as by persistent efforts of scientists and engineers.

The research emphasis of robots is moving from industrial robots to the intelligent service robots. The software and hardware of intelligent robots will be equally important in the next 20 years, and the software will be much more important than hardware 20 years later. The future robots should be a complete intelligent robot with both IQ and EQ.

References

[1] TSK. (2017, Jan. 03). *Depth report in the robotics industry* [Online]. Available: http://sanwen8.cn/p/6299jtY.html
[2] T. Wu. (2015, Aug. 31). *Can China create new century of "made in China"* [Online]. Available: http://www.chinanews.com/it/2015/08-31/7497842.shtml
[3] Machine Design. (2017, Jan. 03). *Robot overview* [Online]. Available: http://www.iw168.cn/jixiesheji/jiqirengaishu
[4] A. Crystal. (2015, October 15). *5 jobs being replaced by robots* [Online]. Available: http://excelle.monster.com/benefits/articles/4983-5-jobs-being-replaced-by-robots
[5] IFR. (2015, Oct. 15). *IFR international federation of robotics* [Online]. Available: http://www.ifr.org/home
[6] IFR. (2015, Oct. 15). *World robotics 2015 industrial robots* [Online]. Available: http://www.ifr.org/industrial-robots/statistics
[7] UNPROFOR. (2016, Nov. 14). *Teach you how to be a lazy in the future* [Online]. Available: http://zjphoto.yinsha.com/bhyszx/10463352.html
[8] IFR. (2015, Oct. 15) *World robotics 2015 service robots* [Online]. Available: http://www.ifr.org/service-robots/statistics
[9] J. Barnard. "Robots in school: games or learning?" Washington, Rep. TR-01-29, 1985.
[10] MCL. (2016, Jan. 03). *Service robot market is in explosive growth, with inestimable developing prospect* [Online]. Available: http://www.robot-china.com/news/201601/26/30920.html
[11] Aiden Burgess. (2015, Sept. 09). *Global industrial robotics market set to rise to $41.17bn by 2020* [Online]. Available: http://www.themanufacturer.com/articles/global-industrial-robotics-market-set-to-rise-to-41-17bn-by-2020
[12] J. Manyika, M. Chui, and J. Bughin, "Disruptive technologies: Advances that will transform life, business, and the global economy," San Francisco, USA, McKinsey Global Institute, Rep., 2013.
[13] T. Collaborative, "From internet to robotics," 2009.
[14] TSK. (2017, Jan. 03). *Automatic intelligence* [Online]. Available: http://baike.baidu.com/item
[15] IFR. (2012, Dec. 27). *Definition of service robots* [Online]. Available: http://www.ifr.org/service-robots
[16] L. Mitgang, "Nova's talking turtle profiles high priest of school computer movement," Gainesville Sun, Rep. TR-10-25, 1983.
[17] F. Ren. (2011). *The REN research lab* [Online]. Available: http://a1-www.is.tokushima-u.ac.jp/member/ren/A1/A1.html
[18] K. N. RoboHon. (2015, Oct. 15). *Cute little robot cum smart phone* [Online]. Available: http://blog.codexify.com/2015/10/robohon-cute-little-robot-cum-smart-phone.html
[19] Y. Myoken, "Research and development for next generation service robots in Japan," Science and Innovation Section British Embassy, 2009.
[20] M. G. Campbell, "*Robots to get their own operating system*," *New Scientist*, vol. 203, no. 2720, pp. 18–19, 2009.
[21] T. McKeough, "The caterpillar self-driving dump truck, " *Fast Company Magazine*, no. 131, pp. 80. Dec./Jan. 2009.
[22] X. Sun, F. Ren, and J. Ye, "Trends detection of flu based on essemble models with emotional factors from social networks", *IEEJ Transactions on Electrical and Electronic Engineering*, to appear in 2017. doi:10.1002/tee.22389.
[23] X. Sun, J. Yi, and F. Ren, "Detecting influenza states based on hybrid model with personal emotional factors from social networks," *Neurocomputing*, vol. 210, pp. 257–268, 2016.
[24] F. Ren and H. Yu, "Role-explicit query extraction and utilization for quantifying user intents," *Information Sciences*, vol. 329, no. 1, pp. 568–580, 2015.
[25] F. Ren and K. Matsumoto, "Semi-automatic creation of youth slang corpus and its application to affective computing", *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 176–189, 2016. doi: 10.1109/TAFFC.2015.2457915.
[26] F. Ren, X. Kang, and C. Quan, "Examining accumulated emotional traits in suicide blogs with an emotion topic model", *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1384–1396, Sept. 2016. doi:10.1109/JBHI.2015.2459683.
[27] F. Ren, Y. Wang, and C. Quan, "TFSM-based dialogue management model framework for affective dialogue systems," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 10, no. 4, pp. 404–410, Jul. 2015. doi: 10.1002/tee.22100.
[28] F. Ren and Y. Wu, "Predicting user-topic opinions in twitter with social and topical context," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 412–424, 2013. doi: 10.1109/T-AFFC.2013.22.
[29] F. Ren and M. G. Sohrab, "Class-indexing-based term weighting for automatic text classification," *Information Sciences*, vol. 236, pp. 109–125, Jul. 2013. doi: 10.1016/j.ins.2013.02.029.
[30] F. Ren and X. Kan, "Employing hierarchical bayesian networks in simple and complex emotion topic analysis," *Computer Speech and Language*, vol. 27, no. 4, pp. 943–968, Jun. 2013. doi: 10.1016/j.csl.2012.07.012.
[31] F. Ren, "From cloud computing to language engineering, affective computing and advanced intelligence," *International Journal of Advanced Intelligence*, vol.2, no.1, pp.1–14, Jul. 2010.
[32] F. Ren, "Affective information processing and recognizing human emotion," *Electronic Notes in Theoretical Computer Science*, vol. 225, pp.39–50, Jan. 2009. doi:10.1016/j.entcs.2008.12.065.
[33] IFR, "Robots today and tomorrow: IFR presents the 2007world robotics statistics survey," World robotics, Oct. 2007.
[34] Reuters. (2007, Dec. 01). *Japan's robots slug it out to be world champ* [Online]. Available: http://www.reuters.com/article/us-robot-fight-idUST32811820071202
[35] AAAI, "Robot ethics," Archive Rep. TR-10-15, 2015
[36] AAAI compilation of articles on robot rights, sources compiled up to2006.
[37] P. Lester. (2009, Jul. 29). *Scientist predicts functional artificial brain in 10 years* [Online]. Available: http://www.gizmag.com/ted-2009-artificial-brain/12362
[38] H. Moravec, *Robot: Mere Machine to Transcendent Mind*. Oxford, UK: Oxford University Press, 2000.
[39] W. Matthew. (2009, Aug. 17). *Robots almost conquering, walking, reading, dancing* [Online]. Available: http://www.koreaittimes.com/story/4668/robots-almost-conquering-walking-reading-dancing

Manuscript received: 2016-07-15

Biographies

REN Fuji (ren@is.tokushima-u.ac.jp) received the PhD degree in 1991 from Faculty of Engineering, Hokkaido University, Japan. He worked at CSK, Japan, where he was a chief researcher of NLP. From 1994 to 2000, he was an associate professor in Faculty of Information Science, Hiroshima City University, Japan. He became a professor in Faculty of Engineering, Tokushima University, Japan in 2001. His research interests include artificial intelligence, language understanding and communication, and affective computing. He is a member of the IEICE, CAAI, IEEJ, IPSJ, JSAI and AAMT, a senior member of IEEE, a fellow of the Japan Federation of Engineering Societies, and the president of International Advanced Information Institute.

SUN Xiao (sunx@hfut.edu.cn) received the double doctorates from University of Tokushima, Japan and Dalian University of Technology, China. He is the director of Affective Computing Institute in Hefei University of Technology, China. His research interests include natural language processing, intelligent man-machine talking, machine translation, and machine learning. He created the natural language processing systems for Chinese. Dr. Sun has published more than 70 papers in the journals and conferences as the first author, including 10 SCI retrieved papers and 30 EI retrieved papers.

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 006
http://www.cnki.net/kcms/detail/34.1294.TN.20161228.1105.002.html, published online December 28, 2016

An Optimization of HTTP/2 for Mobile Applications

DONG Zhenjiang[1], SHUANG Kai[2], CAI Yanan[2], WANG Wei[1], and LI Congbing[1]

(1. Cloud Computing & IT Research Institute, ZTE Corporation, Nanjing 210012, China;
2. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract

In recent years, Hyper Text Transfer Protocol (HTTP) spreads quickly and steadily in the usage of mobile applications as a common web protocol, so that the mobile applications can also benefit from HTTP/2, which is the new version of HTTP based on SPDY developed by Google to speed up the Internet transmission speed. HTTP/2 enables a more efficient use of network resources and a reduced perception of latency by introducing header field compression and allowing multiple concurrent exchanges on the same connection. However, what HTTP/2 focuses on is visiting websites through a browser, and mobile applications are not considered much. In this paper, firstly, mobile applications are classified based on the data flow characteristics. Based on the classification, we propose an optimization of HTTP/2 for mobile applications, called HTTP/2-Advance, which uses multiple Transmission Control Protocol (TCP) connections to multiplex HTTP requests and responses. Then we build a tiny system which simulates actual requests and responses between mobile applications and servers. We figure out the best choice of the number of multiple TCP connections for mobile applications, and compare the performance of HTTP, HTTP/2 and HTTP/2-Advance in both simulated and in-situ experiments in our system.

Keywords

HTTP/2; HTTP optimization; multiple connection; header compression

This work was supported in part by ZTE Industry-Academia-Research Cooperation Funds.

1 Introduction

In recent years, with the growing popularity of smart phones and the rapid prosperous development of mobile Internet, mobile applications present an explosive growth. According to the data provided by Google Play Store, there were more than 370,000 applications available for the Android pads and phones in 2011, and the number rose sharply to 1.5 million in 2014 [1]. Meanwhile, because of an increasing trend on binding mobile applications with rendering engine and a widely acceptance on emerging frameworks like Hybrid [2], Hyper Text Transfer Protocol (HTTP) spreads quickly and steadily in the usage of mobile applications as a common web protocol. However, since the Internet has changed dramatically since 1990 when HTTP was first published, HTTP cannot hold the needs of users gradually.

The optimization of HTTP is imminent. HTTP 1.0 serially builds on Transmission Control Protocol (TCP) connection for each embedded objectives. When requesting a web page with a browser, the TCP 3-way/4-way handshake and slow start significantly increase the latency. To optimize this issue, HTTP1.1 uses multiple, persistent TCP connections to "keep-alive". However, it is still a serial prototype in essence and the multiple connections strategy may cause more bandwidth usage. HTTP/2 [3], [4] is the second major version of HTTP to make web faster. It is based on SPDY [5], which is developed by Google. According to W3Techs, 1.2% of all websites support HTTP/2 in August 2015 to speed up the Internet transmission speed [6].

HTTP/2 enables a more efficient use of network resources and a reduced perception of latency by introducing header field compression and allowing multiple concurrent exchanges on the same connection [3]. This means less competition with other flows and longer-lived connections, which in turn leads to better utilization of available network capacity. As a result, the web browser loads pages more quickly. However, for the mobile, it is much different. The browser is just a very small part of the mobile applications that use HTTP. Besides, the unpredictable network circumstance and data flow characteristics brought about by different applications have great impact on the Internet transmission speed. For example, applications like UC browser have intermittent network connections with diverse connection time and data flow throughout the day depending on the usage patterns, while other applications such as Sina Weibo have persistent network connections to transfer the signaling and user requested resources with light data flow. In this paper, we proposed an optimization of HTTP/2 for the mobile applications, which enables a more efficient way to use HTTP/2 in mobile application scenarios. The main contributions of this paper are: 1) Mobile applications are classified based on the data flow characteristics; 2) an optimization of HTTP/2 for mobile applications, called HTTP/2-Advance, is proposed; 3) a tiny system is built to simulate the actual re-

quests and responses between applications and servers, and 4) the performance of this optimization is tested and analyzed in different categories of applications and different network conditions in both simulated and in-situ experiments.

2 Background and Related Works

2.1 Background

HTTP/2 allows the interleaving of request and response message exchanges on the same connection associated with its own stream. The streams are largely independent of each other; in this way, a blocked or stalled request or response does not prevent any progress on other streams. It can reuse the TCP connections to reduce the number of TCP connections and the latency.

An HTTP header has some same information for a single session, for instance, host and user-agent. HTTP/2 provides the header compression strategy to reduce the unnecessary redundant information, and then decrease the TCP bytes and application layer latency. It is preferable to use no more than two levels of subheadings (primary and secondary), and the levels should be carefully differentiated. Ideally we try not to go below a second level subheading. It should be noted that there is no period after the final number.

2.2 Related Works

Some works have been done in PC since SPDY (which HTTP/2 is based on) was published. The performance of SPDY and HTTP in a variety of settings in publicly available software was tested in [7]. Web page load time under SPDY and HTTP was systematically compared in [8]. A comprehensive evaluation of SPDY's performance was experimented in [9] to find the potential benefits of SPDY. However, all of them focus on the PC side, and are limited by the existing SPDY deployments provided by some famous web service providers.

Some works have been done in the mobile side. J. Khalid et al. [10] proposed two adopting mechanisms to dynamically adjust the overall performance. However, this solution did not evaluate them in the mobile circumstance. J. Erman et al. [11] provided a detailed analysis of the performance of both HTTP and SPDY, and of their interaction with various layers as well. However, only the web pages on mobile browsers in cellular networks have been analyzed. G. Mineki et al. [12] proposed a SPDY accelerator that could considerably accelerate the web access speed by combining the SPDY protocol and cache system, but it only evaluates the performance of web pages, too. So far, none of the current works have examined the performance of SPDY on mobile applications.

In this paper, an optimization of HTTP/2 for mobile applications is proposed, which is called HTTP/2-Advance. Both the simulated experiment and in-situ experiment are also conducted to compare the performance of HTTP, HTTP/2 and HTTP/2-Advance.

3 Application Taxonomy

The taxonomy is based on the data flow characteristics of an application during its usage time. According to the Android operating system, only one app can be shown to the user on the screen, while other applications are only allowed to run in background. The reason we consider app usage time rather than user behavior of multiple applications is that every factor of the user behavior towards multiple applications, such as location, time, user needs, and certain context can be divided into the data flow characteristic of application and the network condition during the app usage time. The data flow characteristic will be introduced in this section, and the network condition influence will be discussed in next section.

Mobile applications are divided into four categories in this paper, and for each category, one Chinese popular app is selected as the test object. The categories are shown in **Table 1**.

The data flow characteristics of these applications are shown as follows.

3.1 Intermittent Connection with Random Data Flow

The application has intermittent connections with diverse connection time, and data flow throughout the day is difficult to predict since it largely depends on the usage pattern. The mobile browser UC Browser is one representative that occupies 41.7% of the Chinese mobile browser market [17].

3.2 Intermittent Connection with Light Data Flow

The application uses intermittent connections to download the resources with diverse connection time. Data flow is generated only when the application is downloading resources. After downloading, the application hardly consumes network resources. One example is Southern Weekly, an online application made by a traditional media company.

3.3 Persistent Connection with Heavy Data Flow

Xunlei Kankan is a popular online video application. Its data flow is heavy, and may occupy the independent channel resources of persistent connections for a long time.

3.4 Persistent Connection with Light Data Flow

The application has persistent network connections to transfer constantly switching of state signaling and download user

▼**Table 1. Application taxonomy based on data flow characteristic**

Application taxonomy	Popular application
Intermittent connection with random data flow	UC Browser [13]
Intermittent connection with light data flow	Southern Weekly [14]
Persistent connection with heavy data flow	Xunlei Kankan [15]
Persistent connection with light data flow	Sina Weibo [16]

requested resources, but the data flow is light. Sina Weibo, the most popular social network app in China, is such an application.

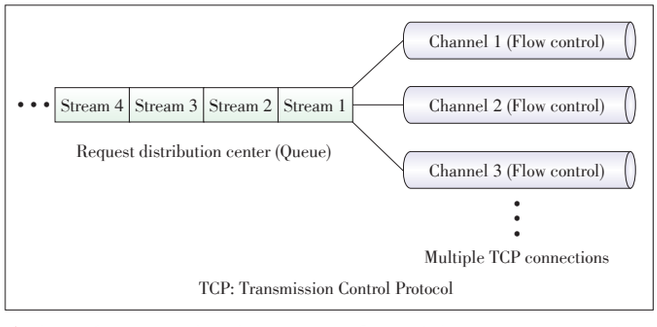
4 Optimization of HTTP/2

HTTP/2 allows interleaving of request and response messages exchange on the same connection associated with its own stream. The client sends an HTTP request on a new stream using a previously unused stream identifier. The server sends an HTTP response on the same stream as the request [6]. HTTP/2 uses multiplexing connection, and since using a single connection, it does not monopolize network resources. This means less competition with other flows and longer-lived connections, which in turn leads to better utilization of available network capacity. As a result, the web browser will load pages more quickly.

However, what this strategy focuses on is visiting websites through a PC browser, and mobile applications are not considered. According to Android operating system, only one application can be shown to the user on the screen, so it is reasonable to assume that there is only one application being able to access to network at the same time. Different from the PC scenario, the network resources can be monopolized by one mobile application. In this case, only using a single TCP connection to interleave HTTP requests and responses is not the best choice, because it cannot make full use of network capacity. Therefore, we propose an optimization strategy of HTTP/2 called HTTP/2-Advance, which uses multiple TCP connections to multiplex HTTP requests and responses. That means that the web elements are loaded in parallel over several TCP connections. The multiplexing strategy (**Fig. 1**) is as follows:

1) The multiple TCP connections between the client and the server are declared as Channel 1, Channel 2, ⋯ , Channel *N*, and these channels are indiscriminate with each other from the HTTP request-response stream perspective.
2) Each HTTP request-response stream is allocated a unique stream ID and put into the request distribution center, which is an orderly queue in essence, by the generation time of each request.
3) The request distribution center distributes the HTTP request-response streams to the multiple TCP connection channels based on the round-robin strategy. Meanwhile, a flow-control [3] scheme ensures that the streams on the same connection do not destructively interfere with each other so that blocked streams are prevented.

The multiple TCP connections provide a parallel data transmission channel, but simultaneously bring more bandwidth consumption and more HTTP/2 multiplexing connection processing time when using our optimization. This is not a "the more, the better" scenario and there should be a tradeoff for each application category. If the application just has light data flow with intermittent connections to transfer, the benefit brought from multiple TCP connections may not cover the process time for that. On the other hand, if the data flow is heavy, the multiple TCP connections benefit the transmission speed more. To figure out how many multiple TCP connections are best for each application category, some experiments have been done, which will be introduced in the next section.

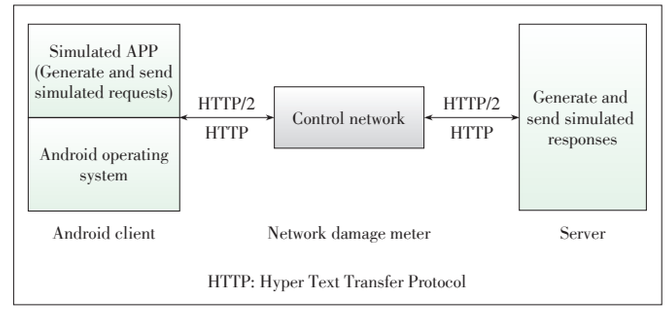


▲**Figure 1. The multiplexing strategy of HTTP-Advance. The HTTP request-response streams are put into the request distribution center by the sequence of requesting time, and distributed to the multiple channels based on round-robin.**

5 Experiments and Evaluation

5.1 Experimental Setup

Since HTTP/2 is a new protocol, almost none of the mobile applications have been built with a HTTP/2 protocol stack, as well as the application servers in China. To simulate actual requests and responses between the applications and servers, we build a tiny system. This system is made up of an Android client with a simulated application, a server, and a network damage meter [18] between the client and the server. The simulated application can generate and send both simulated HTTP and HTTP/2 requests. The server is used to generate and send simulated responses for both HTTP and HTTP/2 requests in each protocol stack. The number of multiplexing connections are alternative between the client and the server. The network damage meter controls the network condition, such like the latency and the packet loss rate. The structure of the system is shown in **Fig. 2**. The Android phone used as the client is Nex-



▲**Figure 2. The simulation system architecture. The client is connected to the server via a network damage meter that controls the network condition.**

us 5 with a 2.3 GHz CPU, 2G RAM, and Android 4.4 OS. The server is a Dell T5810 computer with 3.5 GHz CPU and 32G RAM.

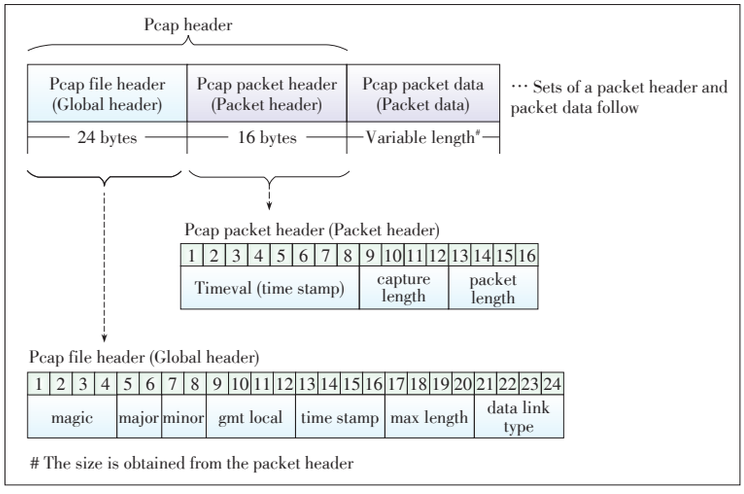
For rigorous controlled trials, simulated user requests based on the actual requests are generated and sent from the Android client to the server many times to compare the performance of HTTP, HTTP/2 and HTTP/2-Advance statistically. Simulated responses based on the actual responses are generated by the server to respond with the client requests. At the same time, network condition is controlled by the parameter configuration on the network damage meter. In order to produce authentic simulated requests and responses:

Firstly, for each application category, application requests are classified into several behavior categories based on our user research, and the packets of actual requests and responses of each behavior category are captured by the Shark [19]. The Shark then generates .pacp files (**Fig. 3**). The pcap packet data stores the captured packet, and pcap packet header stores the time stamp and length of the captured packet. By analyzing the pcap files, we get the time stamp and length from the pcap packet header, and requests and responses information from pcap packet data.

Secondly, the behavior configuration files are generated by analyzing the pcap files of each behavior category. By parsing the pacp file, each realistic request-response is analyzed to get the data of index, request time, dependent resources, GET/POST, URL, header length and body length. Based on these data, the simulated application and server can simulate the actual requests and responses.

Thirdly, the behavior configuration files are grouped to generate the group file. The grouping is based on our user research of user habits when using the application.

Lastly, the simulated application and server send requests and responses based on the group file and the behavior configuration files to guarantee the similarity of the traffic characteristic between the simulated packets and the actual ones.



▲**Figure 3. Pcap file format. Pcap packet data stores the captured packet, and the pcap packet header stores the time stamp and length of the captured packet.**

With the example of Sina Weibo, requests are firstly classified into five behavior categories, that is, login, refresh, read, post, and repost/comment, and actual requests and responses packets of each behavior are then captured by the Shark. Secondly, the behavior configuration files are generated as login file, refresh file, read file, post file, and repost/comment file. Thirdly, based on the research of user habits when using Sina Weibo, the group file is generated by grouping the behavior configuration files. Lastly, the simulated packets based on the group and configuration files are sent by the client and the server.

The network condition is parameterized by the latency and the packet loss rate. The latency and packet loss rate parameters based on the mobile access network like 2G, 3G, 4G and Wi-Fi/802.11 are considered to simulate different network conditions [20]. They are parameterized as (0 ms, 0%), (100 ms, 0%), (300 ms, 0%), (0 ms, 2%), (100 ms, 2%), and (300 ms, 2%). At the same time, the application layer latency $L_{app\ layer}$ in (1), application layer throughput $TP_{app\ layer}$ in (2) and (3), TCP total bytes B_{TCP}, and the total number of TCP connections $\#_{TCP\ conn}$ are recorded to evaluate the performance of HTTP, HTTP/2 and HTTP/2-Advance in different scenarios. The calculation methods of the comparison standards are as follows:

$$L_{app\ layer} = t_{last\ 200\ ok} - t_{1st\ GET/POST} \quad (1)$$

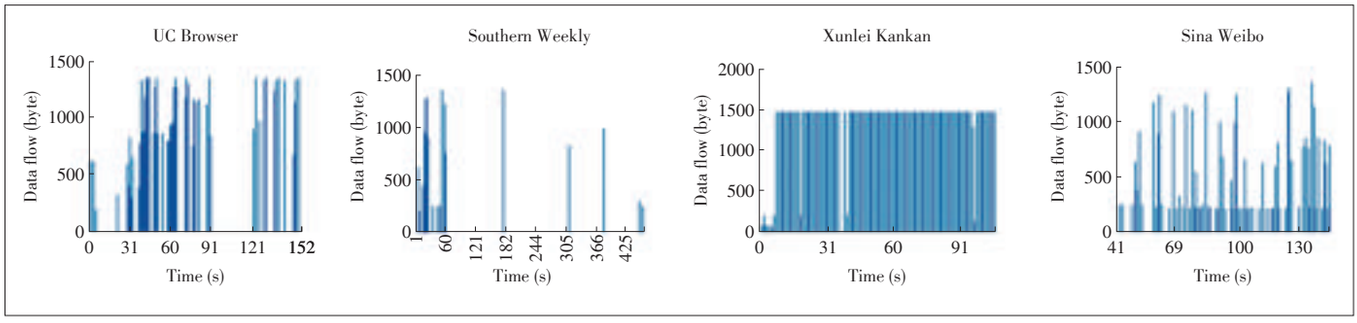
$$TP_{app\ layer} = B_{TCP\ payload} / L_{app\ layer} \quad (2)$$

$$B_{TCP\ payload} = B_{TCP} - B_{TCP\ header} - B_{conn\ control} \quad (3)$$

Since HTTP and HTTP/2 are application layer protocols, the application layer latency $L_{app\ layer}$ and the application layer throughput $TP_{app\ layer}$ are the most important indicators to measure the optimization proportion. The application layer latency is a main standard of user experience, and the application layer throughput reflects network efficiency and usage. The application layer latency $L_{app\ layer}$ equals to the HTTP session duration, from the first sent GET/POST request timestamp $t_{1st\ GET/POST}$ to the last received 200 OK response timestamp $t_{last\ 200\ ok}$. The application layer throughput $TP_{app\ layer}$ equals to the TCP payload bytes $B_{TCP\ payload}$ divided by the application layer latency $L_{app\ layer}$. The TCP payload bytes $B_{TCP\ payload}$ as in (3), equals to the TCP total bytes B_{TCP} decrement the TCP header bytes $B_{TCP\ header}$ and the TCP connection establishment/close control bytes $B_{conn\ control}$ (such as SYN+ACK).

5.2 Data Flow Characteristics of Application Categories

As described in Section 3, applications are divided into four categories based on the data flow characteristic. Before the test, data flow characteristics of each applica-



▲Figure 4. Data flow characteristics of application categories. Different application categories have different data flow characteristics, which influences the performance.

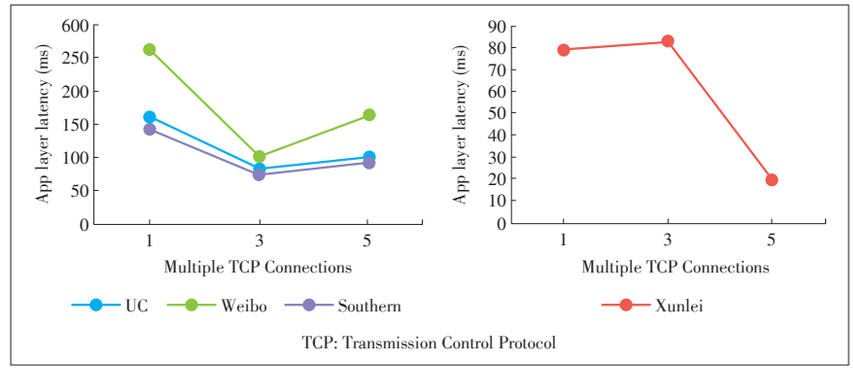
tion category are measured to have more in-depth understanding of the application category. The characteristics are shown in **Fig. 4**, from which we can find that, UC Browser has intermittent network connections with diverse connection time, and the data flow is random. Southern Weekly uses intermittent connections to download with diverse connection time and light data flow. Xunlei Kankan has persistent connection with heavy data flow all the time, and Sina Weibo has persistent connection like Xunlei Kankan, but the data flow is light.

We also find that most of these applications in China use HTTP1.0 (short connection) to request the resources. For every HTTP request-response, the applications serially established multiple TCP connections to improve the resources loading time. Therefore, in our later simulation and experiment, HTTP1.0 (short connection) will be used to evaluate the performance of HTTP, HTTP/2 and HTTP/2-Advance.

5.3 Simulation Results

5.3.1 Number of Connections

The performance of HTTP/2-Advance, using multiple TCP connections to multiplex HTTP requests and responses, is tested and analyzed to figure out how many multiple TCP connections are best for each application category. Application layer latency are measured in single connection, three multiple TCP connections, and five multiple TCP connections (**Fig. 5**). UC Browser and Southern Weekly donnot always have a packet to transfer, and the packets are not big since the web page optimization for mobile phone displaying. Sina Weibo has constantly state switch signaling and heartbeat packet, but the packet is also small. UC Browser, Sina Weibo and Southern Weekly perform best at three multiple TCP connections. Xunlei Kankan has very different data flow characteristics from others. Because its data flow is always heavy and occupying the independent channel resources for a long time, it is not strange that the latency is drastically reduced when three multiple TCP connections is changed into five. In a total, the HTTP/2-Advance enables a more efficient way to use HTTP/2 in mobile applications, and multiplexing three multiple TCP connections is feasible in most cases.



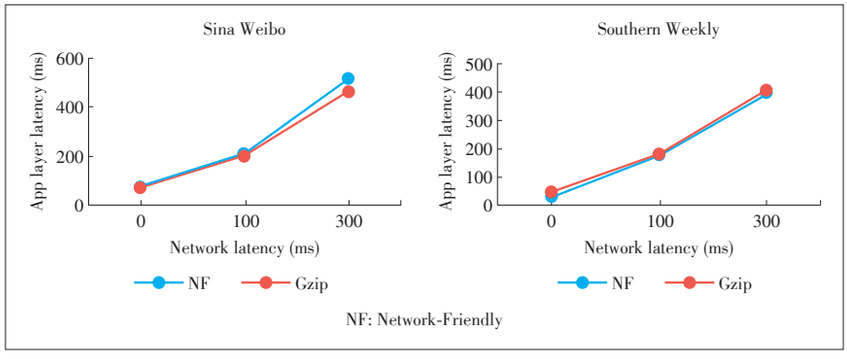
▲Figure 5. The number of multiple TCP connections for each category. More connections increase connection bandwidth and processing time, and a tradeoff is necessary when using HTTP/2-Advance.

5.3.2 Header Compression Strategy

The header compression strategy can reduce redundant information in the HTTP header and further the TCP bytes and application layer latency. The application layer latency is measured in two compression strategies that are Network-Friendly [21] and Gzip [22]. As shown in **Fig. 6,** the difference of header compression strategies has little influence on the application layer latency. The HTTP header occupies a major part of the request packet, but in the response packet, the HTTP header is usually much smaller than HTTP body content. Therefore, in a pair of request-response, the header size of the whole HTTP packets is so small that different header compression strategies have little influence on latency. We finally choose Gzip for our later evaluation.

5.3.3 Evaluation of HTTP/2-Advance

The performance of HTTP, HTTP/2 and HTTP/2-Advance in different network condition are tested and compared with three TCP connections as 3 and Gzip header compression. The results are shown in **Figs. 7** and **8**.



▲Figure 6. Application layer latency in different compression strategy. Difference of compression strategy has little influence on app layer latency.

HTTP/2-Advance performs better than HTTP/2, especially when the network condition is good. The average HTTP/2-Advance optimization proportion of HTTP on application layer latency is 41.67%. The network resources can be better used by multiple TCP connections when the network condition is good, however, these multiple TCP connections will be in conflict while the network condition becomes bad. Even though the advantage of HTTP/2-Advance decreases, HTTP/2-Advance still preforms better.

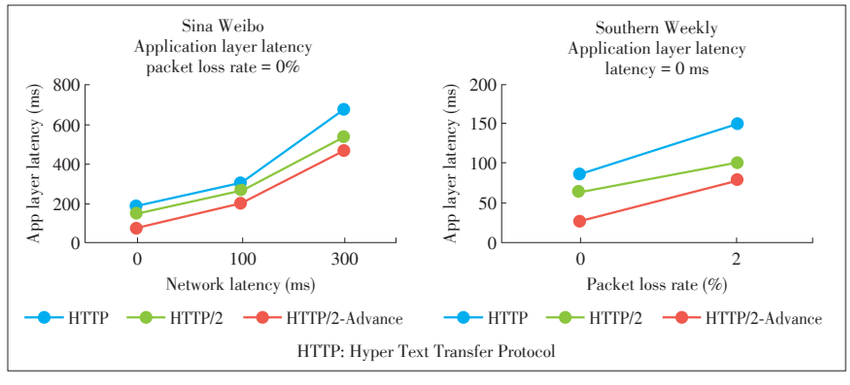
The application layer latency increases when the latency or the packet loss rate increases no matter HTTP, HTTP/2 or HTTP/2-Advance is used. The network latency directly affects the data transmission latency. A larger latency will increase the data transferring time and then increase the application layer request completion time with other conditions remaining unchanged. If latency is a fixed number, a higher packet loss rate stands a higher occurrence of incomplete data, and this relation also exists in retransmission scenario that will cause the increase of the application layer latency.

The application layer throughput declines when latency increases. As $TP_{app\ layer}$ and $B_{TCP\ payload}$ hardly change when using HTTP, HTTP/2 or HTTP/2-Advance, it is obvious that the application layer throughput only depends on the application layer latency, and is inversely proportional to the application layer latency in theory (Fig. 8).

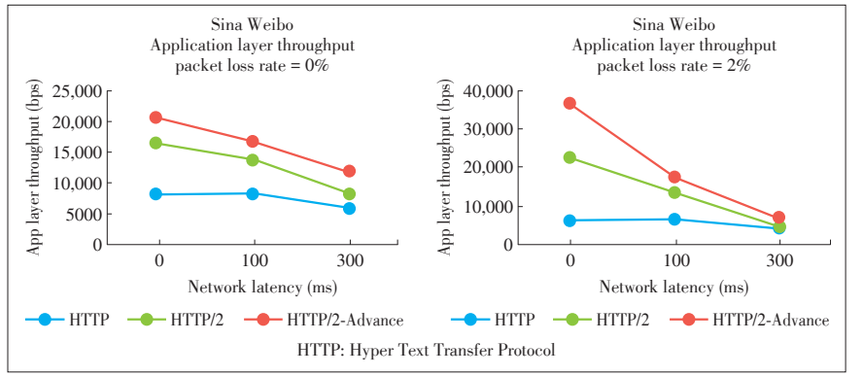
The main reason of the reduced application layer latency using HTTP/2-Advance is that HTTP/2-Advance uses multiplexing connection, which reduces the total number of TCP connections and the cost of TCP connection establishment/close 3-way/4-way handshakes. During the test, the number of total TCP connections and the TCP bytes are recorded (**Tables 2** and **3**). The average reducing bytes of Southern Weekly is 800.44 bytes per connection, and that of UC Browser is 663.92 bytes per connection. Due to the multiplexing connection strategy, the HTTP packet can reuse the previous established TCP connections. As we all know, the establishment of a TCP connection needs three-way handshake, and the close of the connection needs four-way handshake. Theoretically, the connection establishment/close control bytes $B_{conn\ control}$ that is required by TCP establishment and close handshakes are in (4), where n means the total number of TCP connections $\#_{TCP\ conn}$.

$$B_{conn\ control}=\sum_{i=1}^{n}\left(\sum_{j=1}^{3}establi\,\mathrm{sh}_{i,j}+\sum_{k=1}^{4}close_{i,k}\right)\approx$$
$$(74+74+66+66+66+60+60)*n=$$
$$=466n\left(n=\#_{TCPconn}\right) \quad (4)$$

The actual data shown in Tables 2 and 3 are consistent with the magnitude of the theoretical data. Due to the unpredictable factors such as retransmission and HTTP/2 header compression strategy, a little more bytes reduction than the theoretical



▲Figure 7. HTTP/2-Advance preforms better than HTTP/2 and HTTP according to the application layer latency.



▲Figure 8. HTTP/2 and HTTP/2-Advance are both optimize the app layer throughput, and the app layer throughput is inversely proportional to the app layer latency in theory.

▼Table 2. TCP bytes and the total number of TCP connections—Southern Weekly

	Latency	0 ms		100 ms		300 ms	
Packet loss rate is 0%	Protocol	HTTP	HTTP/2-Advance	HTTP	HTTP/2-Advance	HTTP	HTTP/2-Advance
	B_{TCP} (bytes)	186,650	115,712	167,230	100,034	269,855	138,575
	$\#_{TCP\ conn}$	120	3	105	4	146	3
Packet loss rate is 2%	Protocol	HTTP	HTTP/2-Advance	HTTP	HTTP/2-Advance	HTTP	HTTP/2-Advance
	B_{TCP} (bytes)	224,105	144,908	215,676	107,392	280,634	152,482
	$\#_{TCP\ conn}$	121	3	106	3	146	3

HTTP: Hyper Text Transfer Protocol TCP: Transmission Control Protocol

▼Table 3. TCP bytes and the total number of TCP connections—UC Browser

	Latency	0 ms		100 ms		300 ms	
Packet loss rate is 0%	Protocol	HTTP	HTTP/2-Advance	HTTP	HTTP/2-Advance	HTTP	HTTP/2-Advance
	B_{TCP} (bytes)	260,353	165,797	322,514	194,722	292,336	172,530
	$\#_{TCP\ conn}$	186	3	220	4	192	4
Packet loss rate is 2%	Protocol	HTTP	HTTP/2-Advance	HTTP	HTTP/2-Advance	HTTP	HTTP/2-Advance
	B_{TCP} (bytes)	293,155	181,261	365,463	194,722	328,536	181,378
	$\#_{TCP\ conn}$	187	3	219	4	192	3

HTTP: Hyper Text Transfer Protocol TCP: Transmission Control Protocol

data is understandable. Furthermore, besides Southern Weekly and UC Browser, the other application categories also show the same relation.

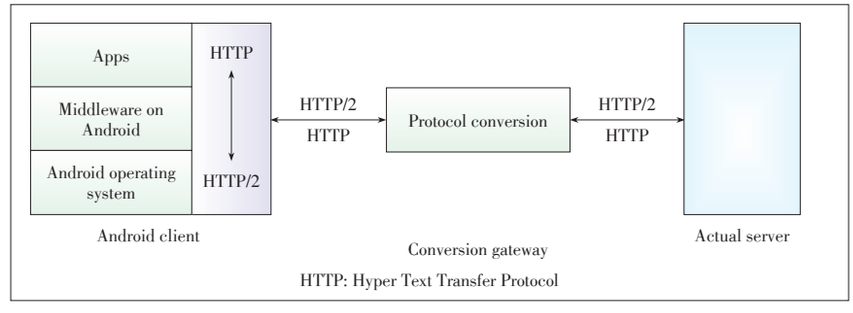
In a total, HTTP/2-Advance performs better than HTTP/2, especially when the network condition is good. HTTP/2-Advance optimizes the application layer latency and throughput and it does reduce the TCP connections and TCP bytes thanks to the multiplexing connection and the header compression strategy.

5.4 In-Situ Experiment

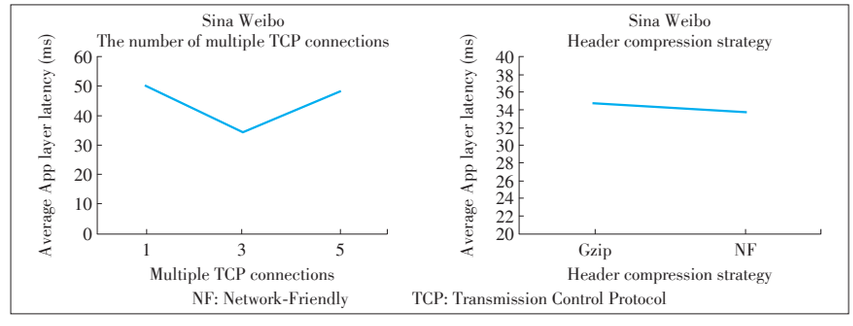
A tiny system is built to test the performance of HTTP-Advance in actual situation (**Fig. 9**). An Android middleware is developed for changing HTTP stack to HTTP/2-Advance stack on the client side. The middleware runs "under" the applications and allows the real applications to send HTTP/2-Advance packet indirectly. Placed in between the actual server and client, a protocol conversion gateway is used to convert the protocol HTTP and HTTP/2-Advance. The server is an actual application server that responds requests from the client. The tests of Sina Weibo and UC Browser were done by 20 persons at different time and different locations. The test results are shown in **Fig. 10** and **Table 4**, from which we can find that the in-situ experiment has the same conclusion with the simulated experiment. HTTP/2-Advance performs better than HTTP/2 when using three multiple TCP connections to multiplex HTTP requests and responses, and the average application layer latency has minute difference between Gzip and Network-Friendly header compression strategies. When the multiple TCP connections is three and the Gzip header compression is usedi, the average HTTP/2-Advance optimization proportion of Sina Weibo is 25.86% and that of UC Browser is 48.39%.

6 Conclusions

In this paper, popular mobile applications are classified by the data flow characteristics. Then, an optimization of HTTP/2 for mobile applications, called HTTP/2-Advance, is proposed for the scenario of mobile applications. Different applications are also simulated and tested in several network circumstances parameterized by the latency and the packet loss rate. By changing the header compressions strategies and the network conditions, the comparisons of HTTP, HTTP/2 and HTTP/2-Advance in different scenarios



▲Figure 9. The actual experiment circumstance. The client is connected to the server via a conversion gateway that converts HTTP/2 and HTTP.



▲Figure 10. The in-situ experiment on the number of multiple TCP connections influence and header compression influence. The in-situ result has the same conclusion with the simulated experiment.

▼Table 4. Application layer latency in in-situ experiment

	Protocols	Mobile applications	
		Sina Weibo	UC Browser
Average app layer latency (ms)	HTTP	57.6983	55.37302
	HTTP/2-Advance	42.7775	28.57672
Average optimization proportion		25.86%	48.39%

HTTP: Hyper Text Transfer Protocol TCP: Transmission Control Protocol

are tested and analyzed. The results indicate that the HTTP/2-Advance enables a more efficient way than HTTP/2 in mobile applications, and it performs better than HTTP/2, especially when the network condition is good. Overall, multiplexing three TCP connections is feasible for most mobile applications.

References

[1] Google Play. (2016). *Google Play for Android Applications* [Online]. Available: http://play.google.com
[2] *Hybird* [Online]. Available: http://www.hybird.org
[3] M. Belshe, M. Thomson, and R. Peon. (2015, May). *Hypertext Transfer Protocol Version 2 (HTTP/2)* [Online]. Available: http://tools.ietf.org/html/rfc7540
[4] M. Thomson (ed.), M. Belshe, and R. Peon. (2015, Feb. 11). *Hypertext Transfer Protocol version 2 - draft-ietf-httpbis-http2-16* [Online]. Available: https://tools.ietf.org/html/draft-ietf-httpbis-http2-16
[5] M. Belshe, and R. Peon. (2013, Nov.). *SPDY protocol—Draft 3.1* [Online]. Available: http://www. chromium. org/spdy/spdyprotocol/spdy-protocol-draft3-1
[6] W3Techs. (2015, Jul.). *Usage of HTTP/2 for websites* [Online]. Available: http://w3techs.com/technologies/details/ce-http2/all/all
[7] J. Padhye and H. F. Nielsen, "A comparison of SPDY and HTTP performance," Microsoft Technical Report MSR-TR-2012-102, 2012.
[8] X. S. Wang, A. Balasubramanian, A. Krishnamurthy, and D. Wetherall, "How speedy is SPDY," in *Proc. 11th USENIX Symposium on Networked Systems Design and Implementation (NSDI'14)*, Seattle, USA, 2014.
[9] Y. Elkhatib, G. Tyson, and M. Welzl. "Can SPDY really make the web faster?" in *IFIP Networking Conference*, Trondheim, Norway, Jun. 2014. doi: 10.1109/IFIPNetworking.2014.6857089.
[10] J. Khalid, S. Agarwal, A. Akella, and J. Padhye. (2016). *Improving the performance of SPDY for mobile devices* [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/sagarwal-hotmobile15-poster.pdf
[11] J. Erman, V. Gopalakrishnan, R. Jana, and K. K. Ramakrishnan, "Towards a SPDY'ier mobile web?" in *Proc. Ninth ACM Conference on Emerging Networking Experiments and Technologies*, Santa Barbara, USA, Dec. 2013, pp. 303–314. doi: 10.1145/2535372.2535399.
[12] G. Mineki, S. Uemura, and T. Hasegawa, "SPDY accelerator for improving web access speed," in *15th International Conference on Advanced Communication Technology (ICACT)*, Pyeong Chang, Korea (South), Jan. 2013, pp. 540–544.
[13] Google Play. (2016). *UC Browser—Google Play for Android Applications* [Online]. Available: https://play.google.com/store/apps/details?id=com.UCMobile
[14] Google Play. (2016). *Nanfang Weekend - Google Play for Android Applications* [Online]. Available: https://play.google.com/store/apps/details?id=net.coollet.infzmreader
[15] Google Play. (2016). *Xunlei Kankan—Google Play for Android Applications* [Online]. Available: play.google.com/store/applications/developer?id=迅雷看看&hl=zh_TW
[16] Google Play. (2016). *Sina Weibo—Google Play for Android Applications* [Online]. Available: play.google.com/store/applications/details?id=com.sina.weibo
[17] Xinhua. (2014, May). *Data of phone browsers in Q1 2014* [Online]. Available: http://www.bj.xinhuanet.com/hbpd/hbrj/rjy/2014-05/16/c_1110729487.htm
[18] WANem. (2016). *The Wide Area Network emulator* [Online]. Available: http://wanem.sourceforge.net
[19] Google Play. (2016). *Shark - Google Play for Android Applications* [Online]. Available: http://play.google.com/store/applications/details?id=lv.n3o.shark
[20] I. Grigorik, "Making the web faster with HTTP 2.0," *Communications of the ACM*, vol. 56, no.12, pp. 42–49, Dec. 2013. doi: 10.1145/2534706.2534721.
[21] W. Tarreau, A. Jeffries, and A. de Croy. (2012, Mar.). *Proposal for a Network-Friendly HTTP Upgrade* [Online]. Available: https://www.ietf.org/archive/id/draft-tarreau-httpbis-network-friendly-00.txt
[22] Gzip. (2016). *The gizp home page* [Online]. Available: http://www.gzip.org

Manuscript received: 2015-07-12

Biographies

DONG Zhenjiang (dong.zhenjiang@zte.com.cn) received his MS degree in telecommunication from Harbin Institute of Technology, China in 1996. He is vice president of the Cloud Computing & IT Research Institute of ZTE Corporation. His main research areas are cloud computing, big data, new media, and mobile Internet technologies.

SHUANG Kai (shuangk@bupt.edu.cn) received his PhD from State Key Laboratory of Networking & Switching Technology, Beijing University of Posts & Telecommunications (BUPT) in 2006. He is currently an associate professor with the BUPT. His research interests include cloud computing and the mobile Internet.

CAI Yanan (522018144@qq.com) is a master candidate of Beijing University of Posts and Telecommunications, China. Her research direction is the mobile Internet.

WANG Wei (wang.wei8@zte.com.cn) received her MS degree from Nanjing University of Aeronautics and Astronautics, China. She is an engineer and project manager in the field of mobile Internet at the Cloud Computing and IT Research Institute of ZTE Corporation. Her research interests include new mobile Internet services and applications, PaaS, terminal application development, and other technologies. She has authored five academic papers.

LI Congbing (li.congbing@zte.com.cn) received his MS degree from Nanjing University of Science and Technology, China. He is currently a technical researcher in the field of mobile Internet and AI at the Service Research Institute of ZTE Corporation. His research interests include WebRTC, HTML5, open platform, and robot AI. He is responsible for researching mobile Internet IM services and mobile robot navigation.

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 007
http://www.cnki.net/kcms/detail/34.1294.TN.20161229.1035.002.html, published online December 29, 2016

One Step Hologram Calculation for Holographic 3D Display Based on Nonuniform Sampled Angular Spectrum Method

CHANG Chenliang[1], XIA Jun[2], and LEI Wei[2]
(1. Department of Physics, Nanjing Normal University, Nanjing 210023, China;
2. School of Electronic Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract

We proposed a method for calculating the computer generated hologram from multi-plane 3D objects by using nonuniform sampled angular spectrum method (NUASM). Both of the hologram plane and the image plane are nonuniform sampled according to the distances and positions of the three-dimensional objects. The nonuniform fast Fourier transform (NUFFT) is used to calculate the angular spectrum propagation from the image plane to the hologram plane and the hologram can be calculated in only one step. Simulation and optical experiment results show that the hologram generated in this way can reconstruct objects on multi-planes simultaneously and separately without axial distortion.

Keywords

holographic display; computer holography; nonuniform fast Fourier transform

This work is supported by Major State Basic Research Development Program of China under Grant No. 2013CB328803, National High Technology Research and Development Program of China under Grant No. 2012AA03A302 and No. 2013AA013904, and ZTE Industry-Academia-Research Cooperation Funds.

1 Introduction

In the field of holographic three dimensional displays, the calculation of computer generated hologram is a big problem for its large amount of points and long calculation time. In order to speed up the calculation, many methods have been proposed, such as the look-up table method [1], [2], the polygon-based method [3], [4] and the wavefront recording plate method [5], [6]. These methods are very useful with those complicated objects generated by computer graphics software, but not suitable for objects with large depth. For some simple situations of hologram calculation, the objects can be divided into several planes and the hologram can be calculated by the diffraction from multi-planes [7], in which the fast Fourier transform (FFT) algorithm can be used to accelerate the calculation between parallel planes.

In the calculation of hologram for multi-plane objects, the diffraction between parallel planes can usually be calculated by either the Fresnel diffraction or the angular spectrum method [8]–[12]. For the Fresnel diffraction, there are two main problems. First, in order to obtain the light distribution in the hologram plane, the diffraction should be calculated from each plane to the hologram separately by many times. This even spends more time when it combined with the iteration algorithm (such as GS iteration) for the kinoform generation. Second, due to the restriction of the sampling rate in the calculation by sampling theory, there is axial distortion in the reconstruction, which will cause the deformation of the objects. For the angular spectrum method, there is no axial distortion in the reconstruction because the sampling rates in both planes are the same in the calculation. However, it still needs many times of diffraction calculation from each plane to the hologram. In this paper, we proposed a method to calculate the hologram from multi-plane objects based on the nonuniform sampled angular spectrum method and the diffraction calculation from objects to hologram is calculated in only one step instead of many steps in the conventional method.

2 Nonuniform Sampled Angular Spectrum Method

2.1 Basic Principle for Changing the Sampling Rate in Angular Spectrum

The angular spectrum method is usually used to calculate the diffraction between parallel planes with the same sampling rate in near field. For example, the calculation from hologram plane to image plane using angular spectrum is expressed as

$$f(x)=FFT^{-1}\left[FFT\left(h(u)\right)\cdot \exp\left(i2\pi z\sqrt{\frac{1}{\lambda^2}-f^2}\right)\right], \quad (1)$$

where FFT and FFT^{-1} represent the Fourier transform and in-

verse Fourier transform, $h(u)$ and $f(x)$ are the distribution on hologram and image planes. z is the distance between two planes and l is the wavelength. f is the spectrum coordinate in the Fourier domain. The calculation of angular spectrum is composited of two FFTs and one exponential multiplication. Now let us analyze the exp term in (1). By using the Binomial approximation, the exp term can be approximated as

$$\exp\left(i2\pi z\sqrt{\frac{1}{\lambda^2}-f^2}\right)\approx\exp\left(\frac{i2\pi z}{\lambda}\left(1-\lambda^2 f^2\right)\right)=$$
$$\exp\left(\frac{i2\pi z}{\lambda}\right)\cdot\exp\left(-i2\pi z\lambda^2 f^2\right). \quad (2)$$

In the right of (2), the first term has nothing to do with the inverse Fourier integral, so it can be discarded and we only consider the second term $\exp(-i2\pi z l^2 f^2)$. Let the sampling rate of the hologram plane be Δu, so according to the sampling theory of the Fourier transform, the sampling rate in the Fourier domain is $\Delta f=1/(N\Delta u)$, where N is the total number of the sampling points. So (1) can be rewritten in the discrete form as

$$f(n\Delta x)=FFT^{-1}\left[FFT\left(h(m\Delta u)\right)\cdot\exp\left(\frac{i2\pi z\lambda^2}{N^2\Delta u^2}\right)\right] \quad (3)$$

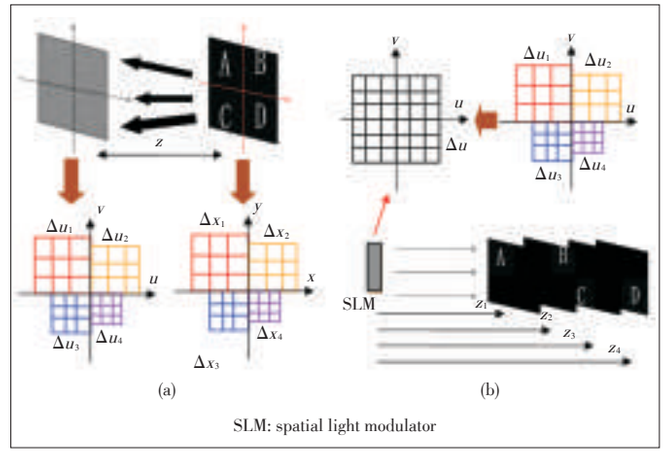
where m and n represent the number of sampling points. Now if the sampling rate in hologram plane is changed from Δu to $\Delta u'=a\Delta u$, (3) can be rewritten as

$$f(n\Delta x)=FFT^{-1}\left[FFT\left(h(m\Delta u)\right)\cdot\exp\left(\frac{i2\pi\alpha^2 z\lambda^2}{N^2\alpha^2\Delta u^2}\right)\right]=$$
$$FFT^{-1}\left[FFT\left(h(m\Delta u)\right)\cdot\exp\left(\frac{i2\pi z'\lambda^2}{N^2\Delta u'^2}\right)\right] \quad (4)$$

Eq. (4) means if the sampling rate of hologram is scaled by a, the image will be reconstructed at a different distance $z'=a^2z$.

2.2 Method of Hologram Design for Multi-Plane Objects

Let us assume a three dimensional object is composed of four letters and each letter is located in different distance from the hologram. The two dimensional pictures with $N\times N$ resolution of this object is taken and located in the image plane. The distance between the image and hologram is z. Both of the hologram plane and image plane (**Fig. 1a**) are decomposed into four parts according to the four quadrants of the coordinate system with the same resolution $N/2\times N/2$. But the sampling in each quadrant is not the same. From the first quadrant to the fourth quadrant, the sampling rate in sequence is $u_2=a_2\Delta u$, $u_1=a_1\Delta u$, $u_3=a_3\Delta u$ and $u_4=a_4\Delta u$, where Δu is the sampling rate of the spatial light modulator (SLM). Similarly the image plane is sampled in the same way as the hologram plane with four different sampling rate Δx_1, Δx_2, Δx_3 and Δx_4 in each quadrant. In this way, the hologram plane and image plane are nonuniformly sampled and the light distribution in the hologram plane can be calculated by diffraction theory from the image.



▲**Figure 1. (a) Sampling method of the hologram plane and the image plane; (b) Reconstruction process from the nonuniform sampled hologram.**

Fig. 1b shows the reconstruction process from the nonuniform sampled hologram. The hologram is loaded into the SLM which has the same sampling rate in each pixel of u. Therefore, this hologram loading would make a mandatory change in the sampling rate of the hologram from the nonuniform sampling to the uniform sampling, which will cause the change of the sampling rate in each quadrant. The sampling rate change in each part will lead to the reconstruction distance change of each letter. For example, the change of sampling rate from Δu_1 to Δu will give rise to the movement of reconstruction position of letter "A" from z to z_1 (Fig. 1b). So are the same effects for the reconstruction of letter "B", "C" and "D". Therefore, the four letters can be reconstructed at different planes simultaneously and separately by nonuniformly sampling the hologram.

2.3 One Step Calculation of Hologram Using Nonuniform Sampled Angular Spectrum Method (NUASM)

We use the nonuniform fast Fourier transform (NUFFT) for the nonuniform sampled angular spectrum method. The principle of calculation is shown in **Fig. 2**. Fig. 2a is the conventional algorithm for angular spectrum method between uniform sampled planes. It includes one FFT and one inverse FFT to calculate the diffraction from uniform sampled image plane $f(x)$ to hologram plane $h(u)$. But if the sampling of $f(x)$ and $h(u)$ are nonuniform (Fig. 2b), the NUFFT is used to calculate the Fourier transform.

NUFFT is an algorithm for calculating the Fourier transform when sampling is nonuniform in either the spatial domain or the frequency domain [13]–[15]. It has been used in scalar diffraction calculation of nonuniform sampled planes by T. Shimobaba et al. [16]–[18]. Generally NUFFT can be categorized into two types: the calculation from nonuniform sampled points to uniform sampled points, and the calculation from uniform sampled points to nonuniform sampled points. In our situation, the calculation of Fourier transform from image plane $f(x)$ to its

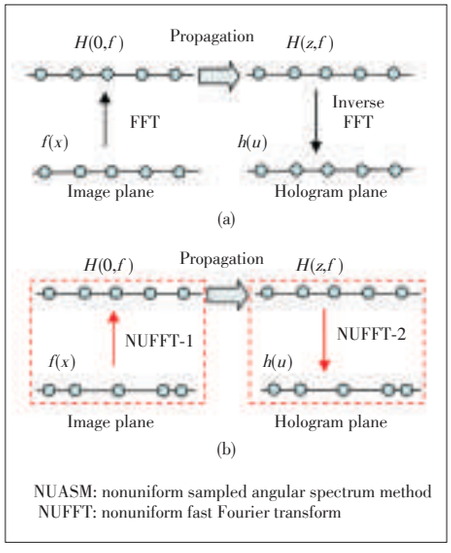


Figure 2. The comparison between (a) conventional angular spectrum method and (b) NUASM.

spectrum domain $H(0, f)$ is the first type. When the spectrum is uniform sampled and it is propagated to the spectrum of distance z to $H(z,f)=H(0,f)\cdot \exp\left(i2\pi z\sqrt{\left(1/\lambda^2\right)-f^2}\right)$, the calculation from $H(z, f)$ to the nonuniform sampled hologram plane $h(u)$ is the second type of the NUFFT . Therefore, the whole calculation of the nonuniform sampled angular spectrum method from the image plane to the hologram plane is expressed as

$$h(u)=NUASM\left[f(x)\right]=$$
$$NUFFT2\left[NUFFT1\left[f(x)\right]\cdot \exp\left(i2\pi z\sqrt{\left(1/\lambda^2\right)-f^2}\right)\right], \quad (5)$$

where *NUASM* is the symbol for nonuniform sampled angular spectrum method, *NUFFT*1 and *NUFFT*2 represent the type 1 and type 2 nonuniform fast Fourier transform respectively. By using 5 the nonuniform sampled hologram which can reconstruct objects on multi-plane simultaneously can be calculated from the two-dimensional image in only one step, rather than the conventional method in which it needs several steps of calculation proportional to the numbers of object planes.

3 Experiments and Results

3.1 Computer Simulation

We first carried out the computer simulation to demonstrate our method of one step hologram calculation. The three-dimensional object is composed of four Chinese characters with different depths so the four characters are located at different distance from the hologram as shown in **Fig. 3**. At first the four characters are combined into one two-dimensional image of 1024 ×1024 resolution and the image is located at the image plane which has the distance z=100 mm from the hologram plane. Both of the image plane and the hologram plane are divided into four parts according to the four quadrants of coordinate system and are nonuniform sampled in the way introduced in Fig. 1a, and the sampling rates are u_1=1.3Δu, u_2=1.1Δu, u_3=0.9Δu and u_4=0.7Δu (Δu is the sampling rate of the SLM). Then the hologram is calculated by (5) from the image plane to the hologram plane.

Then the reconstruction from the complex hologram is calculated by using the FFT based angular spectrum method, and the sampling rate of the hologram is Δu in the calculation. The four characters cannot be reconstructed at the distance z because of the change of sampling rate. Instead, due to the principle described in section 2.1, the change of the sampling rate will result in the reconstruction of four characters in different planes. The relationship between the new distance and the distance z can be calculated. For the first character, the sampling rate changes from 1.3Δu to Δu, so the reconstruction distance is changed from z to $(1/1.3)^2z$. Followed by this, the new reconstruction distance of each character can be calculated by $z_1=z/(1.3)^2$=59 mm, $z_2=z/(1.1)^2$=83 mm, $z_3=z/(0.9)^2$= 123 mm, and $z_4=z/(0.7)^2$=204 mm respectively. The numerical reconstruction result is shown in **Fig. 4a**. It is clear that in each plane the responding Chinese character is clearly focused while the other characters are blurred. Because the sampling rate remains the same in the numerical calculation of FFT based angular spectrum method in reconstruction, the size of the four reconstructed images are the same as the hologram (8 mm×1024). As a comparison, we also presented the simulation results from the compute-generated hologram (CGH) calculated by using the conventional method (**Fig. 4b**). The CGH is calculated and reconstructed by Fresnel diffraction and the size of each reconstructed character is proportional to the distance. It can also be seen from Fig. 4a that all of the four characters are in the same size in each plane and there is no axial distortion with the reconstructed distance which exists in the hologram of multi-plane objects calculated by the Fresnel diffraction. This is because the nonuniform sampling in the image plane makes compensation for the character size in the reconstruction.

3.2 Optical Reconstruction

The optical reconstruction is carried out to show the valida-

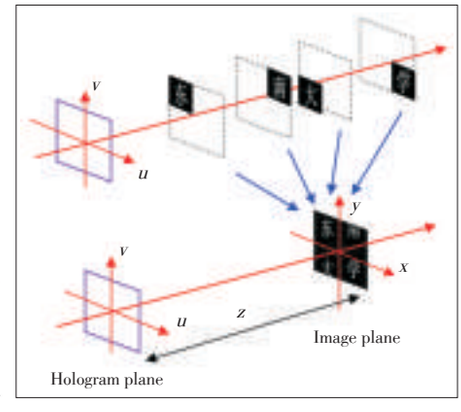
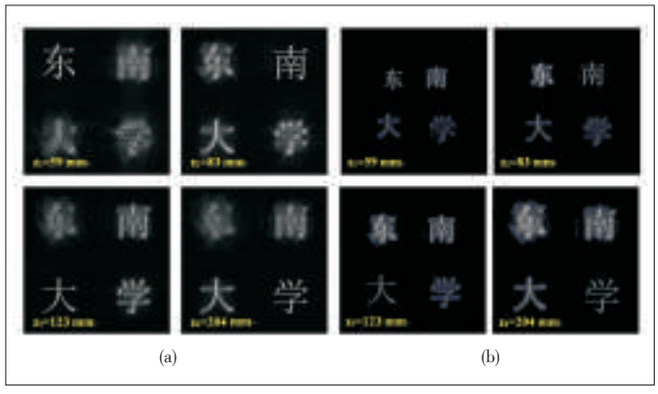


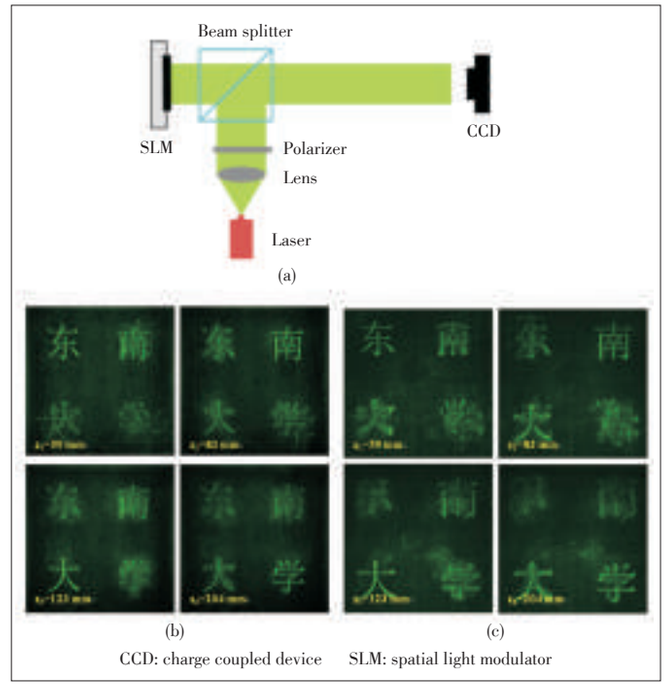
Figure 3. Schematic of one step hologram calculation from 3D object.



▲**Figure 4. Reconstruction from nonuniform sampled hologram at different distances by computer simulation: (a) the proposed method and (b) the conventional method.**

tion of our method in application. Because we use a phase-only type SLM, it is necessary to encode the complex hologram into the phase-only hologram (kinoform). The phase-only hologram can usually be generated by simply discarding the amplitude term of the complex light distribution in the hologram plane, but this will result in much speckle noise in the reconstructed images due to the direct lose of the amplitude information. Here we use the Gerchberg and Saxton (GS) iteration algorithm for optimizing the hologram [19]. The iteration starts by multiplying an initial random phase to the two-dimensional image. The diffraction from nonuniform sampled image to nonuniform sampled hologram is calculated by the nonuniform sampled angular spectrum method in (5), and then we extract only the phase component and impose unity amplitude constraint. The inverse diffraction starting from hologram is calculated by the nonuniform sampled angular spectrum method again back to the image plane, and the amplitude of the image is imposed while remaining the phase component. By repeating this iterative loop with several times, the hologram can be optimized to be used for the phase-only SLM.

Fig. 5a shows the optical setup for the hologram reconstruction on multi-planes. The 532 nm laser is first collimated by the lens and polarizer and then illuminates the SLM after going through the beam splitter. The SLM is holoeye Pluto with 1920×1080 pixel resolutions and 8 mm sampling rate. The reconstructed images are captured directly by the charge coupled device (CCD) camera at different distances. **Fig. 5b** shows the reconstructed images of four Chinese characters at z_1=59 mm, z_2=83 mm, z_3=123 mm and z_4=204 mm respectively. Each character is clearly focused on four separate planes to achieve the multi-plane reconstruction of the three-dimensional object simultaneously. It can also be seen from the results that the four reconstructed characters are the same in size, so there is no axial scale distortion occurred in the reconstruction, which proves the feasibility of our method for true three-dimensional holographic display. **Fig. 5c** shows the reconstructions from the CGH calculated by using conventional method,



▲**Figure 5. (a) Optical setup for holographic display from nonuniform sampled hologram; (b) reconstructed images of four Chinese characters at different distances by our proposed method; (c) reconstructed images of four Chinese characters at different distances by the conventional method.**

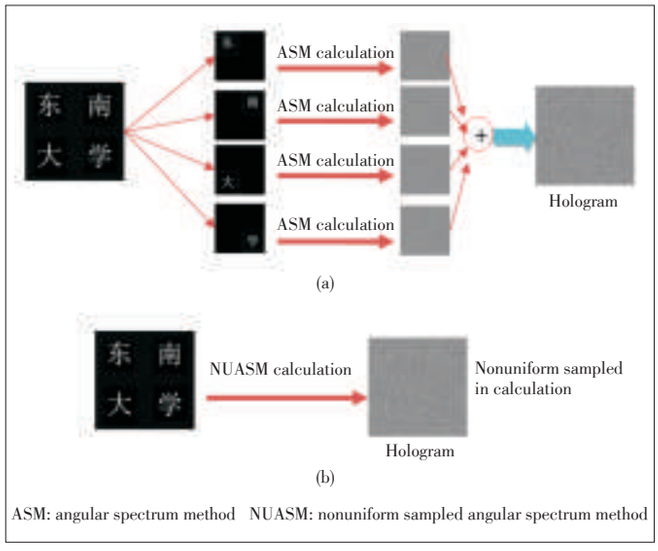
and it can be seen that there is a size distortion of each character due to different reconstructed distances.

4 Discussion

Fig. 6 shows the comparison between the conventional method and our proposed method of hologram calculation for multi-plane 3D objects. In Fig. 6a, the conventional method starts from the decomposition of the three-dimensional objects into several sub-images and then each sub-image is propagated to the hologram plane by the calculation of angular spectrum method. Finally the hologram of multi-plane 3D object is obtained by adding all of the holograms from sub-images. So if the numbers of the object planes are N, it is necessary to calculate the angular spectrum diffraction for N times. In Fig. 6b, by nonuniformly sampling the hologram, we can directly calculate the diffraction from the three-dimensional object to the hologram based on the NUFFT algorithm, and the hologram has the same function for reconstructing at different planes. It is also known that the nonuniform sampled hologram has the advantage of optimizing the numbers of sampling points and eliminating the redundant information properly [20].

5 Conclusions

In conclusion, we proposed a method to calculate the com-



▲**Figure 6. Comparison of the hologram calculation of multi-plane 3D object: (a) hologram calculated by the conventional ASM and (b) hologram calculated by NUASM.**

puter generated hologram of objects on multiple planes by only one step. Both of the hologram plane and the image plane are nonuniformly sampled and the nonuniform sampled angular spectrum method is proposed for diffraction calculation from the image to the hologram. The advantage of our method is that the diffraction calculation from the multi-plane 3D object to the hologram can be calculated in only one step, rather than many steps from each plane to the hologram in the conventional method. Both of the numerical simulation and optical experiment confirmed the practicability of our method. This work has potential applications in the implementation of displaying three dimensional object which can be treated or decomposed into several closely planes.

References

[1] M. Lucente, "Interactive Computation of holograms using a Look-up Table," *Journal of Electronic Imaging*, vol. 2, no. 28, pp. 28–34, Jan. 1993. doi: 10.1117/12.133376.
[2] S.-C. Kim and E.-S. Kim, "Effective generation of digital holograms of 3-D objects using a novel look-up table method," *Applied Optics*, vol. 47, no. 19, pp. 55–62, Jul. 2008. doi: 10.1364/AO.47.000D55.
[3] K. Matsushima and S. Nakahara, "Extremely high-definition full-parallax computer-generated hologram created by the polygon-based method," *Applied Optics*, vol. 48, no. 34, pp. 54–63, Dec. 2009. doi: 10.1364/AO.48.000H54.
[4] Y. Pan, Y. Wang, J. Liu, X. Li, and J. Jia, "Fast polygon-based method for calculating computer-generated holograms in three-dimensional display," *Applied Optics*, vol. 52, no. 1, pp. 290–299, Jan. 2013. doi: 10.1364/AO.52.00A290.
[5] T. Shimobaba, N. Masuda, and T. Ito, "Simple and fast calclulation algorithm for computer-generated hologram with wavefront recording plane," *Optics Letters*, vol. 34, no. 20, pp. 3133–3135, Oct. 2009. doi: 10.1364/OL.34.003133.
[6] J. Weng, T. Shimobaba, N. Okada, H. Nakayama, M. Oikawa, N. Masuda, and T. Ito, "Generation of real-time large computer generated hologram using wavefront recording method," *Optics Express*, vol. 20, no. 4, pp. 4018–4023, Feb. 2012. doi: 10.1364/OE.20.004018.
[7] J. Xia, and H. Yin, "Three-dimensional light modulation using phase-only spatial light modulator," *Optical Engineering*, vol. 48, no. 2, pp. 020502, Feb. 2009. doi: 10.1117/1.3076211.
[8] T. Haist, M. Schonleber, and H. J. Tiziani, "Computer-generated holograms from 3D-objects written on twisted-nematic liquid crystal display," *Optics Communications*, vol. 140, no. 4, pp. 299–308, Aug. 1997. doi: 10.1016/S0030-4018(97)00192-2.
[9] M. Makowski, M. Sypek, A. Kolodziejczyk, G. Mikula, and J. Suszek, "Iterative design of multiplane holograms: experiments and applications," *Optical Engineering*, vol. 46, no. 4, pp. 045802, Apr. 2007. doi: 10.1117/1.2727379 .
[10] M. Makowski, M. Sypek, and A. Kolodziejczyk, "Colorful reconstructions from a thin multi-plane phase hologram," *Optics Express*, vol. 16, no. 15, pp. 11618–11623, Jul. 2008. doi: 10.1364/OE.16.011618.
[11] G. Sinclair, J. Leach, P. Jordan, et al., "Interactive application in holographic optical tweezers of a multi-plane Gerchberg-Saxton algorithm for threedimensional light shaping," *Optics Express*, vol. 12, no. 8, pp. 1665–1670, Apr. 2004. doi: 10.1364/OPEX.12.001665.
[12] H. Zheng, Y. Yu, T. Wang, and L. Dai, "High-quality three-dimensional holographic display with use of multiple fractional Fourier transform," *Chinese Optics Letters*, vol. 7, no. 12, pp. 1151– 1154, Dec. 2009. doi: 10.3788/COL20090712.1151.
[13] A. Dutt and V. Rokhlin, "Fast Fourier transforms for nonequispaced data," *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1368–1393, Nov. 1993. doi: 10.1137/0914081.
[14] Q. H. Liu and N. Nguyen, "An accurate algorithm for nonuniform fast Fourier transforms (NUFFTs)," *IEEE Microwave and Guided Wave Letters*, vol. 8, no. 20, pp. 18–20, Feb. 1998. doi: 10.1109/75.650975.
[15] L. Greengard and J. Y. Lee, "Accelerating the Nonuniform Fast Fourier Transform," *SIAM Review*, vol. 46, no. 3, pp. 443–454, Sep. 2004. doi: 10.1137/S003614450343200X.
[16] T. Shimobaba, K. Matsushima, T. Kakue, N. Masuda, and T. Ito, "Scaled angular spectrum method," *Optics Letters*, vol. 37, no. 19, pp. 4128–4230, Oct. 2012. doi: 10.1364/OL.37.004128.
[17] T. Shimobaba, N. Masuda, and T. Ito, "Arbitrary shape surface Fresnel Diffraction," *Optics Express*, vol. 20, no. 8, pp. 9335–9340, Apr. 2012. doi: 10.1364/OE.20.009335.
[18] T. Shimobaba, T. Kakue, M. Oikawa, et al., "Nonuniform sampled scalar diffraction calculation using nonuniform fast Fourier transform," *Optics Letters*, vol. 38, no. 3, pp. 5130–5131, Dec. 2013. doi: 10.1364/OL.38.005130.
[19] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, no. 2, pp. 237–246, Nov. 1972. doi: 10.1070/QE2009v039n06ABEH013642.
[20] Z. Zhang, J. Liu, J. Jia, et al., "Tunable nonuniform sampling method for fast calculation and intensity modulation in 3D dynamic holographic display," *Optics Letters*, vol. 38, no. 15, pp. 2676– 2679, Aug. 2013. doi: 10.1364/OL.38.002676.

Manuscript received: 2015-10-21

Biographies

CHANG Chenliang (changchenliang@hotmail.com) received his BS and PhD degrees from Southeast University, China in 2009 and 2015 respectively. He is now a lecturer at Department of Physics of Nanjing Normal University, China. His current research interests include phase retrieve algorithm in holographic display, digital holographic microscopy, design and computation of computer-generated hologram. He has published more than 10 journal papers and authorized two Chinese patents.

XIA Jun (xiajun@seu.edu.cn) received his MS and PhD degrees from School of Electronic Science and Engineering, Southeast University in 1999 and 2004 respectively. He was a visiting professor with the Delft University of Technologyin 2007. He is now the professor at the Display Center of Southeast University. His current research interests include signal processing, novel three-dimensional display, and flexible display based on electrowetting technology. He has published more than 60 scientific papers and has more than 30 patterns.

LEI Wei (lw@seu.edu.cn) received his BS, MS and PhD degrees from School of Electronic Science and Engineering, Southeast University in 1987, 1989 and 1993 respectively. He visited the Philips Company (Netherland) in 1998. He is now the professor at the Display Center of Southeast University. His current research interests include 3D display and field emission display. He has published more than 200 scientific papers and has more than 20 patterns.

Research Paper

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 009
http://www.cnki.net/kcms/detail/34.1294.TN.20170106.1827.002.html, published online January 6, 2017

Research on Interference Cancellation for Switched-on Small Cells in Ultra Dense Network

SUN Yang[1], CHANG Yongyu[1], WANG Chao[1], ZHANG Lu[1], ZHANG Yu[2], and WANG Xinhui[2]
(1. Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. ZTE Corporation, Shanghai 201203, China)

Abstract

In ultra-dense heterogeneous networks, the co-channel interference between small cells turns to be the major challenge to cell throughput improvement, especially for cell edge users. In this paper, we propose a distributed frequency resource allocation approach for interference cancellation, which allocates appropriate frequency resources when a small cell is switched on to reduce the co-channel interference to its neighboring small cells. This frequency resource pre-allocation aims at avoiding co-channel interference between small cells and improving users' throughput. The simulation results show that our proposed scheme can effectively reduce the co-channel interference and achieve considerable gains in users' throughput.

Keywords

ultra-dense network; interference cancellation; frequency resource allocation; cell switch

1 Introduction

With the exploding increase of portable devices in wireless communications, the wireless industry has stepped in the 5th generation (5G). As estimated in [1], the future wireless network will face 1000x rate challenge beyond 2020. Ultra-dense heterogeneous networks (UDN) are proved to be one of key 5G technologies.

Heterogeneous network (HetNet) is a network that contains macro cells, small cells and users. In HetNet, users can select their serving cell adaptively based on their location and received signal power. Macro cells and small cells can all be served as users' serving cell. In addition, macro cells with high transmission power are deployed to provide wide coverage, whereas some small cells are deployed for coverage extension or offloading. There exist a lot of mature research studies in 3GPP HetNet [2]–[4].

UDN is an enhancement of HetNet. In UDN, there are often dozens of small cells even hundreds of small cells in a macro cell. UDN can provide much more network capacity. In addition, because of low power cost of small cells, it offers a great improvement for energy efficiency. UDN can also ensure full coverage of the service area [5].

However, a large number of small cells deployed in one macro cell lead to a more serious co-channel interference issue between small cells. In sparse HetNet, there are often a few small cells deployed in a macro cell. Because of the large distance between neighboring cells and the low transmission power, co-channel interference is tolerable. On the other hand, due to the dense deployment of small cells, the distance between neighboring cells in UDN is greatly reduced. As a result, the co-channel interference turns to be more severe than before. If the frequency resource cannot be allocated reasonably to every cell, the strong co-channel uplink-downlink interference between neighboring small cells would lead to the whole network performance decline [6].

In order to reduce co-channel interference between neighboring cells, collaboration of a variety of interference cancellation technologies to improve the quality of network service is required. Frequency reuse technology can guarantee the neighboring cells using orthogonal frequency resources, so quality of service (QoS) can be ensured and co-channel interference be alleviated [7]. The cell switching method can bring a more efficient resource management and obviously eliminate the interference [8], [9].

However, existing frequency reuse schemes are generally focused on on-service small cells. When a new cell is informed to be switched on in UDN, unplanned frequency allocation will introduce serious interference to its neighbor cells and impact the system throughput performance. In order to tackle this problem, in this paper, we combine the cell switching operation with the frequency reuse strategy, and propose an advanced interference cancellation method based on frequency resource allocation for candidate switched-on small cells.

In the following parts of this paper, we give an overview of some related resource allocation technologies in section 2, and propose an advanced interference cancellation method based on frequency resource allocation in section 3. In section 4, we demonstrate the effectiveness of our advanced frequency allocation method through system simulation. Finally, we give the

This work was partly supported by ZTE Industry-Academia-Research Cooperation Funds.

conclusion in section 5.

2 Overview of Resource Allocation Techniques

2.1 Frequency Reuse in HetNet

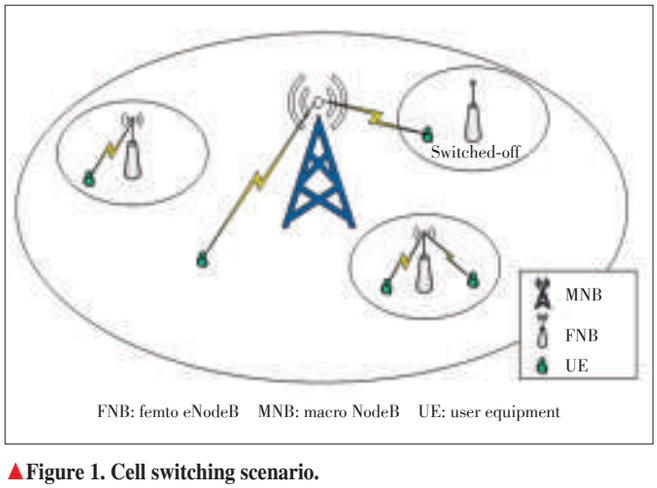
In HetNet, the introduction of small cells brings an obvious improvement for hotspots, and gives a supplement for the macro cell's coverage. In order to achieve a further gain on network capacity, there are much more small cells deployed in a macro cell.

Due to the inter-frequency deployment of Macro eNodeB (MNB) and Femto eNodeB (FNB), the interference between them is negligible. Therefore, we should pay attention to co-channel interference among small cells. In addition, fractional frequency reuse (FFR) and soft frequency reuse (SFR) are useful for macro cells in LTE standards. However, because of random deployment of small cells, FFR and SFR will be complex and difficult to plan in UDN. Therefore, we take traditional frequency reuse technology into account for the UDN scenario [10].

In UDN, if we take other cells' frequency resource and interference information into account to make a more reasonable frequency allocation, co-channel interference will be further decreased and resource utilization will be more increased [11].

2.2 Cell Switching

In HetNet, cell switching is a new radio resource management method which is proposed recently. As **Fig. 1** shows, there are some small cells switched off by using this method, where UE means User Equipment. There are several factors that influence the switch decision of small cells, like interference from neighboring cells, cell load, and packet arrival rate [12]. The cell switching technology can save energy and signaling overhead, and also improve resource utilization of system. As a consequence, it plays a significant role in co-channel interference avoidance [8].



▲**Figure 1. Cell switching scenario.**

Due to the dense deployment of small cells in UDN, the propagation environment of network is more complex. In addition, dense deployment of small cells does not equal to dense deployment of users. When users are distributed sparsely, there will be a large proportion of small cells at low load [13]. To save energy and signaling overhead, it is better to switch off a part of small cells; moreover, as mall cell produces strong interference to its adjacent cells for a while, it should be switched off to improve quality of network service. So the cell switch scheme will take a great boost for UDN which has more complex network deployment and propagation environment [7].

However, if frequency resource cannot be allocated reasonably when small cells are turned on, it will introduce serious co-channel interference to the system, and QoS will be negatively influenced. Taking above situations into consideration, we propose a frequency allocation method for switched-on small cells to reduce the interference.

3 Proposed Frequency Pre-Allocation Scheme

3.1 System Model

A downlink resource allocation scheme is proposed in an orthogonal frequency division multiple access (OFDMA) time division duplex system. The system is composed of multiple cells which share the same frequency band with the bandwidth of 10 MHz. In the macro cell's coverage area, there are M small cells. Let K denote the set of user equipment (UEs), S denote the set of small cells and U denote the set of physical resource blocks (PRBs). The system time is divided into fixed-length frames. Since the resource allocation is performed on a per cell basis, we will focus on the small cell that is switched on recently (the target cell) in describing our scheme. Meanwhile, the interference from other cells is considered in designing the scheme.

We assume that the transmission modes of all UEs are selected before resource allocation. That is, it is assumed that each UE has determined its best transmitter among macro cells and small cells as the serving transmitter by using an appropriate mode selection method [14]. UEs select their serving cells by Reference Signal Received Power (RSRP), which can be achieved as follows:

$$RSRP = P_u^i \cdot H_{k,u}^i, i \in S, u \in U, k \in K, \quad (1)$$

where $H_{k,u}^i$ denotes the desired channel gain that UE k experiences from eNB i at PRB u including all long-term and short-term channel-fading characteristics. P_u^i is the transmit power allocated to PRB u by eNB i.

In addition, the formulation of the instantaneous Signal-to-

Interference and Noise Ratio (SINR) $\gamma_{k,u}^{i}$ is given as follows:

$$\gamma_{k,u}^{i}=\frac{P_{u}^{i}\cdot H_{k,u}^{i}}{\sum_{j\neq i}P_{u}^{j}\cdot H_{k,u}^{j}+N_{w}},\quad i,j\in S,\ u\in U,\ k\in K, \tag{2}$$

where $H_{k,u}^{j}$ denotes the interfering link *and* P_{u}^{j} is the transmit power allocated to PRB u by eNB j. To convert the achievable SINR $\gamma_{k,u}^{i}$ into an effective data rate $r_{k,u}^{i}$, we assume a log-linear function as follows:

$$r_{k,u}^{i}=B\cdot\log_{2}\left(1+\gamma_{k,u}^{i}\right), \tag{3}$$

where B is the band width allocated for UE k.

3.2 Problem Formulation

The optimization problem is formulated in the context of OFDMA as in the LTE standard. There are U effective subchannels—PRBs in LTE—available in the system each with a bandwidth of 10 MHz. In our scheme, frequency resources are divided into three parts: available resources, shared resources, and low-prior resources. As **Fig. 2** shows, available resources are the bandwidth which has been not used by small cells. Shared resources are the bandwidth which has been reused by small cells with reuse factor N=1. In addition, the bandwidth with reuse factor N=3 is called low-prior resources, which meets the small cells' minimum resource requirements.

We let $B=\{B_1,B_2,...,B_n\}$ denote the set of bandwidth. After a femtocell is switched on, it is called the target cell i. Our goal is to minimize the interference of the femtocell to its neighboring cells. Then, the interference I is formulated as follows:

$$I(B_i)=\begin{cases}0, & B_i\neq B_j\\ P_{u}^{i}\cdot H_{k,u}^{i}, & B_i=B_j\end{cases},\quad i,j\in S,\ u\in U,\ k\in K, \tag{4}$$

where B_i and B_j are the frequency resource of the target cell and neighboring cell, respectively.

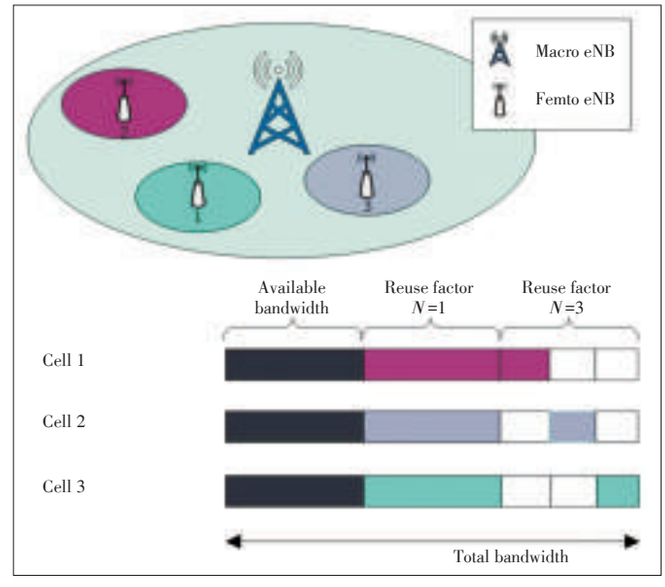
Next, the frequency resource allocation problem is formulated for the target cell as follows:

$$\arg\left(\min\sum I(B_i)\right). \tag{5}$$

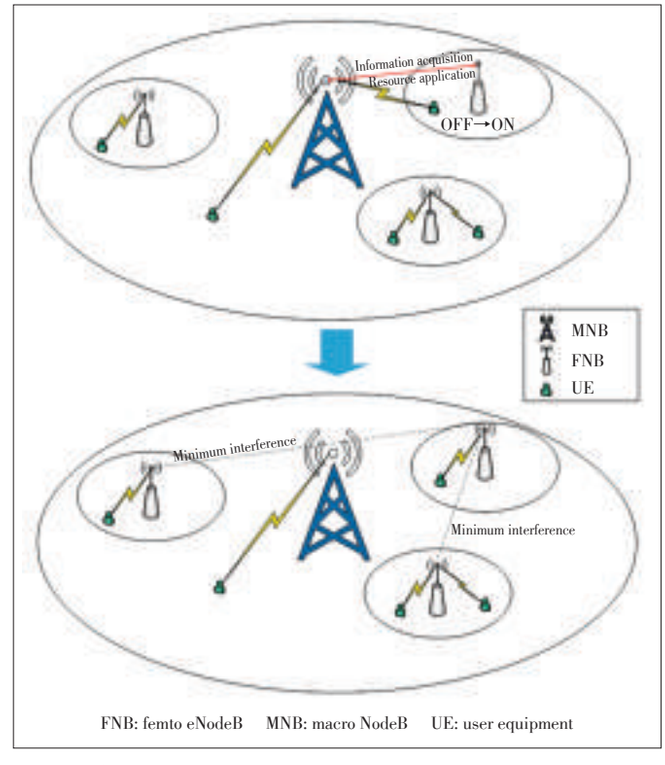
As (4) and **Fig. 3** show, the frequency resource that is orthogonal with other cells' is the best solution to minimize the interference of the target cell. If there are available resources, the target cell will use them. $I(B_i)$ can be zero by this method. If there are no available resources but shared resources, the target cell can occupy for them, and $I(B_i)$ can also be zero. Otherwise, if there are only low-prior resources, the target cell will apply for B_i, which will lead $I(B_i)$ is minimum.

3.3 Process of the Method

When a small cell becomes ON state from OFF state, an initial frequency resource application process is beginning. To minimize the co-channel interference from neighboring cells after frequency resources are allocated, it's better to apply for



▲**Figure 2. Frequency resources in our scheme.**



▲**Figure 3. Frequency pre-allocation in our scheme.**

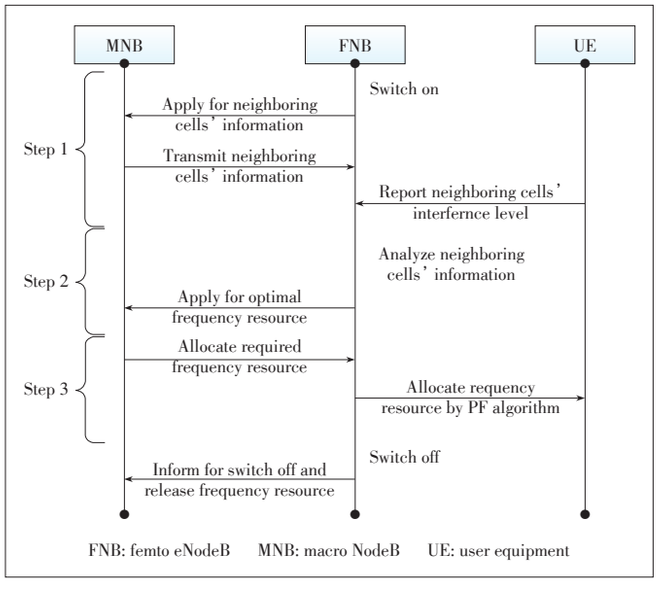
the frequency resource for the cell according to RSRP and neighboring cells' information of frequency resource usage.

This method can be divided into three steps. As **Fig. 4** shows, the first step is the information acquisition. After a small cell is switched on, it becomes the target cell in our scheme and it should obtain its neighboring cells' on/off status and information of frequency resource usage through backhaul links between macro cells and small cells. Then the small cell commands its UEs to measure downlink RSRP of all neighboring cells which are on-state. After that, UEs report the measuring result to the target cell, and the target cell will calculate every neighboring cell's average RSRP. In this way, it can estimate the interference strength of each cell according to its RSRP. The higher RSRP is, the stronger interference between the target cell and neighboring cell is.

The second step is frequency resource allocation. First, the target cell should find if there are any available bands. It will apply for the spare frequency resources if they are available. Otherwise, it will apply for some frequency resource reusing with other cells according to frequency resource usage of these cells. In our system models, there are shared frequency resources and low-prior frequency resources. If there are the shared resources used by some neighboring cells, the target cell can apply for the shared resources directly. Otherwise, if all the adjacent cells only use low-prior resources, the target cell should apply for reusing the frequency resources with the neighboring cell which has the lowest RSRP. In this way, we can guarantee that the introduction of the target cell produces the minimal interference to other cells.

In the third step, the target cell obtains the frequency resources from network. If the resources are the available resources or shared resources, the system allocates them to the target cell directly. If the resources are low-prior resources, the system allocates them to the target cell and informs the original cell for frequency reusing.

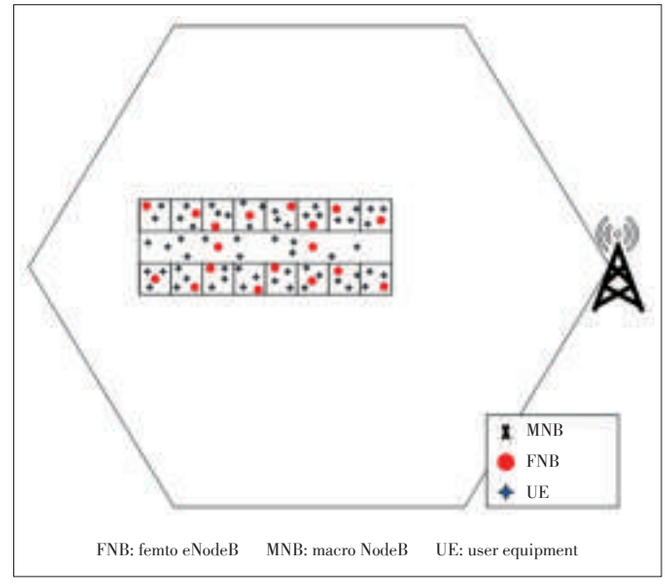
After the small cell is switched off, its frequency resources will go back to the resource pool waiting for scheduling next time.



▲Figure 4. Process of the proposed scheme.

4 Simulation Scenario and Results

We evaluated the performance of the proposed scheme. The simulation is based on the LTE standard [4]. The downlink transmission scheme for an LTE system is based on OFDMA. As **Fig. 5** shows, an inter-frequency indoor office scenario is used in our simulation. We consider FNBs as small cells. Resources are allocated to users according to proportional fair scheduling. **Tables 1** and **2** list the simulation parameters and the scenario parameters we used. Because of the inter-frequency deployment of macro cells and femtocells, their cross-tier in-



▲Figure 5. Simulation scenario.

▼Table 1. Simulation parameters

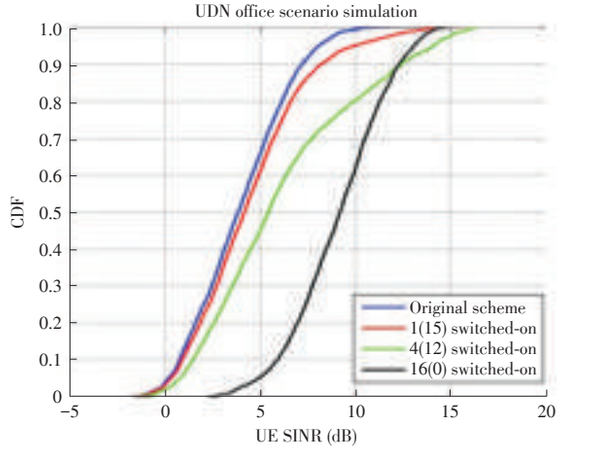
	Macro	Femto
Carrier frequency	1.8 GHz	2.5 GHz
Bandwidth	20 MHz	10 MHz
Antenna gains	17 dBi	5 dBi
Antenna models	3D sectors	Omnidirectional antenna
Transmission power	46 dBm	20 dBm/30 dBm
Inter-site distance	500 m	none
Log-normal shadow	10 dB	3 dB
Minimum access distance	35 m	3 m
Maximum service users	10	3
Traffic model	Full buffer	

▼Table 2. Scenario parameters

Cell model	7 cells model, 3 sectors per cell
Building model	1 one-layer building per sector (16 rooms, 1 corridor)
Building scale	Room: 15 m×15 m, Corridor: 20 m×120 m
Users per room	3
Corridor users	16
Number of FNB	2 FNBs in corridor, 1 FNB per room
FNB: Femto eNodeB	

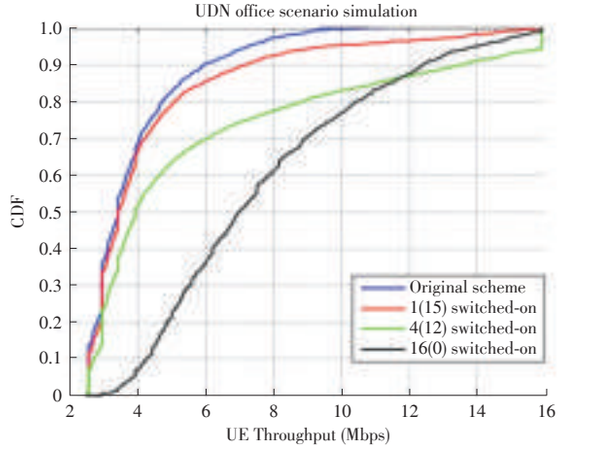
terference is negligible and we only focus on the co-channel interference between the femtocells.

According to the simulation in the office scenario, we get SINR and the cumulative distribution function (CDF) of users' throughput (**Figs. 6** and **7**). In the figures, the original scheme uses the unplanned frequency allocation method of 16 small cells without cell switch and 1(15)/4(12)/16(0) switched-on denote switched-on 1/4/16 small cells based on 15/12/0 existing small cells with unplanned frequency allocation in the proposed advanced scheme.



CDF: cumulative distribution function
SINR: Signal-to-Interference and Noise Ratio
UDN: ultra-dense heterogeneous network
UE: user equipment

▲Figure 6. Users' SINR.



CDF: cumulative distribution function
UDN: ultra-dense heterogeneous network
UE: user equipment

▲Figure 7. Users' Throughput.

Fig. 6 shows that the SINR of the advanced method is better than the original method. According to SINR cumulative distribution function (CDF) curves showed in Figs. 6 and 7, we can see that after a small cell is switched on, the interference level has been apparently alleviated. What's more, it turns out that with more small cells switched on, the interference level is more improved in Fig. 6. Based on the improvement of system capacity from switched-on cells, our proposed scheme provides higher performance for system.

Fig. 7 gives the femto users' average throughput and shows that the trend of throughput curves are consistent with SINR curves. In our system models, there are different numbers of small cells switched on during a drop. Compared with the original scheme, UEs' average throughput is improved visibly because of interference cooperation of our proposed scheme. In addition, after more small cells are switched on, the system capacity is more enhanced and UEs' average throughput is also improved more evidently.

Frequency pre-allocation can reduce co-channel interference after a small cell is turned on. With this method, users can get a higher data rate because of low interference. In addition, the introduction of this method has no negative effect on the whole system service. The cell switch method can save system energy and signaling overhead. In a word, our proposed scheme for switched-on cells can provide a higher system performance.

5 Conclusions

In this paper, we proposed an advanced frequency allocation method for co-channel interference cancellation via analysis of existing resource management methods. It turns out that the improvement of frequency resource allocation method can guarantee the quality of system service according to our simulation results. The proposed method can also play the role of interference cancellation. However, band utilization of this method is lower than the whole frequency reuse, and cell center users' throughput will be a little reduced. In the future, we will further optimize the method and enhance the service quality of the system.

References

[1] B. S. Hwang, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20–27, Jun. 2013. doi: 10.1109/MCOM.2013.6525591.
[2] A. Damnjanovic, J. Montojo, Y. Wei, *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, Jun. 2011. doi: 10.1109/MWC.2011.5876496.
[3] B. M. Hambebo, M. M. Carvalho, and F. M. Ham, "Performance evaluation of static frequency reuse techniques for OFDMA cellular networks," in *Proc. 11th*

IEEE International Conference on Networking, Sensing and Control, Miami, USA, 2014, pp. 355–360. doi: 10.1109/ICNSC.2014.6819652.

[4] S. Kumar, S. Kalyani, and K. Giridhar, "Spectrum allocation for ICIC-based picocell," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 8, pp. 3494–3504, Aug. 2015. doi: 10.1109/TVT.2014.2360454.

[5] M.-C. Chuang, M. C. Chen and Y. Sun, "Resource management issues in 5G ultra dense small-cell networks," in *2015 International Conference on Information Networking (ICOIN)*, Cambodia, 2015, pp. 159–164. doi: 10.1109/ICOIN.2015.7057875.

[6] J. B. Rao and A. O. Fapojuwo, "A survey of energy efficient resource management techniques for multi-cell cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 154–180, 2014. doi: 10.1109/SURV.2013.042313.00226.

[7] F. Ahmed, A. A. Dowhuszko, and O. Tirkkonen, "Distributed algorithm for downlink resource allocation in multicarrier small cell networks," in *2012 IEEE International Conference on Communications (ICC)*, Ottawa, Canada, 2012, pp. 6802–6808. doi: 10.1109/ICC.2012.6364716.

[8] D. P. Siewiorek and E. J. McCluskey, "An iterative cell switch design for hybrid redundancy," *IEEE Transactions on Computers*, vol. C-22, no. 3, pp. 290–297, Mar. 1973. doi: 10.1109/T-C.1973.223709.

[9] T. Beitelmal and H. Yanikomeroglu, "A set cover based algorithm for cell switch-off with different cell sorting criteria," in *2014 IEEE International Conference on Communications Workshops (ICC)*, Sydney, Australia, 2014, pp. 641–646. doi: 10.1109/ICCW.2014.6881271.

[10] H. E. E. O. M. Elfadil, M. A. I. Ali, and M. Abas, "Fractional frequency reuse in LTE networks," in *2nd World Symposium on Web Applications and Networking (WSWAN)*, Sousse, Tunisia, 2015, pp. 1–6. doi: 10.1109/WSWAN.2015.7210297.

[11] A. R. Elsherif, W. P. Chen, A. Ito, and Z. Ding, "Adaptive resource allocation for interference management in small cell networks," *IEEE Transactions on Communications*, vol. 63, no. 6, pp. 2107–2125, Jun. 2015. doi: 10.1109/TCOMM.2015.2420676.

[12] L. Wang, X. Feng, X. Gan, J. Liu, and H. Yu, "Small cell switch policy: a reinforcement learning approach," in *Sixth International Conference on Wireless Communications and Signal Processing (WCSP)*, Hefei, China, 2014, pp. 1–6. doi: 10.1109/WCSP.2014.6992126.

[13] G. Cili, H. Yanikomeroglu, and F. R. Yu, "Cell switch off technique combined with coordinated multi-point (CoMP) transmission for energy efficiency in beyond-LTE cellular networks," in *IEEE International Conference on Communications (ICC)*, Ottawa, Canada, 2012, pp. 5931–5935. doi: 10.1109/ICC.2012.6364869.

[14] M. Rodziewicz, "Location-based mode selection and resource allocation in cellular networks with D2D underlay," in *Proc. 21th European Wireless Conference*, Budapest, Hungary, 2015, pp. 1–6.

Manuscript received: 2015-12-19

Biographies

SUN Yang (sunyangemail@bupt.edu.cn) received the BS degree from the Beijing University of Posts and Telecommunications (BUPT), China in 2011. She is currently working toward the PhD degree with the School of Information and Communication Engineering, BUPT. Her current research interests focus on ultra-dense heterogeneous network, massive MIMO and green telecommunications.

CHANG Yongyu (yychang@bupt.edu.cn) received her PhD from BUPT in July 2005. Since then, she has worked with the Wireless Theory and Technology Laboratory of School of Information and Communication Engineering there. She has been engaged in the research of theories and key technologies in mobile communication for many years. She has made remarkable achievements by leading or participating in many research programs, some of which are supported by the government, such as the China National "973" Program and the China National "863" Program. Her current research is focused on the key technologies and system performance evaluation for 4G/5G wireless networks.

WANG Chao (874699078@qq.com) received his BE degree in communications engineering from BUPT in 2015. Currently, he is a postgraduate with the Wireless Theory and Technology Laboratory in BUPT. His research focuses on M2M communication, radio resource management and massive MIMO.

ZHANG Lu (zhanglu.lucas@outlook.com) received his BE and MS degree in communications engineering from BUPT in 2013 and 2016 respectively. His research interests include mobility management and radio resource management in 5G ultra-dense network.

ZHANG Yu (Zhang.yu38@zte.com.cn) received her master's degree in signal and information processing from Nanjing University of Post and Telecommunication, China in 2006. She is a senior standard engineer with ZTE Corporation, engaged in wireless high-level protocol standard research from 2008 with a focus on 5G related topics such as UDN and Massive MIMO from 2014. She is the co-authors of "White Paper of 5G Future Forum" and "White Paper of IMT-2020(5G) Wireless Technology" of ZTE Corporation.

WANG Xinhui (wangxinhui@zte.com.cn) received his master's degree in database processing and design from Northeastern University, China in 1998. He is the director of standardization of ZTE Corporation, engaged in standard research and industry relationship from 2006. He has led the 5G Standardization team to work on 3GPP RAN/3GPP SA/IEEE/ITU specifications in different radio aspects from 2013. He is the first authors of ZTE 5G white paper "Driving the Convergence of the Physical and Digital Worlds", 5G paper "Focusing on User Experience to Build in Converged, Innovative 5G Networks", and CCSA 5G speech "5G Vision and Challenges".

Research Paper

DOI: 10.3969/j. issn. 1673-5188. 2016. S1. 008
http://www.cnki.net/kcms/detail/34.1294.TN.20170105.1512.002.html, published online January 5, 2017

Action Recognition in Surveillance Videos with Combined Deep Network Models

ZHANG Diankai[1], ZHAO Rui-Wei[2], SHEN Lin[1],
CHEN Shaoxiang[2], SUN Zhenfeng[2], and JIANG Yu-Gang[2]
(1. ZTE Corporation, Nanjing 210012, China;
2. Fudan University, Shanghai 201203, China)

Abstract

Action recognition is an important topic in computer vision. Recently, deep learning technologies have been successfully used in lots of applications including video data for sloving recognition problems. However, most existing deep learning based recognition frameworks are not optimized for action in the surveillance videos. In this paper, we propose a novel method to deal with the recognition of different types of actions in outdoor surveillance videos. The proposed method first introduces motion compensation to improve the detection of human target. Then, it uses three different types of deep models with single and sequenced images as inputs for the recognition of different types of actions. Finally, predictions from different models are fused with a linear model. Experimental results show that the proposed method works well on the real surveillance videos.

Keywords

action recognition; deep network models; model fusion; surveillance video

1 Introduction

Action recognition in surveillance videos has long been an important research topic in the computer vision community. Automatic or machine aided surveillance technologies can be widely used in public areas like airport, banks, shops, etc.

The currently used action analysis methods usually contain the following steps: detection, tracking and recognition. Traditionally, some handcrafted visual features are required to be extracted from the video in order to make further computation for recognition. These traditional features include color histograms, scale-invariant feature transform (SIFT), histogram of oriented gradient (HoG), etc. The detection, tracking and recognition algorithms usually rely on these kinds of computed feature values. However, these manually designed features may suffer from their limitation in describing complicated actions in complex environment. Therefore, recognition algorithms relying on this kind of features do not always work very well in many real applications [1]–[3]. In recently years, features automatically learnt from deep neural networks are widely used in the computer vision community because of their great successes in many real applications [4], [5]. So far, there have been some work concentrating on applying deep learning methods on video analysis applications, for example, 3D convolutional neural networks (3D-CNN) [6], recurrent neural network [7] and two-streams models [8], [9]. All these methods highlight on general framework for video analysis. While for the specific problem of action recognition in the surveillance videos, there exists room for further improvement.

For example, the recognition algorithm for different types of human actions could be treated in different ways. Some actions could be comparatively easy to be recognized by single frames because of their characteristic appearance. Such actions include fighting with others or riding a bicycle. However, some other types of actions may not be easy to distinguish by merely single frames. For instance, some frames of walking and running, especially jogging, may look very similar in appearance. In order to distinguish these actions, the temporal information that describing the motion of the actions is more helpful. Therefore, the single model is not adequate to make classification for all types of human actions in surveillance videos. A composite method that combines both single image based recognition and images sequence based models could potentially be a good solution to this problem.

Motivated by the issues discussed above, we propose novel action recognition for surveillance videos in this paper. The proposed method contains improvement modifications in human detection and tracking methods, as well as a novel integrated action recognition strategy with the help of three different types of deep neural network models. The novelty and advantages of the proposed algorithm are as follows.

The detection algorithm in the proposed algorithm uses true detections from the detector and the motion points extracted from foreground motion information to compensate for miss detections. This modification could tackle with the issue of rapid appearance changes and occlusion of objects that often occur in the surveillance videos.

In the proposed network framework for action recognition, a novel fusion strategy from the spatial frames based network model and from the temporal frames based network model is

This work is supported by ZTE Industry-Academia-Research Cooperation Funds.

used to make final action recognition. Specifically, models of different structures are used to tackle with the different types of actions in surveillance videos, considering their static and dynamic properties. In the end, recognition results from both the models are fused to produce the final action classification result. Specifically, the proposed recognition algorithm relies more on a sufficiently sophiscated network model based on single frame input for some action that is distinguishable from static image frame like fighting among people or riding a bicycle. While for human actions not very distinguishable from single input image, the proposed algorithm would switch to rely more on a temporal model based on stacked image frames to provide more accurate action prediction results. For both spatial and temporal recognition models, the input image patches are expanded to the minimal square regions that completely cover the detection boxes of the objects.

Moreover, the cascade classification strategy is applied to the spatial recognition network model and temporal network model to improve the algorithm efficiency. To realize this, some thresholds are set on some predicted scores from the spatial recognition network. If the predicted scores on these layers are higher than the specified thresholds, the corresponding target object belongs to that specific action. In this way, the recognition process of the succeeding temporal network is skipped by the algorithm. The advantage is that in the prediction process, more computing resources could be saved on recognizing these simple actions and runtime efficiency could be improved.

Besides, a special network was applied to provide more accurate judgment on discerning whether detected image regions contain real foreground content or are false alarms from the detection algorithm instead. To realize this, the proposed algorithm applies a comparatively simple structured network to the detection regions in the image frame and its only aim is to check whether the image regions are real foreground or not. We intentionally set low threshold to this model so that only background regions with very high confidence are filtered out and the chances of missing any real foreground objects in the video are accordingly reduced. As this simple network model is applied only on the regions returned by the detection and tracking algorithm, the computation burden introduced by this background filtering network is in fact very limited.

2 The Approach

2.1 General Framework

In general, the proposed method first receives the captured video data as input. Then improved human detection and tracking algorithms are used to detection human regions in the original videos. Based on the detection and tracking results, proposed action recognition algorithms are applied to predict the action labels for every object obtained.

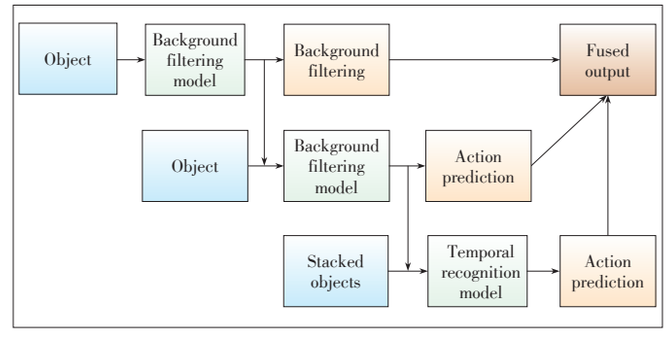
The first step is to detect and track persons in the video. We adopt the widely used tracking-by-detection method to achieve this. Specially, we calculate the motion points of the figures during the person detection and use these motion points extracted from foreground motion information to compensate for miss detections in the video.

The recognition module includes three different neural networks (**Fig. 1**). The proposed recognition algorithm is based on the neural network models to distinguish different actions in the videos. All the contextual information of the target person in the video is utilized. The proposed recognition algorithm uses both spatial and temporal information in the video sequences to achieve better recognition results.

The first network model (the upper row of Fig. 1) is used to further filter non-human patches returned by the detection and tracking algorithm. This network is comparatively simple in structure and takes non temporal data as its input. If the obtained prediction by this network shows that the input image patch is of high confidence of background image, then the algorithm will skip all the rest part of the recognition module.

If the obtained prediction by the first background filtering network shows that the input image patch is not of high confidence of background image, this image data will be sent into the second action recognition network model in the middle row in Fig. 1. This recognition model is much more complicated in structure to help discern different types of actions in each image patch received. Different from the background filtering network, this recognition network expands the input image patch to square size region on the original frame, so that it includes more contextual information to improve recognition actions on some action classes. If the output of this network shows that the input data belongs to some action with a high confidence, then the algorithm will take the corresponding action as the final prediction.

If the output of the spatial recognition network fails to give prediction to some action with high confidence, the designed algorithm will rely on the results of object tracking and use a temporal action recognition model that takes stacked sequence of image patches of some object to recognize its action, as shown in the bottom row in Fig. 1. Because this model takes the stacked sequence of image patches as input, it is better at capturing motion information of the target object. Similar to the



▲**Figure 1. Network fusion for action recognition.**

spatial network, this temporal network requires the input image patches to be expanded to the minimal square bounding box covering the original detection region to include more contextual information.

Finally, the results from the spatial and temporal networks are fused by a linear model to get the final predicted scores on each action class concerned.

Each component of the proposed method is described in detail in the rest of this section.

2.2 Foreground Motion Information Extraction

The first step in the designed algorithm is to detect and track human regions in a surveillance video. The purpose here is to track every person appeared in the surveillance video in real time. Our method is based on the tracking-by-detection philosophy and consists of three parts: foreground motion information extraction, object detection and object tracking. Our object detection and tracking is aided by foreground motion information.

First, we adopt a background subtraction method called ViBe [10] to filter the static background of video frames. Then edge and contour detection is performed on the resulting image. We only keep the centers of contours to represent the detected objects for simplicity.

So far, these centers represent the moving parts of our objects in a single frame. We made the centers more informative by tracking them. We adopt the Kalman Filter [11] to track these center points so that every point is assigned by an ID during its life time. These center points with ID are called "motion points" and play an important role in improving the performance of detection and tracking.

2.3 Object Detection

There are many successful person detectors proposed in previous years [12]. In this work, we adopt the aggregated channel features (ACF) [13] detector due to its advantages on both speed and accuracy. We train the detector on pedestrian dataset consisting INRIA Person and our labeled video. We also propose two methods, filtering and compensation, to improve the detector in our application.

We use filtering to deal with obvious yet common false positives, such as lamp posts, trees and traffic cones. Because our surveillance cameras are stationary, the size of a person is usually linearly related to where he appears in the region. Therefore, objects like traffic cones and trees can be filtered out easily.

In video, rapid appearance change and occlusion of objects often happen. So our person detector cannot detect a person in every frame of a video. We propose a novel method by using the true detections from the detector and the motion points extracted from foreground motion information to compensate for missed detections.

Supposing there is detection D from the detector in frame F, we use the following procedure to determine if the object contained in it is lost in frame $F+1$ or not.

1) Let N be the set of motion points contained in D, $n=|N|$, $k=0$.

2) For every motion point in N, if we can find the motion point in frame $F+1$ with the same ID and belongs to some bounding box, put it into set S, otherwise $k=k+1$.

3) If $k/n<0.5$, most of the motion points are still in some bounding box in frame $F+1$. In this case, we think the bounding box that contains the most points from S is the detection that matches D (**Fig. 2a**).

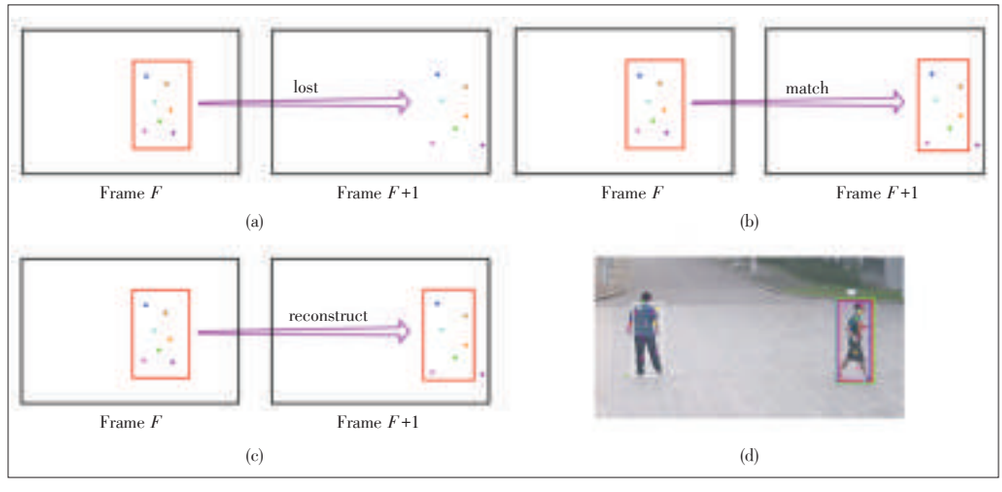
4) If $k/n>0.5$, it is very likely that detection D is lost in frame $F+1$ (**Fig. 2b**). We use set S to reconstruct it.

5) S contains the "moved" motion points of D. For every point in S, we put a bounding box that has the same relative position to it as D around it. These boxes represent the place where D should be in frame $F+1$ with respect to the movement of every motion point. We then find a bounding box that covers all these boxes, rescale it to the size of D, and make it the reconstructed detection of D in frame $F+1$ (**Fig. 2c**).

In **Fig. 2d**, the red box is the detection from last frame, green box is the covering box and blue box is the compensated detection. The idea here is similar to object tracking, however, it only works for moving objects.

2.4 Object Tracking

Our person tracker is an online tracker that integrates the detector's responses and appearance templates. The tracking is performed by assigning every target a tracker with the Kal-



▲**Figure 2. Illustration of human detection with motion information.**

man Filter and the target's appearance templates. At each frame, the tracker will predict the target's new position with the Kalman Filter, calculate a voting map with appearance templates, do mean-shift tracking on the voting map and update the Kalman Filter. Then detections are associated with every tracker. Every tracker updates its appearance template when assigned a new detection. We also adopt the idea of tracker hierarchy [14] to select the most effective tracking strategy.

The result of human detection and tracking is saved as a sequence of object ID and object bounding box, denoted as

$$O(i,t)=\left\{I_t^{(i)},R_t^{(i)}\right\}, \quad (1)$$

where $O(i,t)$ is the target information of object ID i at time step t in the video, and $I_t^{(i)}$ is the image content of this object detected at time step t. While $R_t^{(i)}$ is the bounding box information about this object at time step t in the video. The values in $R_t^{(i)}$ is a four dimensional vector (x,y,w,h) marking the upper left corner location, width and height of the bounding box.

2.5 Background Filtering by Simple Spatial Neural Networks

During the real time recognition process, the extracted target image region is input into the background filtering network model. As shown in **Fig. 3**, the background filtering network is based on single image input, which is the image data returned by the detection and tracking algorithm. Sample input images patches for this network (**Fig. 4**) may include both real human figures and some false alarms from the previous detection and tracking step.

After the input layer, there exist several convolution layers to extract the visual features from the image. To improve the training effect, the rectified linear units (ReLU) and max pooling layer are connected to each convolution layer in the network. Afterwards, several fully connection layers are appended to the last convolution layer to further encode the learnt visual features. In all fully connection layers except the last one, ReLU layers and dropout layers are applied to achieve non-linear transformation of feature values and generalization improvement. The last fully connection layer outputs are connected to a sigmoid layer and transform the output values to class probabilities. In this case, it outputs whether the input image data is foreground human figure or background image. The number of convolution and fully connection layers could be set to meet the computation capacity of the deploying hardware.

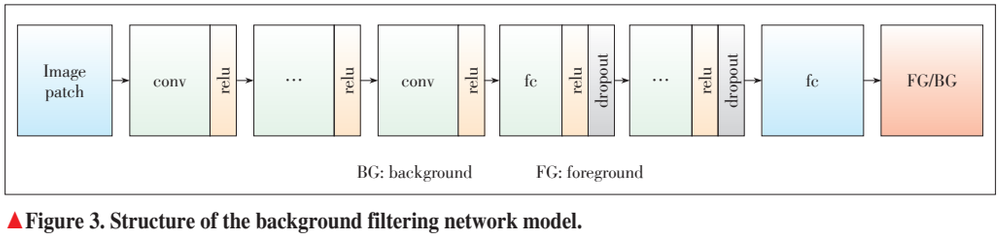
Figure 4. ▶ Sample input data to the background filtering network model.

After this step, the algorithm could filter out the non-foreground areas returned by the detection and tracking algorithm. Therefore, when used in combination, the detection and tracking algorithm could adjust the decision threshold to allow more potential foreground images to pass in order to reduce the missing rate of real human objects. Because this model is computed only on the image regions returned by the detection and tracking algorithm rather than the sliding windows on the whole image, no intensive computation consumption is introduced by this network model during running time.

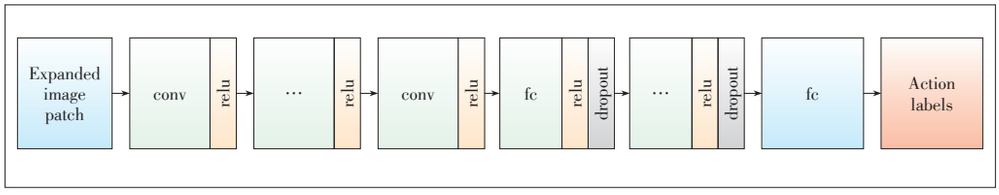
2.6 Spatial Action Recognition Model Based on Single Image

After the background filtering, only foreground figure images are supposed to be obtained. In this step, the algorithm introduces a more complicated recognition network in structure to recognize the action type in the image region. In this network (**Fig. 5**), context information about the target region is also taken into consideration to improve the recognition performance.

The general network framework is very similar to the background filtering network, but with more complicated network structures like more convolution layers and fully connection layers. The input layer of this network expands the input image regions to cover more context information about the target object. Specifically, it expands the detection region to



▲**Figure 3. Structure of the background filtering network model.**



▲Figure 5. Structure of the spatial recognition network model.

the minimal square areas that cover the original object regions. Some sample input image data to the recognition network are shown in **Fig. 6**. As we can see, when expanded to square regions, the input image covers more contextual information to the target human. For example, in the upper image of Fig. 6, the person in red shirt is included as the context of the target person in green at the center of the whole image patch. This contextual information would be very helpful to recognize that the target person's action is to fight with another person. In the lower image of Fig. 6, the expanded square box image region would cover more road context with regard to the person in center, giving more indication that person is walking alone the road and not possible in fighting with someone else.

After the input, the algorithm also uses convolution layers to extract the visual features from the image, together with ReLU and the max pooling layer connected to each convolution layer in the network. Afterwards, fully connection layers are appended to the last convolution layer to further encode the learnt visual features. Also, in all fully connection layers except the last one, ReLU layers and dropout layers are applied to achieve non-linear transformation of feature values and generalization improvement. The last fully connection layer outputs are connected to a sigmoid layer and transform the output values to class probabilities. In this case, it outputs the probability of some specified action classes. The number of convolution and fully connection layers could be set to meet the computation capacity of the deploying hardware. In order to learn more complicated feature representation, this network uses more convolution layers and fully connection layers than the background filtering network model.

As mentioned, the output of the spatial recognition network is a vector of action class probabilities. It is then possible to set certain a decision threshold. When the output value is greater than the specified decision threshold, the algorithm will take this high confidence class label as the final predicted action class for the target person.

▲Figure 6. Sample input data to the spatial recognition network model.

2.7 Temporal Action Recognition Model Based on Stacked Images

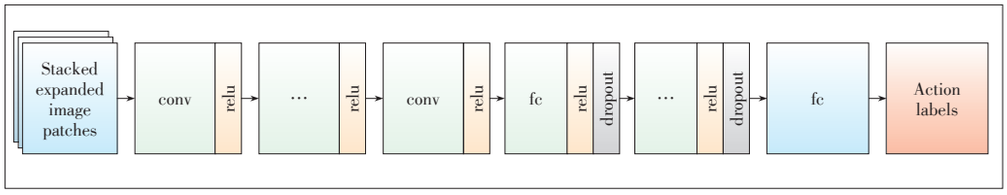
If the previous spatial action recognition model fails to output the action class prediction probabilities higher than the specified threshold, the temporal action recognition model will be introduced to help action predictions. Different from the spatial network model, this temporal model takes stacked image patches as its input data. The stacked image patches are ordered sequence of the target object in the video. Because these sequenced images contain the object information at different time, this temporal recognition model has more advantages at motion information capture. This would be advantageous in distinguishing those actions featured by human motion at a small period of time. The general framework of the proposed temporal model is depicted in **Fig. 7**.

In order to get the stack motion images of a certain object, the temporal recognition model utilizes the tracking results obtained by the tracking algorithm. The stacked image patches could be represented as

$$input(M_3)=\left[I_{t-k}^{(i)};\ldots;I_t^{(i)}\right], \tag{2}$$

where $I_t^{(i)}$ is the image content of this object detected at time step t. This means that the stacked image patches come from the current and history object appearances in a short time interval. In the temporal network, each convolution filter size in the first convolution layer is h by w by c, k, where h, w is the height and width of the filter, c is the number of input image channels, and k is the number of stacked image patches in the input data. Some sample stacked image patches are illustrated in **Fig. 8**. These image patches are also expanded to minimum covering square areas to include contextual information.

As the image patches input into the convolution layers are stacked according to the ordered temporal sequence, content of the first few channels in the input data must contain actions happening earlier than those on the bottom channels. This makes the parameters of the first few channels in the convolution layers always to be applied to action patches happening earlier than those of the last few channels during the training phase. As the interactions of earlier behavior and later behavior are modeled by weighted addition of products between convolution parameters and image content at different time stages in sequenced order, temporal information is automatically and properly learned by this layer architecture in the process of error back-propagation algorithm during network training. Besides, the time span between the image patches also affects the



▲Figure 7. Structure of the temporal recognition network model.

▲Figure 8. Sample input data to the temporal recognition network model.

tuning of convolution parameters, which is part of the learning of temporal information.

After the input, the algorithm also uses convolution layers to extract the visual features from the image, together with ReLU and the max pooling layer connected to each convolution layer in the network. Afterwards, fully connection layers are appended to the last convolution layer to further encode the learnt visual features. Also, in all fully connection layers except the last one, ReLU layers and dropout layers are applied. The last fully connection layer outputs are connected to a sigmoid layer and transform the output values to class probabilities. In this case, it outputs the probability of some specified action classes.

2.8 Fusion of Recognition Models

After the spatial and temporal recognition models are trained, we apply a fusion strategy to better combine the action predictions from them. The basic idea is to use a linear function to aggregate both classification scores returned by the two models. Therefore, for each action to be recognized, the final prediction score after fusion could be written as

$$p(action) = w_S p_S(action) + w_T p_T(action) + b, \quad (3)$$

where $p_S(action)$ and $p_T(action)$ are the action prediction scores returned by the spatial recognition model and temporal recognition model respectively. Besides, w_S and w_T are the weights balancing these two scores, and b is the bias parameter. The values $of\ w_S$, w_T $and\ b$ could be learnt on the validation dataset by using the linear regression model.

3 The Experiment

All the data used in the experiment are collected from outdoor surveillance videos. These videos include actions of walking, running, kicking, fighting and riding.

In order to train the models for recognition, we manually labeled the human actions in a set of videos. In total, the labeled video length is more than 30 hours and we collected more than 5000 labeled trails of human figures in the videos. The tool we used to annotate the videos is the open-sourced ViPER software.

The recognition models are trained from scratch on some existing large scaled image and video datasets and fine-tuned on the surveillance data. For the spatial and temporal recognition models, we first use ImageNet image data to train the structured models from randomly initialized network parameters. Separate and stacked image frames extracted from UCF-101 dataset are then used to fine-tune the networks for the first round. Finally, the pre-trained networks are used to fine-tune our surveillance data with the action labels of interest. During the training, we adopt image patch mirroring, different scaling ratios and random cropping techniques to improve model generalization.

In our experiment, we set five convolution layers and three fully connection layers for the first background filtering network. For the second spatial recognition network model, we set 16 fully connection layers for the three in the first background filtering network. For the third temporal recognition network model, we set five fully connection layers for the three in the first background filtering network to balance recognition performance and time efficiency.

The test set contains videos of around 22 minutes. The obtained detection rate of the proposed method is 0.9751 on the test set. The average recognition precision of the recognition algorithms on the test set is 0.9251. **Table 1** compares recognition precisions on the tested action classes and the performance of the baseline algorithm with ordinary single image based CNN model. As we can see in the table, our proposed method outperforms the baseline model in predicting all the evaluated action classes. For those actions with comparatively distinguishing static appearances such as fighting and riding,

▼Table 1. Comparision of recognition precisions

Action	CNN	Our method
Walk	0.7961	0.8500
Run	0.8332	0.8859
Kick	0.9246	0.9535
Fight	0.9383	0.9596
Ride	0.9558	0.9764
Average	0.8896	0.9251

CNN: convolutional neural networks

the performance of the baseline CNN method is comparatively close to ours. While for some ambiguous actions more dependent on temporal information to distinguish, such as walking and running, our method outperforms the baseline method by a larger margin. Therefore, fusion with temporal information with the stacked images based model is able to improve the classification performance for different kinds of actions.

Fig. 9 shows some recognition results in the test. The red bounding boxes in the figure marks the detection area on the human figures. Different characters stand for different action label predicted. Here "w" stands for walking, "r" stands for running, "k" stands for kicking, "f" stands for fighting, and "b" stands for riding.

4 Conclusions

In the paper, a novel action recognition algorithm is proposed to deal with the automatic video surveillance problems.

▲**Figure 9. Examples of the recognition results.**

The proposed algorithm mainly features in motion compensation during detection, background filtering before recognizing, different fusion settings of spatial and temporal network models for different types of action to recognize. The experiment shows that the proposed method produces good results on the surveillance video data.

References

[1] O. Russakovsky, J. Deng, H. Su, et al. (2014, Sept. 02). *ImageNet large scale visual recognition challenge* [Online]. Available: https://arxiv.org/abs/1409.0575
[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012. doi: 10.1109/TASL.2011.2134090.
[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. (2013, Oct. 17). *Distributed representations of words and phrases and their compositionality* [Online]. Available: https://arxiv.org/abs/1310.4546
[4] A. Karpathy, G. Toderici, S. Shetty, et al., "Large-scale video classification with convolutional neural networks," presented at IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[5] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. (2015, Apr. 07). *Modeling spatial-temporal clues in a hybrid deep learning framework for video classification* [Online]. Available: https://arxiv.org/abs/1504.01561
[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan. 2013. doi: 10.1109/TPAMI.2012.59.
[7] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. (2015, Jun. 05). *Beyond temporal pooling: recurrence and temporal convolutions for gesture recognition in video* [Online]. Available: https://arxiv.org/abs/1506.01911
[8] K. Simonyan and A. Zisserman. (2014, Jun. 09). *Two-stream convolutional networks for action recognition in videos* [Online]. Available: https://arxiv.org/abs/1406.2199
[9] H. Ye, Z. Wu, R.-W. Zhao, et al., "Evaluating two-stream cnn for video classification," presented at 5th ACM on International Conference on Multimedia Retrieval, New York, USA, 2015.
[10] O. Barnich and M. Van Droogenbroeck, "ViBe: a universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011. doi: 10.1109/TIP.2010.2101613.
[11] E. V. Cuevas, D. Zaldivar, and R. Rojas. (2005). *Kalman filter for vision tracking* [Online]. Available: http://www.diss.fu-berlin.de/docs/servlets/MCRFileNodeServlet/FUDOCS_derivate_000000000473/2005_12.pdf
[12] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, Apr. 2012. doi: 10.1109/TPAMI.2011.155.
[13] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532– 1545, Aug. 2014. doi: 10.1109/TPAMI.2014.2300479.
[14] J. Zhang, L. Lo Presti, and S. Sclaroff, "Online multi-person tracking by tracker hierarchy," in *IEEE Ninth International Conference on Advanced Video and Signal - Based Surveillance (AVSS)*, Beijing, China, pp. 379– 385, 2012. doi: 10.1109/AVSS.2012.51.

Manuscript received: 2015-12-08

Biographies

ZHANG Diankai (zhang.diankai@zte.com.cn) received his BE degree in electronic information engineering and MS degree in signal and information processing from Nanjing University of Posts and Telecommunications (NUPT), China in 2006 and 2009. He is a senior video and image algorithm engineer of ZTE Corporation. His research interests include video and image processing, pattern recognition, and computer vision.

ZHAO Rui-Wei (rw.du.zhao@gmail.com) received BS degree in 2005 and MS degree in 2009 from Tongji University, China. He is currently a PhD candidate in the School of Computer Science at Fudan University, China. His research interests include deep learning methods for image and video recognition.

SHEN Lin (shen.lin2@zte.com.cn) received her BE degree in communication engineering and MS degree in computer application from Nanjing University of Science and Technology (NUST), China in 2007 and 2009. She is a senior video and image algorithm engineer of ZTE Corporation. Her research interests include video and image processing, pattern recognition, and computer vision.

CHEN Shaoxiang (forwchen@gmail.com) is currently a research student in the School of Computer Science at Fudan University. His research interests include object detection and tracking in video data.

SUN Zhenfeng (zf_sun@foxmail.com) is currently studying at School of Computer Science, Fudan University towards the degree of Doctor of Engineering. His research interests include multimedia software systems and computer vision.

JIANG Yu-Gang (yugang.jiang@gmail.com) received the PhD degree in computer science from the City University of Hong Kong, China in 2009. During 2008–2011, he was with the Department of Electrical Engineering, Columbia University, USA. He is currently a full professor of computer science at Fudan University. His research interests include multimedia retrieval and computer vision. He is one of the organizers of the annual THUMOS Challenge on Large Scale Action Recognition, and served as a program chair of ACM ICMR 2015. He is the recipient of many awards, including the prestigious ACM China Rising Star Award (2014).

ZTE Communications
Table of Contents, Volume 14, 2016

Volume-Number-Page

SPECIAL TOPICS

ZTE Communications
Table of Contents, Volume 14, 2016

Volume–Number–Page

Multi-Gigabit Millimeter-Wave Wireless Communications

REVIEW

RESEARCH PAPER

ZTE Communications Guidelines for Authors

• Remit of Journal

ZTE Communications publishes original theoretical papers, research findings, and surveys on a broad range of communications topics, including communications and information system design, optical fiber and electro-optical engineering, microwave technology, radio wave propagation, antenna engineering, electromagnetics, signal and image processing, and power engineering. The journal is designed to be an integrated forum for university academics and industry researchers from around the world.

• Manuscript Preparation

Manuscripts must be typed in English and submitted electronically in MS Word (or compatible) format. The word length is approximately 3000 to 8000, and no more than 8 figures or tables should be included. Authors are requested to submit mathematical material and graphics in an editable format.

• Abstract and Keywords

Each manuscript must include an abstract of approximately 150 words written as a single paragraph. The abstract should not include mathematics or references and should not be repeated verbatim in the introduction. The abstract should be a self-contained overview of the aims, methods, experimental results, and significance of research outlined in the paper. Five carefully chosen keywords must be provided with the abstract.

• References

Manuscripts must be referenced at a level that conforms to international academic standards. All references must be numbered sequentially in-text and listed in corresponding order at the end of the paper. References that are not cited in-text should not be included in the reference list. References must be complete and formatted according to ZTE Communications Editorial Style. A minimum of 10 references should be provided. Footnotes should be avoided or kept to a minimum.

• Copyright and Declaration

Authors are responsible for obtaining permission to reproduce any material for which they do not hold copyright. Permission to reproduce any part of this publication for commercial use must be obtained in advance from the editorial office of *ZTE Communications*. Authors agree that a) the manuscript is a product of research conducted by themselves and the stated co-authors, b) the manuscript has not been published elsewhere in its submitted form, c) the manuscript is not currently being considered for publication elsewhere. If the paper is an adaptation of a speech or presentation, acknowledgement of this is required within the paper. The number of co-authors should not exceed five.

• Content and Structure

ZTE Communications seeks to publish original content that may build on existing literature in any field of communications. Authors should not dedicate a disproportionate amount of a paper to fundamental background, historical overviews, or chronologies that may be sufficiently dealt with by references. Authors are also requested to avoid the overuse of bullet points when structuring papers. The conclusion should include a commentary on the significance/future implications of the research as well as an overview of the material presented.

• Peer Review and Editing

All manuscripts will be subject to a two-stage anonymous peer review as well as copyediting, and formatting. Authors may be asked to revise parts of a manuscript prior to publication.

• Biographical Information

All authors are requested to provide a brief biography (approx. 100 words) that includes email address, educational background, career experience, research interests, awards, and publications.

• Acknowledgements and Funding

A manuscript based on funded research must clearly state the program name, funding body, and grant number. Individuals who contributed to the manuscript should be acknowledged in a brief statement.

• Address for Submission

magazine@zte.com.cn
12F Kaixuan Building, 329 Jinzhai Rd, Hefei 230061, P. R. China

ZTE COMMUNICATIONS

ZTE Communications has been indexed in the following databases:

- Cambridge Scientific Abstracts (CSA)
- China Science and Technology Journal Database
- Chinese Journal Fulltext Databases
- Inspec
- Ulrich's Periodicals Directory
- Wanfang Data—Digital Periodicals

ZTE COMMUNICATIONS
Vol. 14 No. S1 (Issue 54)
Quarterly
First English Issue Published in 2003

Supervised by:
Anhui Science and Technology Department

Sponsored by:
Anhui Science and Technology Information Research Institute and ZTE Corporation

---

Editorial Correspondence:
Add: 12F Kaixuan Building, 329 Jinzhai Road,
Hefei 230061, P. R. China
Tel: +86-551-65533356
Fax: +86-551-65850139
Email: magazine@zte.com.cn

Published and Circulated (Home and Abroad) by:
Editorial Office of *ZTE Communications*

Printed by:
Hefei Tiancai Color Printing Company

Publication Date:
December 25, 2016

Publication Licenses:
ISSN 1673-5188
CN 34-1294/ TN

Advertising License:
皖合工商广字0058号

Annual Subscription:
RMB 80