MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
http://www.comsoc.org/~mmc

MMTC Communications - Frontiers

Vol. 12, No. 4, July 2017

CONTENTS

Message from the MMTC Chair

Dear MMTC friends and colleagues:

It is my pleasure to provide a message for the July issue of MMTC Communications-Frontiers. I joined the service for MMTC from 2010. Within the past few years, I have been witnessing the development and growth of MMTC. I am very proud of having been serving in MMTC and deeply enjoy working with MMTC team members for these years. I would like to take this opportunity to express my sincere appreciation to MMTC friends and colleagues! Without your participation and support, MMTC cannot have such a success.

MMTC has fourteen interesting groups (IGs) focusing on different topics in the area of multimedia communications. In addition, MMTC has six boards in charge of award, publication, review, membership, publicity, and advisor, respectively. The number of MMTC members is already above 1000. With the efforts from these IGs and boards, MMTC provides very efficient channels to share, exchange, and discuss information and enhance the visibility of its members.

MMTC will hold TC meeting several times each year during the period of some main conferences such as ICC, GLOBECOM, ICME, etc. The next TC meeting will be hold at ICME 2017 at Hong Kong on July 13, 2017. All are welcome to join this meeting.

If you want to join the MMTC, please visit our Membership Board page at http://committees.comsoc.org/mmc/membership.asp. I have no doubt that you will benefit from being a member of MMTC. Finally, MMTC Communications Frontiers provides the readers the timely update on the start-of-the-art development and hot research topics. I hope you will enjoy reading this issue of MMTC Communications Frontiers!

Sincerely yours,

Fen Hou
Vice Chair for Asia
Multimedia Communications Technical Committee
IEEE Communications Society

SPECIAL ISSUE ON Recent Activities in Mobile Edge Computing and Edge Caching

Guest Editor: Melike Erol-Kantarci, University of Ottawa, Canada

{melike.erolkantarci}@uottawa.ca

The rapid increase in powerful mobile devices along with the demand for rich multimedia applications has escalated the need for efficient computing and caching techniques more than ever. Adding to this, low-latency demand of many Internet of Things (IoT) applications and mobile Augmented Reality and Virtual Reality (AR/VR) leads the mobile network system operators to position themselves more than just communication facilitators but also facilitators of computing and caching closer to the users. On the other hand, device-to-device communications expand the boundaries of computing and caching from the operator equipment to user devices and even cars. Caching of popular contents at the network edge can significantly improve latency performance while mobile edge computing (MEC) or fog computing makes it convenient to access shared pool of services and resources that are location independent. Therefore, MEC and edge caching are important components of the research in 5G and beyond networks. This is discussed in detail in our recent paper "Caching and Computing at the Edge for Mobile Augmented Reality and Virtual Reality in 5G," to be published in ADHOCNETS 2017.

The five papers included in this special issue on "Recent Activities in Mobile Edge Computing and Edge Caching" aim to provide points of views of renowned researchers in this field and to provide the readers cutting-edge results from their groups. The included papers are briefly introduced below.

X. Sun and N. Ansari, in their research "Cloudlet Networks: Empowering Mobile Networks with Computing Capabilities" propose a cloudlet architecture that aims to address three important questions: i) How does MEC incentivize mobile users to upload mobile data to edge computing entities? ii) How does MEC leverage and coordinate the highly distributed edge computing entities at the network edge to analyze the data streams from mobile users? iii) How do mobile users associate with different edge computing entities when the mobile users roam over the network? The authors tackle the virtual machine placement problem in mobile edge computing and propose a delay-aware optimization-based solution.

The paper entitled, "A Dynamic Task Scheduler for Computation Offloading in Mobile Cloud Computing Systems," by H. Shah-Mansouri, V. W.S. Wong, and R. Schober introduce an efficient task scheduler using an optimization framework that takes the energy consumption and delay into account. A task scheduler dynamically makes an offloading decision upon arrival of a task. Therefore the task scheduler has a fundamental role in exploiting the advantages of mobile cloud computing systems. The proposed scheduler is shown to arrive at the optimal offloading decision while maximizing the utility obtained by using the cloud computing services.

In "Online Optimization Techniques for Effective Fog Computing under Uncertainty," authored by G. Lee, W. Saad, and M. Bennis, the problem of operating in a dynamic environment is addressed. In mobile cloud computing, fog nodes can dynamically join and leave a network. Therefore the full information on the location and the future availability of different fog nodes might not be available at all times. However, most of the studies in the literature propose optimization-based approaches with an assumption of all the information being available. In their paper, the authors propose, using online optimization, to capture the dynamically varying and largely uncertain environment of fog networks. The paper introduces use cases for online optimization, as well as discussing its use jointly in caching and fog computing.

In "Human-enabled Edge Computing: When Mobile Crowd-Sensing meets Mobile Edge Computing," the authors L. Foschini and M. Girolami report on recent their findings on Human-driven Edge Computing (HEC). HEC relies on continuously monitoring humans and their mobility patterns to dynamically re-identify hot locations and to use a human-in-the-loop approach. The proposed approach leverages human sociality and mobility to broaden the coverage of the fixed mobile edge computing architectures.

The research in "Mobile Edge Computing: Recent Efforts and Five Key Research Directions," by T. X. Tran, M.-P. Hosseini, and D. Pompili presents the recent state-of-the-art in mobile edge computing. The authors begin by introducing proofs of concepts and standardization efforts. Then they discuss the research on computation offloading as well as edge caching. Finally, they outline future research directions as a valuable guideline for the researchers who are interested in the area.

The purpose of this special issue is to introduce several state-of-the-art research efforts in mobile edge computing and edge caching rather than giving a complete coverage of the area. The contributions of the widely recognized researchers make the special issue a valuable source for the readers. The guest editor is thankful for all the authors for their valuable contributions and the help from the MMTC Communications – Frontiers Board.

Melike Erol-Kantarci is an assistant professor at the School of Electrical Engineering and Computer Science at the University of Ottawa, ON, Canada. She is the founding director of the Networked Systems and Communications Research (NETCORE) laboratory. She is also a courtesy assistant professor at the Department of Electrical and Computer Engineering at Clarkson University, Potsdam, NY, where she was a tenure-track assistant professor prior to joining University of Ottawa. She received her Ph.D. and M.Sc. degrees in Computer Engineering from Istanbul Technical University in 2009 and 2004, respectively. During her Ph.D. studies, she was a Fulbright visiting researcher at the Computer Science Department of the University of California Los Angeles (UCLA). She is an editor of the IEEE Communications Letters and IEEE Access. She is the co-editor of the book "Smart Grid: Networking, Data Management, and Business Models". Her articles are continuously among the top cited and top accessed papers on IEEE and Elsevier databases. She has acted as general chair or technical program chair for many international conferences and workshops. She is a senior member of the IEEE and the past vice-chair for Women in Engineering (WIE) at the IEEE Ottawa Section. She is currently the vice-chair of Green Smart Grid Communications special interest group of IEEE Technical Committee on Green Communications and Computing. She is also the research group leader for IEEE Smart Grid and Big Data Standardization. Her main research interests are 5G and beyond wireless networks, smart grid, cyber-physical systems, electric vehicles, Internet of things and wireless sensor networks.

Cloudlet Networks: Empowering Mobile Networks with Computing Capabilities

Xiang Sun and Nirwan Ansari
Advanced Networking Laboratory
Helen and John C. Hartmann Department of Electrical & Computer Engineering
New Jersey Institute of Technology, Newark, NJ 07102, USA
{xs47, nirwan.ansari}@njit.edu

1. Introduction

Mobile devices are currently embedded with various sensors to sense the environment over time. Analyzing these sensed data can substantially transform how we do business and conduct our lives. For instance, analyzing the data generated by on-body sensors can enable early detection of unusual activities or abnormalities, thus improving our health [1]; analyzing the photos/videos captured by mobile users can detect and track terrorists to safeguard the whole society. Traditionally, these mobile data would be uploaded to a remote data center, which has been demonstrated to provision resources flexibly and efficiently, for further processing [2]. However, this would burden the network to conduct data aggregation from mobile users to the remote data center, thus significantly increasing the response time of generating high-level knowledge (or providing services) by analyzing the mobile data. The response time (of generating high-level knowledge) is very important for mobile data analytics [3]-[5], e.g., identifying the terrorists and obtaining their locations along with related timestamps (by analyzing the photos/videos from users) in a timely fashion is very critical in deterring terrorism.

Mobile Edge Computing (MEC) has been proposed to enable computing entities (e.g., cloudlets and fog nodes) to process mobile data streams at the network edge [6], [7]. This can tremendously reduce the time for uploading the mobile data from mobile users to the computing entities, thus potentially reducing the response time accordingly. However, the MEC concept is still in the phase of proof of concept and many issues need to be addressed:

- *Issue-1: How does MEC incentivize mobile users to upload mobile data to edge computing entities?*
  Mobile users would like to share their original data to applications in order to receive the corresponding services, which are provided by the applications. However, sharing original data may provide personal information of mobile users to the application providers. This may discourage mobile users from uploading mobile data. For instance, the terrorist detection application is to identify and track terrorists by comparing the photos of the terrorists with the ones captured by mobile users. Thus, mobile users need to upload their photos (which contain the personal information of mobile users) to the terrorist detection application, which can be placed at the edge computing entities. Therefore, it is beneficial to design a data sharing mechanism tailored for MEC such that mobile users can obtain services provided by the applications while preserving privacy of mobile users.
- *Issue-2: How does MEC leverage and coordinate the highly distributed edge computing entities at the network edge to analyze the data streams from mobile users?*
  Edge computing entities may be highly distributed in the mobile network. Each edge computing entity provides computing resources to process data streams from its local mobile users. Different edge computing entities may need to coordinate with each other in order to provide services to mobile users with low delay. For instance, the terrorist detection application may need to collect and analyze the photos/videos captured by different mobile users in a large area, which includes a number of distributed edge computing entities. Transmitting all the photos/videos (captured by the different mobile users) to the terrorist detection application (which is located in a specific edge computing entity) may not provide a low response time in identifying the terrorists because of the high network delay for transmitting high volume data (i.e., photos/videos) to the terrorist detection application (in a specific edge computing entity). Thus, we need to design a distributed computing architecture tailored for MEC such that different edge computing entities can coordinate with each other to reduce the response time.
- *Issue-3: How do mobile users associate with different edge computing entities when the mobile users roam over the network?*
  In order to minimize the delay between a mobile user and an edge computing entity as well as the traffic in the core network, a mobile user may associate with the closest edge computing entity. That is, a mobile user may

need to change its associated edge computing entity to process its mobile data when the mobile user roams from one area into another. However, associating mobile users with their closest edge computing entities may result in insufficient resource provisioning of edge computing entities (i.e., some edge computing entities do not have enough resources to process the data streams from their local mobile users). Meanwhile, changing associated edge computing entity incurs extra computing and communications overheads. Therefore, designing an efficient edge computing entity association strategy is critical to speed up mobile data processing.

In this paper, we introduce a cloudlet network to address these three issues. The rest of the paper is organized as follows. In Sec. 2, we introduce the cloudlet network architecture to resolve Issue-1 and Issue-2. In Sec. 3, in order to resolve Issue-3, we propose to associate mobile users with different edge computing entities by migrating mobile users' Avatars (i.e., Virtual Machines (VMs)) among edge computing entities based on mobile users' locations. We formulate the Avatar placement problem and demonstrate its performance via simulations.

2. The Cloudlet Network Architecture

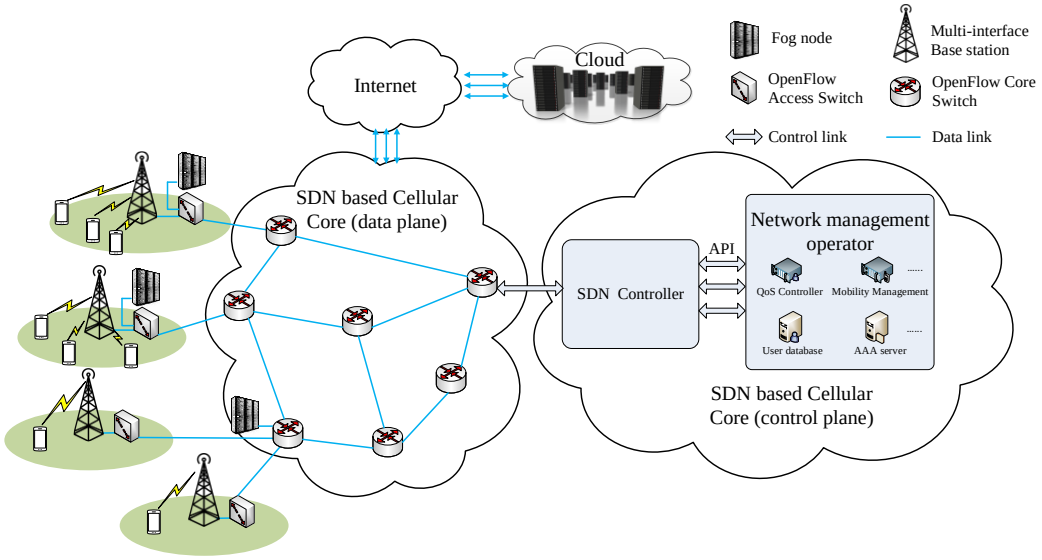


Figure. 1: The cloudlet network architecture (cf. Fig. 3 in [7]).

The cloudlet network architecture, as shown in Figure 1, comprises three parts, i.e., distributed cloudlets in the mobile network, hierarchical structure of a cloudlet, and the Software Defined Networking (SDN) based mobile core network [8]. We will next detail these three parts.

2.1 Distributed cloudlets in the mobile network

A tremendous number of Base Stations (BSs) have already been deployed in the mobile network and provide high radio coverage, i.e., every mobile user can communicate with a BS everywhere. Meanwhile, with the development of 5G technologies, the speed of the mobile access network would be much higher as compared to the existing 4G LTE system. These facts justify that deploying cloudlets (i.e., edge computing entities) at BSs in the mobile network would be a suitable solution to provide computing resources to mobile users with high availability and low latency. Specifically, each BS is connected to a cloudlet [9], which comprises a number of interconnected Physical Machines (PMs). The deployment of cloudlets is flexible, i.e., a BS can access to its local cloudlet via an access switch or many BSs can be connected the same cloudlet, which is located at the edge of the mobile core network.

Data centers are located at remote sites (which are commonly connected to the core network directly) to provide the scalability and availability of the system. Specifically, the computing and storage capacities of the local cloudlets are limited, and thus they may not have enough capacities to efficiently analyze mobile data streams. Data centers, which supply sufficient and flexible resource provisioning, can be considered as backup units to process mobile data streams.

2.2 Hierarchical structure of a cloudlet

As shown in Figure 2, a cloudlet consists of two logical layers, i.e., Avatar layer and Application VM layer. The

Avatar layer comprises a number of Avatars. An Avatar is considered as a private VM associated with a specific mobile user [3]. Thus, each mobile user can upload its generated data to its Avatar[1] periodically or upon requests. The Avatar would pre-process the received mobile data in order to generate metadata, in which the personal information from raw data has been parsed and trimmed, and then send them to the application VM upon requests. The application VM layer comprises a number of application VMs, which are deployed by the application providers. The application VMs are to retrieve metadata from Avatars, analyze the received metadata to generate high-level knowledge, and provide the corresponding services to mobile users.

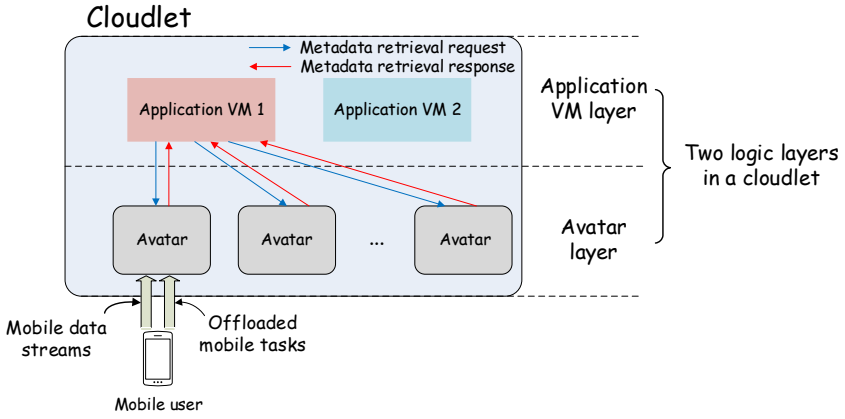


Figure. 2: The hierarchical structure of a cloudlet (cf. Fig. 4 in [7]).

There are two methods for uploading and processing mobile data streams in a hierarchical cloudlet structure, i.e., *passive uploading* and *proactive uploading*. Here, we provide the terrorist detection application to illustrate how the *passive uploading* and *proactive uploading* methods work.

- The *passive uploading* method comprises five steps: 1) if a mobile user is interested in the terrorist detection application, it can install the application in its Avatar. The application installed in the Avatar is to conduct face matching by comparing different photos. 2) If the terrorist detection application VM tries to find a terrorist, it would send a metadata retrieval request, which contains a set of photos of the terrorist, to the Avatars (which have installed the application). 3) After receiving the metadata retrieval request, the Avatar would send a data retrieval request to its mobile user in order to obtain the recent captured photos and videos from the mobile user. 4) The mobile user would upload the corresponding photos and videos via the BS after it receives the data retrieval request. 5) The Avatar would execute the face-matching algorithm by comparing the terrorist's photos with the photos and videos uploaded from the mobile user. If matched, the Avatar would respond to the application VM with the related metadata, i.e., the location information and time stamps of the matched photos and videos.
- In the *proactive uploading* method, after the application has been installed in the Avatar (i.e., Step-1 in *passive uploading*), the mobile user is to proactively upload its data in terms of captured photos and videos (i.e., the photos and videos would be immediately uploaded and stored in the Avatar once they are captured). Once the application VM sends the metadata retrieval request to the Avatar (i.e., Step-2 in *passive uploading*), the Avatar would execute the face-matching algorithm immediately (i.e., Step-5 in *passive uploading*).

The above mentioned two methods have their own pros and cons: the *proactive uploading* method would improve the response time to detect terrorists because data (e.g., photos and videos) of mobile users have already been in their Avatars when the Avatars receive the metadata retrieval requests.. However, Avatars need to store all the data streams generated by their mobile users. Avatars thus become heavily loaded VMs, which would significantly increase the storage resource requirement in the cloudlets and increase the overheads of migrating Avatars among cloudlets (which will be discussed in the next section). On the other hand, the *passive uploading* method may incur

[1] It is worth to note that each Avatar can not only analyze the data generated by its mobile user but also execute the tasks offloaded from its mobile user [10], i.e., an Avatar acts two roles in the network: the data stream analyzer and the mobile task outsourcer. This may incentivize mobile users to subscribe their own Avatars.

higher response time as compared to the *proactive uploading* method. However, Avatars do not have to store data streams generated by their mobile users in the *passive uploading* method. Hence, Avatars are considered to be lightly loaded VMs, thus generating less overheads in migrating Avatars among cloudlets. Our previous study [3] focused on the implementation of the *proactive uploading* method, but we will focus on the *passive uploading* method in this paper, i.e., Avatars do not proactively store data streams from their mobile users.

The proposed hierarchical cloudlet architecture structure addresses Issue-1 and Issue-2. Specifically, each mobile user's Avatar is considered as a private VM to facilitate resource isolation and access control. The Avatar would analyze the raw data (which are generated from its mobile user) and provide metadata (which do not contain personal information) to the application VM. This resolves Issue-1, i.e., mobile users can obtain services provided by the application VMs while preserving privacy of mobile users. In addition, each Avatar can be a worker node of an application VM, which acts as a master node to distribute workloads to the Avatars (which have been installed the corresponding application) and aggregate metadata from the Avatars, and provide services to users. This distributed computing structure resolves Issue-2 by fully utilizing the distributed computing resources in the cloudlets and significantly reducing the traffic load of the network as compared to the current way in which the application VM directly retrieves the data from mobile users, analyzes them, and provides services to mobile users. Moreover, the distributed computing structure provisions the flexibility for the application in placing their application VMs. For instance, if the terrorist detection application tries to determine whether the terrorists appeared in area-1 and area-2, it can create two application VMs in the two cloudlets, which are located in area-1 and area-2, respectively. Each application VM can send the metadata retrieval requests to their Avatars in their cloudlets. After the search is completed, the application VMs can be removed accordingly.

2.3 SDN based mobile core network

Instead of applying the traditional cellular core network architecture, which leads to inefficient, inflexible, and unscalable packet forwarding, the SDN based mobile core network is adopted in the cloudlet network. The SDN based mobile core network is essentially decoupling the control plane from the switches, which only run data plane functionalities. The control plane is offloaded to a logical central controller, which transmits the control information (e.g., flow tables) to the OpenFlow switches by applying the OpenFlow protocol [11], monitors the traffic statistics of the network, and provides Application Programming Interfaces (APIs) to network management operators. Thus, different mobile network functionalities, such as mobility management, user authentication, authorization and accounting, network virtualization, and QoS control, can be added, removed, and modified flexibly.

3. Adaptive Avatar Placement

Mobile users are roaming among BSs over time, and statically positioning mobile users' Avatars in their original cloudlets may significantly increase the delay between Avatars and their mobile users. Note that delay is an important factor to determine the QoS of many mobile applications. On the other hand, if the Avatar always follows its mobile user's movement (e.g., once a mobile user roams from BS-1's coverage area into BS-2's coverage area, its Avatar would migrate from the cloudlet, which is associated with BS-1, into the cloudlet, which is associated with BS-2, accordingly), a significant amount of migration overhead will be generated during the Avatar migration. The migration overhead may degrade the performance of the Avatar in executing applications because the Avatar migration requires a significant amount of bandwidth resource and non-negligible computing as well as memory resource of the Avatar [12], [13]. This may deprive the applications (which run in the Avatar) with less computing, memory as well as bandwidth resource during the Avatar migration, thus degrading the performance of the applications. The migration overhead of an Avatar is then modeled as a function of Avatar migration time and the resource requirements of the Avatar (before the migration) [14]. Note that longer migration time implies that the migration consumes more resource of the Avatar, and thus may degrade the performance of applications (which run in the Avatar).

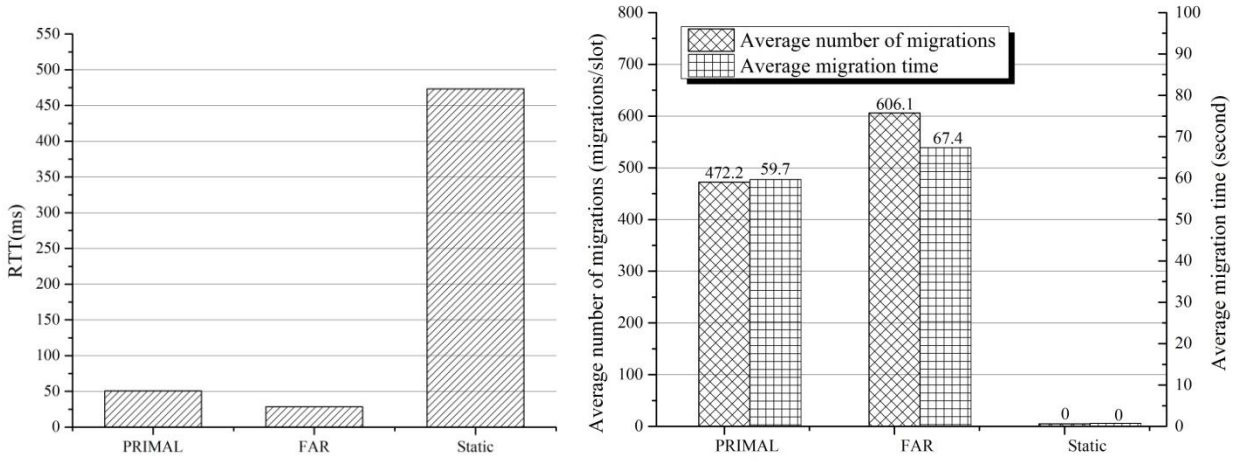
We consider the gain of the Avatar migration to be the End-to-End (E2E) delay reduction between the Avatar and its mobile user and the cost of the Avatar migration to be the migration overhead. Thus, we can formulate the Avatar placement problem[2] as follows:

Given: 1) each mobile user's location indicator (i.e., the location of the BS that covers the mobile user) in the next

[2] The Avatar placement problem is to determine the location of each Avatar (i.e., each Avatar is placed in which cloudlet) for each mobile user in the next time slot. If the location of an Avatar in the next time slot is different from that in the current time slot, we say the Avatar should be migrated into in the new location in the next time slot.

time slot; 2) the average E2E delay between each cloudlet and each BS; 3) the capacity limitation of each cloudlet (i.e., the maximum number of Avatars hosted by a cloudlet).
Obtain: the placement indicator of each Avatar x_{ij} (i.e., $x_{ij} = 1$ indicates mobile user i's Avatar is hosted by cloudlet j; else $x_{ij} = 0$) in the next time slot.
Objective: maximize the profit (the gain minus the cost) of all the Avatar migrations.
Constraints: 1) each Avatar should be placed in only one cloudlet (i.e., $\sum_j x_{ij} = 1$); 2) the total number of Avatars assigned to each cloudlet cannot exceed its capacity (i.e., $\sum_i x_{ij} = s_j$, where s_j is the capacity of cloudlet j).

The detailed formulation of the Avatar placement can be found in [14]. Note that the Avatar placement problem can be formulated as a mixed integer linear programming problem and can be solved by applying CPLEX [15]. In order to demonstrate the performance of the PRofIt Maximization Avatar pLacement (PRIMAL) strategy (i.e., the solution of the mentioned Avatar placement problem), we compare it with two other strategies, i.e., Static and Follow me AvataR (FAR), via simulations. The idea of the FAR strategy is to minimize the E2E delay between an Avatar and its mobile user by assigning the Avatar to the available cloudlet (i.e., the cloudlet has enough space to host the Avatar), which yields the lowest E2E delay. The Static strategy is to avoid the migration cost, i.e., the locations of Avatars do not change over time after they are initially deployed. The detailed simulation setups can be found in [14]. Figure 3(a) shows the average Round Trip Time (RTT) between mobile users and their Avatars incurred by PRIMAL, FAR, and Static, respectively. Figure 3(b) shows the average number of migrations and the average migration time incurred by PRIMAL, FAR, and Static, respectively. We conclude that, from Figure 3(a) and 3(b), PRIMAL can achieve the similar average RTT as compared to FAR, but it incurs fewer number of Avatar migrations and shorter average migration time. Note that although Avatar migration is not triggered by Static, the RTT incurred by Static is unbearable.



(a) (b)

Figure 3: Simulation results

4. Conclusion

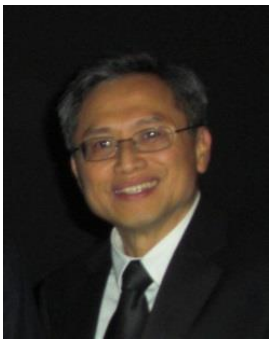
In this paper, we have introduced the cloudlet network to bring the computing resource from centralized cloud to mobile users in order to minimize the delay between mobile users and computing resource. We have also designed the hierarchical structure to address Issue-1 and Issue-2. Furthermore, we have proposed to migrate mobile users' Avatars among edge computing cloudlets to resolve Issue-3. We have also proposed the green cloudlet network by introducing green energy into the cloudlet network to reduce the operational cost of maintaining the distributed cloudlets [16] and we have designed the energy-aware Avatar migration strategy to fully utilize green energy [17], [18].

References

[1] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain and K. S. Kwak, "The Internet of Things for Health Care: A Comprehensive Survey," *IEEE Access*, vol. 3, pp. 678-708, 2015.
[2] X. Sun, N. Ansari and R. Wang, "Optimizing Resource Utilization of a Data Center," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2822-2846, Fourth Quarter 2016.
[3] X. Sun and N. Ansari, "Adaptive Avatar Handoff in the Cloudlet Network," *IEEE Transactions on Cloud Computing*, doi: 10.1109/TCC.2017.2701794, early access.

[4] K. Ha, *et al.*, "Adaptive VM handoff across cloudlets," Technical Report CMU-CS-15-113, CMU School of Computer Science, 2015.
[5] X. Sun and N. Ansari, "Latency Aware Workload Offloading in the Cloudlet Network," *IEEE Communications Letters*, doi: 10.1109/LCOMM.2017.2690678, early access.
[6] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys & Tutorials*, doi: 10.1109/COMST.2017.2682318, early access.
[7] X. Sun and N. Ansari, "EdgeIoT: Mobile Edge Computing for the Internet of Things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22-29, December 2016.
[8] X. Jin, L.E. Li, L. Vanbever, and J. Rexford, "Softcell: Scalable and flexible cellular core network architecture," in *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, Santa Barbara, CA, Dec. 09-12, 2013, pp. 163-174.
[9] M. Satyanarayanan, P. Bahl, R. Caceres and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14-23, Oct.-Dec. 2009.
[10] X. Sun and N. Ansari, "Energy-optimized bandwidth allocation strategy for mobile cloud computing in LTE networks," in *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, New Orleans, LA, 2015, pp. 2120-2125.
[11] A. Lara, A. Kolasani and B. Ramamurthy, "Network Innovation using OpenFlow: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 493-512, First Quarter 2014.
[12] A. Anand, J.Lakshmi, and S.K. Nandy, "Virtual Machine Placement Optimization Supporting Performance SLAs," in *2013 IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom)*, Bristol, UK, Dec. 2-5, 2013, vol. 1, pp. 298-305.
[13] L. Haikun, H. Jin, C.Z. Xu, and X. Liao, "Performance and energy modeling for live migration of virtual machines," *Cluster computing*, vol. 16, no. 2 pp. 249-264, 2013.
[14] X. Sun and N. Ansari, "PRIMAL: PRofIt Maximization Avatar pLacement for mobile edge computing," in *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, 2016, pp. 1-6.
[15] CPLEX Optimizer. Available. [Online]: https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/.
[16] X. Sun and N. Ansari, "Green Cloudlet Network: A Distributed Green Mobile Cloud Network," *IEEE Network*, vol. 31, no. 1, pp. 64-70, January/February 2017.
[17] X. Sun, N. Ansari and Q. Fan, "Green Energy Aware Avatar Migration Strategy in Green Cloudlet Networks," in *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, Vancouver, BC, 2015, pp. 139-146.
[18] Q. Fan, N. Ansari, and X. Sun, "Energy Driven Avatar Migration in Green Cloudlet Networks," *IEEE Communications Letters*, doi: 10.1109/LCOMM.2017.2684812, early access.

Xiang Sun [S'13] received a B.E. degree in electronic and information engineering and an M.E. degree in technology of computer applications from Hebei University of Engineering, Hebei, China. He is currently working towards the Ph.D. degree in electrical engineering at the New Jersey Institute of Technology (NJIT), Newark, New Jersey. His research interests include mobile edge computing, big data networking, green edge computing and communications, and cloud computing.



Nirwan Ansari [S'78,M'83,SM'94,F'09] is Distinguished Professor of Electrical and Computer Engineering at the New Jersey Institute of Technology (NJIT). He has also been a visiting (chair) professor at several universities.

Professor Ansari has authored Green Mobile Networks: A Networking Perspective (John Wiley, 2017) with T. Han, and co-authored two other books. He has also (co-)authored more than 500 technical publications, over 200 published in widely cited journals/magazines. He has guest-edited a number of special issues covering various emerging topics in communications and networking. He has served on the editorial/advisory board of over ten journals. His current research focuses on green communications and networking, cloud computing, and various aspects of broadband networks.

Professor Ansari was elected to serve in the IEEE Communications Society (ComSoc) Board of Governors as a member-at-large, has chaired ComSoc technical committees, and has been actively organizing numerous IEEE International Conferences/Symposia/Workshops. He has frequently been delivering keynote addresses, distinguished lectures, tutorials, and invited talks. Some of his recognitions include IEEE Fellow, several Excellence in Teaching Awards, a few best paper awards, the NCE Excellence in Research Award, the ComSoc AHSN TC Outstanding Service Recognition Award, the IEEE TCGCC Distinguished Technical

Achievement Recognition Award, the COMSOC TC Technical Recognition Award, the NJ Inventors Hall of Fame Inventor of the Year Award, the Thomas Alva Edison Patent Award, Purdue University Outstanding Electrical and Computer Engineer Award, and designation as a COMSOC Distinguished Lecturer. He has also been granted over 30 U.S. patents.

A Dynamic Task Scheduler for Computation Offloading in Mobile Cloud Computing Systems

Hamed Shah-Mansouri[*], *Vincent W.S. Wong*[*], *and Robert Schober*[†]
[*]*Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada*
[†]*Institute for Digital Communications, Friedrich-Alexander University of Erlangen–Nuremberg, Germany*
Email: [*]*{hshahmansour, vincentw}@ece.ubc.ca,* [†]*robert.schober@fau.de*

1. Introduction

Mobile cloud computing (MCC) provides computing services to mobile devices to enable them to perform their computation-intensive applications. By utilizing cloud computing services, the limited processing power of mobile devices is no longer a barrier for the rapid development of future applications. The mobile devices can offload their computation tasks to the cloud servers to benefit from powerful computing resources, save their battery power, and expedite the task execution. There are various cloud-assisted mobile platforms including ThinkAir [1] and MAUI [2] that realize computation task offloading in mobile environments. The module making the offloading decision is called the *task scheduler*. The task scheduler dynamically makes an offloading decision upon arrival of a task. In order to fully exploit the advantages of MCC systems, the design of an efficient task scheduler is crucial.

Computation task offloading in MCC has been widely studied in the literature. In [3], the authors developed an offloading decision strategy that aims to minimize the energy consumption of a mobile device while meeting a latency deadline. In [4], the authors proposed an energy-delay aware mechanism for computation task offloading in MCC systems. Although the proposed mechanism addresses the energy-delay tradeoff, all computing tasks face the same tradeoff between energy consumption and delay regardless of their different sensitivities to delay. In [5], the authors designed a centralized controller hosted in cloud servers and optimized the mobile users' offloading decisions by minimizing the overall cost. An application-aware computation offloading mechanism that balances the tradeoff between energy efficiency and responsiveness of mobile applications was proposed in [6]. In [7], the authors studied energy-efficient computation offloading under a completion time deadline constraint.

An efficient task scheduler should take the energy saving obtained from task offloading and the delay of a task into account to arrive at the optimal offloading decision. However, the aforementioned task schedulers do not consider the heterogeneous latency requirements of different delay-sensitive applications. In this paper, we design an efficient task scheduler by using an optimization framework that takes the energy consumption and delay into account. We consider both delay-sensitive and delay-tolerant applications and address their different latency requirements. We then evaluate the performance of our proposed task scheduler through numerical studies. We further compare the proposed task scheduler with ThinkAir [1] and MAUI [2] and illustrate its superiority in terms of energy consumption and delay for delay-sensitive and delay-tolerant applications.

2. Dynamic Task Scheduler

In this section, we first introduce the system model of a mobile device. We then design the task scheduler using a utility maximization framework and obtain the optimal offloading strategy.

To model the task scheduler, we consider a queuing system with two servers, as illustrated in Fig. 1. The first server is the centralized processing unit (CPU) of the mobile device and represents the local task execution. The second server, which is the wireless interface (e.g., WiFi, Long-Term Evolution (LTE)), is used to model the task offloading to the cloud servers. We classify the mobile user's workload into three categories: Offloadable computing tasks, CPU workload, and network traffic. We assume that they arrive according to independent Poisson processes with rate λ^{o}, λ^{c}, and λ^{nt}, respectively. The offloadable computing tasks can either be processed by the local CPU or be offloaded to the cloud servers, while the other workloads have to be processed by their allocated servers. We denote the size of each task by z (in bits) and its processing density by γ (in cycles/bit), which is the number of CPU cycles required to process a unit bit of data. We model z and γ as random variables that follow probability density functions (pdf) $f_Z(z)$ and $f_\Gamma(\gamma)$, respectively. The service time in the CPU is $\gamma z/C$, where C is the CPU processing capacity (in cycles/time unit). Similarly, the service time in the wireless interface is z/μ, where μ (in bits/sec) is the data rate of the wireless interface. We consider a slow flat fading wireless channel model and assume that the data rate of the wireless interface remains constant during the transmission of a task. However, μ is a random variable with pdf $f_M(\mu)$. We further assume that the data rate is known to the scheduler upon arrival of the tasks.

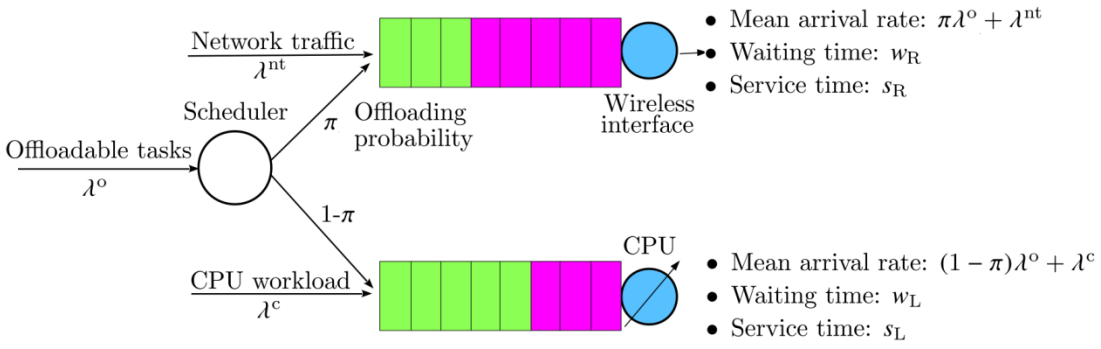


Figure 1. The task scheduler and queuing system of a mobile device.

We model the sensitivity of each task to delay by a parameter θ, which depends on the type of application [8]. The value of θ follows pdf $f_\Theta(\theta)$ and is known to the scheduler upon arrival of the task. A large value of θ represents applications with stringent delay requirements, whereas applications with $\theta = 0$ are tolerable to delay.

We now introduce the offloading decision indicator to indicate whether or not a task is offloaded. We denote the offloading decision indicator for a task of size z with processing density γ and delay parameter θ when the data rate is μ as $\delta(z,\gamma,\theta,\mu) \in \{0,1\}$. We set $\delta(z,\gamma,\theta,\mu) = 1$ when the task is offloaded to the cloud servers. We also denote the probability of $\delta(z,\gamma,\theta,\mu) = 1$ by $\pi(z,\gamma,\theta,\mu) \in [0,1]$. We further define the offloading probability π as the average probability that a task is offloaded to the cloud servers. We have

$$\pi = \int_{\mathbb{R}_+^4} \pi(z,\gamma,\theta,\mu) \mathrm{d}F_{Z,\Gamma,\Theta,M}(z,\gamma,\theta,\mu), \quad (1)$$

where $\mathrm{d}F_{Z,\Gamma,\Theta,M}(z,\gamma,\theta,\mu)$ is the joint cumulative distribution function of random variables z, γ, θ, and μ.

The task scheduler makes the offloading decisions based on a utility maximization framework. The utility is the weighted sum of the energy consumption saving and delay improvement obtained by offloading the tasks to the cloud servers. It reflects the benefit of computation task offloading. We first define the energy consumption saving as the energy consumed in the CPU to execute the task minus the transmission energy required to submit the task to the cloud servers. We follow the energy consumption models used in [4] and [9]. The energy consumed in the CPU to execute a task of size z and processing density γ is $\kappa\gamma z/C$, where κ is a constant and depends on the CPU model. Similarly, the energy consumed to transmit a task of size z to the cloud servers is $\beta z/\mu$, where β depends on the type of the wireless interface.

We further determine the delay improvement, which is the difference between the time required to complete a task locally and the time spent to process the task in the cloud servers. We first obtain the delay a task experiences if it is offloaded to the cloud servers. We assume that the cloud servers are located in close proximity of the mobile device. Thus, the delay consists of the following terms: the queuing delay of the wireless interface, the time required to transmit the task to the cloud servers, and the processing time in the cloud servers to complete the task. Notice that the time required to retrieve the results from the cloud servers is negligible since the downlink rate is usually much higher than the uplink rate. As shown in Fig. 1, the arrival at the wireless interface queue consists of two workloads: the offloaded tasks that arrive according to a Poisson process with rate $\pi\lambda^{o}$ when the offloading probability π is given, and the network traffic. Combining these two independent Poisson processes forms another Poisson process [10] with arrival rate $\pi\lambda^{o} + \lambda^{nt}$. Thus, when π is a given constant, we model the wireless interface queue as an *M/G/1* queuing system. Let $w_R(\pi)$ and $s_R(z,\mu) = z/\mu$ denote the waiting time and the service time of this queue, respectively. The queue is stable if $\pi\lambda^{o} + \lambda^{nt} < 1/\mathbb{E}[s_R]$ holds, where $\mathbb{E}[s_R]$ denotes the mean service time. In this case, the mean waiting time can be obtained using the Pollaczek-Khinchin formula [10] and is as follows:

$$\mathbb{E}[w_R(\pi)] = \frac{(\pi\lambda^{o} + \lambda^{nt})\mathbb{E}[s_R^2]}{2(1-(\pi\lambda^{o} + \lambda^{nt})\mathbb{E}[s_R])}.$$

In addition to the queuing delay and service time, we consider the time required to process the task in the cloud servers and denote it by $s_C(z,\gamma) = \gamma z/C_R$, where C_R denotes the processing capacity (in cycles/unit time) of the cloud servers. Thus, given π, if a task is offloaded, its total delay is $\mathbb{E}[w_R(\pi)] + s_R(z,\mu) + s_C(z,\gamma)$.

Similar to the wireless interface queue, we obtain the delay to perform the task locally in the CPU queue. The CPU queue is an *M/G/1* queuing system with arrival rate $(1-\pi)\lambda^{\mathrm{o}}+\lambda^{\mathrm{c}}$, when π is given. We also define $w_L(\pi)$ and $s_L(z,\gamma)$ as the waiting time and service time, respectively. According to the Pollaczek-Khinchin formula, when the queue is stable (i.e., $(1-\pi)\lambda^{o}+\lambda^{c}<1/\mathbb{E}[s_L]$), the mean waiting time in the CPU queue is given by:

$$\mathbb{E}[w_L(\pi)]=\frac{\left((1-\pi)\lambda^{\mathrm{o}}+\lambda^{\mathrm{c}}\right)\mathbb{E}[s_L^2]}{2\left(1-\left((1-\pi)\lambda^{\mathrm{o}}+\lambda^{\mathrm{c}}\right)\mathbb{E}[s_L]\right)}.$$

Moreover, the service time of a task of size z and processing density γ is $s_L(z,\gamma)=\gamma z/C$. Thus, given π, the delay introduced by performing the task locally is $\mathbb{E}[w_L(\pi)]+s_L(z,\gamma)$.

We now calculate the delay improvement obtained by offloading the task to the cloud servers. Given offloading probability π, the delay improvement, denoted by $\tau(z,\gamma,\mu,\pi)$, is as follows:

$$\tau(z,\gamma,\mu,\pi)=\mathbb{E}[w_L(\pi)]+s_L(z,\gamma)-\left(\mathbb{E}[w_R(\pi)]+s_R(z,\mu)+s_C(z,\gamma)\right).$$

As mentioned earlier, the utility consists of the energy consumption saving and delay improvement. Given π, we denote the utility obtained from offloading a task of size z with processing density γ and delay parameter θ as $u^{\pi}(z,\gamma,\theta,\mu)$ when the data rate is μ. We have

$$u^{\pi}(z,\gamma,\theta,\mu)=\frac{\kappa\gamma z}{C}-\frac{\beta z}{\mu}+\theta\tau(z,\gamma,\mu,\pi).$$

The scheduler makes the offloading decision upon arrival of each task with the goal of maximizing the utility obtained from offloading the task. Notice that if a task is performed locally (i.e., $\delta(z,\gamma,\theta,\mu)=0$), the utility is zero. Thus, the task scheduler solves the following problem:

$$\begin{aligned}&\text{maximize}\quad \delta(z,\gamma,\theta,\mu)u^{\pi}(z,\gamma,\theta,\mu)\\&\text{subject to}\ \ \delta(z,\gamma,\theta,\mu)\in\{0,1\}.\end{aligned}\tag{2}$$

Let $\delta^{*}(z,\gamma,\theta,\mu)$ denote the optimal offloading decision indicator. By solving problem (2), we have

$$\delta^{*}(z,\gamma,\theta,\mu)=\begin{cases}1, & \text{if } u^{\pi}(z,\gamma,\theta,\mu)\geq 0\\ 0, & \text{otherwise.}\end{cases}\tag{3}$$

To obtain the optimal offloading decision indicator, we need to know the offloading probability π. According to (3), if the utility obtained from offloading a task is non-negative, the scheduler offloads the task to the cloud servers. This happens with probability $\pi(z,\gamma,\theta,\mu)$. Thus, $\pi(z,\gamma,\theta,\mu)=1$ if $u^{\pi}(z,\gamma,\theta,\mu)\geq 0$. We define the offloading region $\mathcal{O}(\pi)$ as the set of (z,γ,θ,μ) such that $u^{\pi}(z,\gamma,\theta,\mu)\geq 0$, i.e.,

$$\mathcal{O}(\pi)=\{(z,\gamma,\theta,\mu)\in\mathbb{R}_{+}^{4}|\ u^{\pi}(z,\gamma,\theta,\mu)\geq 0\}.$$

We now obtain the optimal offloading probability in the following theorem.

Theorem 1. *The optimal offloading probability*, denoted by π^{*}, *can be uniquely obtained by solving the following equation when we substitute* $\pi(z,\gamma,\theta,\mu)$ *into (1):*

$$\pi^{*}=\int_{\mathbb{R}_{+}^{4}}\pi(z,\gamma,\theta,\mu)\mathrm{d}F_{Z,\Gamma,\Theta,M}(z,\gamma,\theta,\mu)=\int_{\mathcal{O}(\pi^{*})}\mathrm{d}F_{Z,\Gamma,\Theta,M}(z,\gamma,\theta,\mu).\tag{4}$$

Proof. Please refer to [11] for the proof of Theorem 1.

3. Performance Evaluation

In this section, we investigate the performance of the proposed task scheduler. We assume that the mobile device has a CPU with clock speed $C=1.4$ GHz. We further assume that z,γ, and θ follow uniform distributions in [100 B, 1 MB], [100, 3000] cycles/bit, and [0, 5], respectively, while we fix $\mu=2$ Mbps. The other simulation parameters are $\kappa=1005$ mJ/sec, $\beta=2605$ mJ/sec [4], $\lambda^{\mathrm{c}}=0.08$, $\lambda^{\mathrm{nt}}=0.02$, and $C_R=4$ GHz.

We compare the performance between our proposed task scheduler and task scheduling policies ThinkAir–E, ThinkAir–D, and ThinkAir–ED proposed in [1]. ThinkAir–E prioritizes energy conservation and offloads the tasks if the energy consumption is expected to improve. ThinkAir–D optimizes the offloading decision in order to expedite the execution of the tasks. However, ThinkAir–ED is an energy-delay aware offloading mechanism and offloads the computing tasks only if both the energy consumption and the execution time are expected to improve.

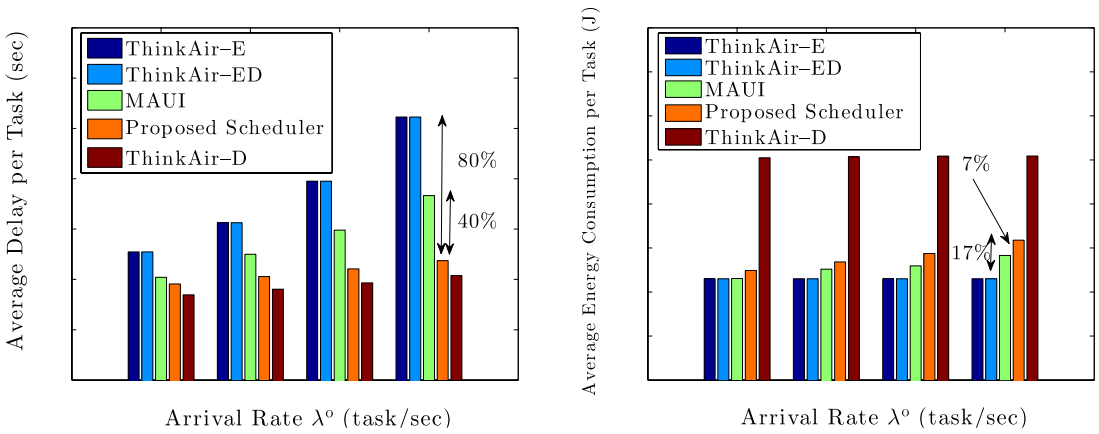


Figure 2. The average delay per task and the corresponding average energy consumption per task versus different arrival rates of the computing tasks.

We further compare our proposed task scheduler with MAUI [2]. To make an offloading decision, the MAUI solver aims to minimize the mobile device's energy consumption subject to a latency constraint. The solver ensures that the total delay experienced by each task does not exceed an application-dependent constant delay, denoted by L. In order to compare our proposed scheduler with MAUI, we set $L = 1/\theta$ for each application.

Fig. 2 illustrates the average delay and the average energy consumption for different arrival rates of the computing tasks. The proposed scheduler substantially outperforms ThinkAir–D in terms of energy consumption. Furthermore, our proposed scheduler consumes slightly more energy than ThinkAir–E, however, it is better able to meet the requirements of delay-sensitive tasks. As can be observed from Fig. 2, our proposed scheduler reduces the delay by 80% and 40% compared to ThinkAir–E and MAUI, while it only consumes 17% and 7% more energy, respectively.

Fig. 3 shows the average energy consumption per task for different schemes and the corresponding average delay. We vary the delay parameter θ from 0 to 10^3 to investigate the tradeoff between the energy consumption and delay for different applications. The proposed scheduler consumes more energy than ThinkAir-E to expedite the execution of delay-sensitive tasks, while it behaves similar to ThinkAir–D when θ is very large. Moreover, for delay-tolerant tasks, the proposed task scheduler consumes the same amount of energy as ThinkAir-E.

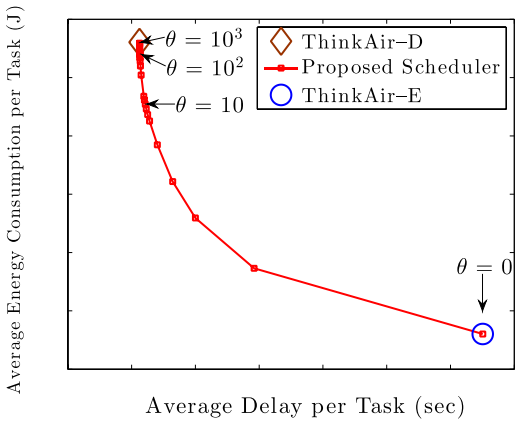


Figure 3. The tradeoff between the energy consumption and delay.

4. Conclusion

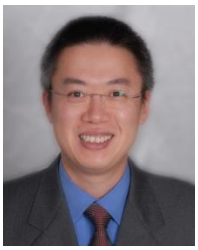
In this paper, we proposed a dynamic task scheduler for computation task offloading in MCC systems. We considered both delay-sensitive and delay-tolerant applications and designed the task scheduler based on a utility maximization framework. The proposed scheduler arrives at the optimal offloading decision when maximizing the

utility obtained by using the cloud computing services. We further investigated the performance of the proposed task scheduler through numerical experiments. Our results showed that the proposed task scheduler outperforms existing task scheduling policies in terms of energy consumption and delay. More details of the design of the task scheduler as well as additional simulation results can be found in [11].

References

[1] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *Proc. of IEEE INFOCOM*, Orlando, FL, Mar. 2012.
[2] E. Cuervo, A. Balasubramanian, D.-K. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making smartphones last longer with code offload," in *Proc. of MobiSys*, San Francisco, CA, Jun. 2010.
[3] W. Zhang and Y. Wen, "Energy-efficient task execution for application as a general topology in mobile cloud computing," accepted for publication in *IEEE Trans. Cloud Computing*, 2015.
[4] Y. Kim, J. Kwak, and S. Chong, "Dual-side dynamic controls for cost minimization in mobile cloud computing systems," in *Proc. of WiOpt*, Mumbai, India, May 2015.
[5] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading decision and resource allocation for multi-user multi-task mobile cloud," in *Proc. of IEEE ICC*, Kuala Lumpur, Malaysia, May 2016.
[6] L. Tong and W. Gao, "Application-aware traffic scheduling for workload offloading in mobile clouds," in *Proc. of IEEE INFOCOM*, San Francisco, CA, Apr. 2016.
[7] S. Guo, B. Xiao, Y. Yang, and Y. Yang, "Energy-efficient dynamic offloading and resource scheduling in mobile cloud computing," in *Proc. of IEEE INFOCOM*, San Francisco, CA, Apr. 2016.
[8] N. Tran, C. S. Hong, Z. Han, and S. Lee, "Optimal pricing effect on equilibrium behaviors of delay-sensitive users in cognitive radio networks," *IEEE J. Select. Areas Commun.*, vol. 31, no. 11, pp. 2566–2579, Nov. 2013.
[9] J. Kwak, O. Choi, S. Chong, and P. Mohapatra,"Dynamic speed scaling for energy minimization in delay-tolerant smartphone applications," in *Proc. of IEEE INFOCOM*, Toronto, Canada, Apr. 2014.
[10] D. P. Bertsekas and R. G. Gallager, *Data Networks*, 2nd ed. Prentice-Hall Inc., 1992.
[11] H. Shah-Mansouri, V. W.S. Wong, and R. Schober, "Joint optimal pricing and task scheduling in mobile cloud computing systems," accepted for publication in *IEEE Trans. Wireless Commun.*, 2017.

Hamed Shah-Mansouri received the B.Sc., M.Sc., and Ph.D. degrees from Sharif University of Technology, Tehran, Iran, in 2005, 2007, and 2012, respectively all in electrical engineering. Currently, Dr. Shah-Mansouri is a Post-doctoral Fellow at the University of British Columbia, Vancouver, Canada. His research interests are in the area of stochastic analysis, optimization and game theory and their applications in cellular networks, mobile cloud computing systems, and Internet of Things. He has served as the publication co-chair for the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) 2016 and as the technical program committee (TPC) member of the IEEE Globecom'15, IEEE VTC–Fall ('16, '17), and IEEE PIMRC'17.



Vincent W.S. Wong is a Professor in the Department of Electrical and Computer Engineering at the University of British Columbia, Vancouver, Canada. His research areas include protocol design, optimization, and resource management of communication networks, with applications to wireless networks, smart grid, mobile cloud computing, and Internet of Things. Dr. Wong is an Editor-at-Large of the *IEEE Transactions on Communications*. He has served as a Guest Editor of *IEEE Journal on Selected Areas in Communications* and *IEEE Wireless Communications*. He is the Chair of the IEEE Communications Society Emerging Technical Sub-Committee on Smart Grid Communications and the IEEE Vancouver Joint Communications Chapter. He received the 2014 UBC Killam Faculty Research Fellowship. He is a fellow of the IEEE.

Robert Schober is an Alexander von Humboldt Professor and the Chair for Digital Communication with the Friedrich-Alexander Universität Erlangen-Nürnberg (FAU), Erlangen, Germany. From 2002 to 2011, he was a Professor and Canada Research Chair with The University of British Columbia, Vancouver, Canada. His research interests include the broad areas of communication theory, wireless communications, and statistical signal processing. From 2012 to 2015, Dr. Schober served as an Editor-in-Chief of the *IEEE Transactions on Communications* and since 2014, he is the Chair of the Steering Committee of the *IEEE Transactions on Molecular, Biological, and Multiscale Communications*. Furthermore, he is a member-at-large of the Board of Governors of the IEEE Communications Society. He is a fellow of the IEEE, fellow of the Canadian Academy of Engineering, and a fellow of the Engineering Institute of Canada.

Online Optimization Techniques for Effective Fog Computing under Uncertainty

Gilsoo Lee[1], Walid Saad[1], and Mehdi Bennis[2]
[1] Department of Electrical and Computer Engineering, Virginia Tech, USA, {gilsoolee, walids}@vt.edu
[2] Centre for Wireless Communications, University of Oulu, Finland, bennis@ee.oulu.fi

1. Introduction

The Internet of Things (IoT) environment will encompass billions of devices that are expected to generate more than two Exabyte of data per day [1]. Fog computing is a promising approach to perform distributed computation and caching for supporting IoT applications such as self-driving vehicle communications and drone flight control as well as enabling augmented reality (AR) or virtual reality (VR) services [2]. To meet the ultra-low latency communication and computing requirements of such applications, relying on cloud computing will no longer be possible due to the round-trip delay needed to reach the cloud data center. Thus, fog computing has been proposed as an extension of the cloud computing. In fog computing, some of the cloud's functionalities such as caching, control, and computing are migrated to edge fog nodes [2]. Therefore, by pooling the computing resources of fog nodes located in proximity of one another at the edge of a wireless network, fog computing can achieve low-latency data transmission and computation.

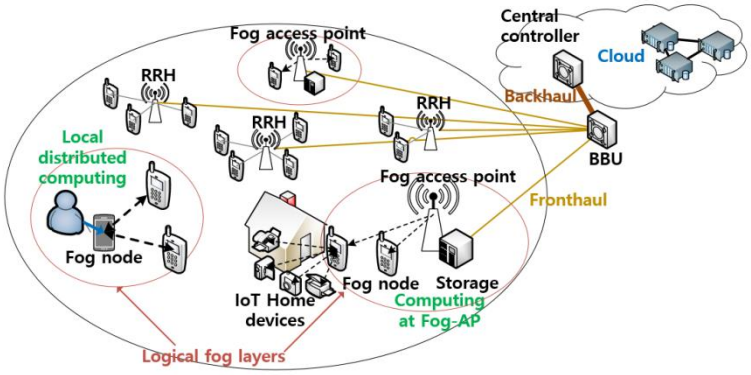


Fig. 1: Example of distributed fog computing and its integration with cloud computing.

To enable low-latency fog computing, a fog-centric radio access network (FRAN) architecture was introduced in [3]. Fig. 1 shows an illustration of an FRAN-based fog computing network consisting of fog nodes and a central cloud that operates a cloud-based radio access network (CRAN). In Fig. 1, fog nodes can be clustered for pooling their resources. Then, *distributed fog computing* is used to process the computing tasks of a fog node with low latency. For instance, the user of a fog node can run a big data application that requires highly intense computing tasks. If the fog node does not have sufficient computing resources, it can find other neighboring fog nodes that have idle computing resources that they are willing to share. After discovering the neighboring nodes, the computing tasks can be distributed to the neighbors and computed in a distributed way to achieve low computational latency. Also, *caching at the fog layer* can be used to reduce the computational latency. For example, in Fig. 1, fog nodes can be associated with a fog access point (AP) that has a local data storage that it can use for caching. Whenever a computing task of a fog node needs to have input data, it is possible that the data is not stored at the fog node. In this case, caching the data at the fog AP can reduce the data transmission latency, thus resulting in a low computational latency. Moreover, in an FRAN, caching can be done by any fog node having a data storage. Thus, whenever fog nodes perform the functions of caching or computing for other devices including many IoT home appliances or sensors, low computational latency can be achieved, if the system is properly designed.

2. Online Optimization Frameworks for Fog Computing

Fog computing will generally need to operate in highly dynamic networking environments. For instance, distributed computing can be performed among heterogeneous fog nodes such as smartphones, tablets, and other IoT devices and embedded systems. Therefore, co-existing fog nodes may have different computing resources. When those fog nodes participate in distributed computing, each fog node is able to individually control and manage its computing resources. In consequence, the amount of shared computing resources of a fog node can be controlled by each individual fog node. In this case, the fog network will generally not be able to know, in advance, the available computing resource shared by each fog node. Also, fog nodes such as handheld devices and vehicles can be mobile. While fog nodes are moving, if a fog node moves beyond a maximum communication distance, it will no longer be

able to join the fog computing network. Therefore, fog nodes can dynamically join and leave a network and, thus, it is challenging to know, a priori, the full information on the location and the future availability of different fog nodes. Thus, complete information on the fog computing environment is not always known to the involved fog nodes. Under such lack of information availability, fog nodes will not be able to use conventional offline optimization techniques, for computing, task distribution, or caching purposes, as such techniques typically require at least some form of information availability (full information or statistical) at the level of each fog node. Thus, to capture the dynamically varying and largely uncertain environment of fog networks, one can rely on the powerful tools of *online optimization* [4]. In online optimization problems, newly updated information is revealed to the system in a sequential manner. The sequentially arriving information becomes the input of an online problem. Thus, unlike offline optimization problems, online problems can be updated according to the input. Therefore, when an input is initially unavailable and revealed sequentially, an online algorithm must be used for decision making and optimization purposes. Hence, for an online cost minimization problem *P*, when an input set *I* is given in an online manner and the online algorithm yields feasible output O_A, the objective function can be shown as $C(I, O_A)$ where *C* is a function of the online input and the output of the algorithm [5].

For an online algorithm solving the online problem, competitive analysis can be used to measure the performance of an online algorithm. Using competitive analysis, the performance of the optimal offline algorithm is compared to the performance of the online algorithm, and the ratio between online and offline algorithms is known as a *competitive ratio* [5]. For instance, for a cost minimization problem, the costs of an online algorithm and optimal offline algorithm are denoted by $ALG(I)=C(I,O_A)$ and $OPT(I)=$ min $C(I,O_A)$, respectively. Then, the competitive ratio is *c* satisfying $c \geq ALG(I)/OPT(I), \forall I$. Whenever $c=1$, then the online algorithm and offline algorithm achieve the same performance. However, optimality cannot be readily achieved in online since the online algorithm is run without having the complete knowledge about the future events. Thus, the goal of developing online algorithms is to minimize the value of *c* in order to achieve a suboptimal solution whose performance is as close as possible to the optimal, offline approach. Also, when an input is given, the decision should be determined by an online algorithm before the next input arrives. In practice, the input arrival interval can be very short. In that case, the online algorithm needs to have low time complexity to make a decision in real time. Particularly, using a low time-complexity algorithm can further reduce the decision-making latency, thus helping to achieve the ultra-low latency in fog computing. This provides a key rationale for adopting online optimization solutions for distributed and real-time fog computing. In the next section, we introduce fog computing applications in which online optimization tools are a natural choice.

3. Applications of Online Optimization

3.1 Fog Computing for IoT Devices

The IoT environment will include many small sensors for smart home. smart building control system, smart transportation, and even wearables that can deliver a wide range of services to the end-users. Considering a network that consists of a sensor layer and a fog layer, the IoT sensors with low computing power will generally seek to offload their computational tasks to other fog nodes in the fog layer such as smartphones or APs. By doing so, the tasks from the sensors can be computed with more powerful computing resources provided by the more capable fog nodes. The tasks are first offloaded from sensors to a certain fog node. Then, the fog node that received the tasks initiates distributed fog computing. This node will be referred to as the initial fog node. To compute the tasks over the fog network, the initial fog node distributes the received tasks to other neighboring fog nodes. During task distribution, the transmission latency can be defined as the sum of the waiting time before the tasks are transmitted and the wireless transmission delay incurred by the transmission from the initial fog node to another fog node. Once a neighbor successfully receives the tasks from the initial fog node, the computation procedure will further incur a computational latency that includes the waiting time in the computational queue and the actual processing time.

In this typical use case of fog computing, the initial fog node first needs to select an optimal set of neighboring fog nodes and distribute the tasks to this selected set so as to minimize the latency. However, in practice, neighboring fog nodes can dynamically join and leave the fog computing network. Therefore, a given fog node will typically not be able to know when neighboring fog nodes will join the network and also where those neighboring nodes are located. Thus, when the arrival of neighboring fog nodes is uncertain, the information cannot be known to the initial fog node. This uncertainty of the location and the order of arrival of neighboring fog nodes can be modeled as a sequential input of an online optimization problem. Then, an online optimization problem can be formulated to minimize the total computational latency defined as the sum of the transmission latency and computational latency.

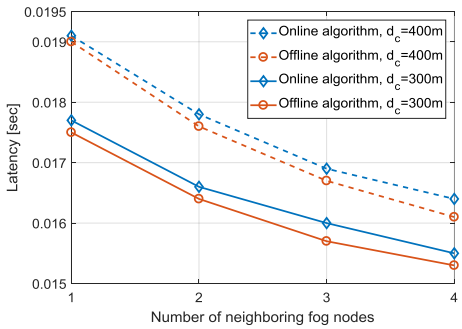


Fig. 2: Maximum latency (computation and transmission) when tasks are offloaded to neighboring fog nodes and cloud. The transmission latency to the cloud increases with d_c that is the distance between the initial fog node and the base station that is connected to the cloud.

In [6], the fog network formation and task distribution problems are jointly studied to minimize the maximum total computational latency when the initial fog node can offload tasks to neighboring nodes and cloud. To enable the initial fog node to form a local fog network when the arrival process of neighboring nodes is randomly determined, an online framework based on the *online secretary problem* is proposed in [6]. In secretary problems, while an interviewer meets the applicants in a random sequential order, the interviewer should determine whether to hire the applicant or not. If a candidate is not hired, the interviewer cannot recall the rejected applicant. By using the analogy between the secretary problem and the fog network formation problem, in [6], we proposed an online algorithm that first observes the performance of some fog nodes and select the fog nodes by using the information gathered during the observation stage. This algorithm also repeatedly optimizes the task distribution whenever new fog node join fog computing. The simulation result in Fig. 2 shows that the online algorithm can successfully minimize the latency while achieving the a latency that is similar to the optimal latency that can be found by the offline algorithm using complete information of neighboring nodes. Fig. 2 also shows that when the transmission latency to the cloud decreases by changing the distance to the cloud d_c from 400m to 300m, the maximum total latency can be reduced for all network sizes. In this work, to formulate the online optimization problem, the uncertain location and future availability of neighboring fog nodes are modeled as an online input. This online approach can be extended and applied for networks with moving fog nodes, e.g., autonomous cars or unmanned vehicles. In such scenarios, due to mobility, fog nodes are unable to know the future location of neighbors as well aso the time duration that each fog node can dedicate for participation in fog computing. Thus, online optimization can be used for moving fog nodes to solve fog network formation and task distribution problems. Naturally, other similar IoT scenarios with uncertainty can be modeled via online optimization.

3.2 Caching Fog Computation

In fog computing, fog nodes can further reduce their latency by caching the input data needed to process their computational operations. The computational operation of the application running on a fog node can have one corresponding input data (e.g. one file). In this case, the goal is to fetch the necessary input data, and this caching technique can be viewed as *data caching*. For such data caching, an interesting information-theoretic latency analysis is provided in [7] when the contents are delivered through a fronthaul and a wireless channel. In this analysis, it is assumed that the set of popular file is static; thus, it can be seen as offline caching. However, if the popularity of data changes within a short time interval, the data requests can be modeled as an online input. For scenarios in which fog nodes are unable to know full knowledge on user requests, online optimization can be used to develop online caching strategy. For instance, the work in [8] proposes an online scheme to minimize the data delivery time by replenishing the fog node's cache and scheduling the delivery of the requested files.

In contrast to data caching in which any given operation has a single input set, a more realistic scenario will include multiple files that can possibly be used for processing a computational operation at a fog node. Then, from those files, an operation can select a specific file as input data. When each input file represents the computational intermediate result (IR) of an operation, fog nodes can cache the computation by storing the possible input files, and this caching technique can be viewed as *computational caching* [9]. If the computational operations allow reusing the IR from previous computational operations, computational caching can be used to reduce the computational latency at a fog node. Therefore, by downloading IRs from neighbors, the fog node can avoid redundant computation, and the computational latency can be reduced. However, since downloading cached computation from

neighbors incurs an additional cost due to transmission latency, the fog node should properly decide whether or not to download IRs. Also, in computational caching, the uncertain arrival order of different computational operations can be modeled as an online input. Therefore, under uncertain future computation, we can consider the online computational caching problem that minimizes the transmission and computational latency where an online algorithm makes the downloading and caching decisions. This online computational caching can be applied to video applications. When the randomness of content requests is modeled as an online input, the online data caching can be used to fetch the requested data with low latency. Furthermore, by using online computational caching, the fog nodes can download the requested data in the form of IR from the neighboring nod; thus further optimizing the computational latency.

3.3 Caching at Energy Harvesting Fog Nodes

To reduce energy consumption of fog computing, mobile network operators can deploy self-power fog APs that use the harvested energy from ambient energy sources, e.g., solar or wind. However, when self-powered fog APs are not connected to the conventional power grid, they can be turned off due to the unexpected energy outage. Thus, the fog network may no longer be able to use the data cached at APs that are no longer operational due to power outage. This uncertainty on energy harvesting makes it very challenging to operate self-powered fog computing networks. To partially address this challenge, the problem of maximizing the caching payoff of self-powered APs is studied in [10]. In this work, the proposed algorithm enables self-powered fog APs to decide whether to accept the arriving requests from users and also whether to cache the downloaded data. While it is assumed that the energy level of APs increases by a constant unit in [10], in practice, the energy arrival process can be randomly determined. Thus, it can be difficult to know the future energy status. Here, the randomness of energy arrival can be modeled as online input [11]. By doing so, an online problem can be formulated to minimize the latency when self-powered fog APs decide whether to cache the data. Similarly, if fog nodes use energy harvesting, the online energy arrival can be applied to other applications. For example, self-powered fog nodes can be used to relay the data traffic to a server. Then, the uncertainty of the harvested energy is modeled as the online energy arrival. Therefore, one can consider an online problem that optimizes the data routing path when the energy states of fog nodes are uncertain. As such, online optimization provides powerful mechanisms for handling uncertain energy arrivals in energy harvesting fog networks.

4. Summary

In this paper, we have introduced the framework of online optimization as a powerful tool for operating distributed fog computing networks in dynamic and uncertain environments. In particular, we have discussed three key applications of online optimization for fog computing. First, we have shown how online optimization can be used to study distributed fog computing when the location and presence of fog nodes have uncertainty. Then, we have shown that online optimization can be used to enable caching in fog computing when future caching requests are unknown to the fog node. Finally, we have shown that online optimization can be used to perform caching at energy harvesting APs when the energy arrival at the self-power fog AP is random. In summary, online optimization tools are expected to play a key role in future fog computing networks primarily due to the need for low latency operation and the online models that capture the dynamic and uncertain environments.

Acknowledgement

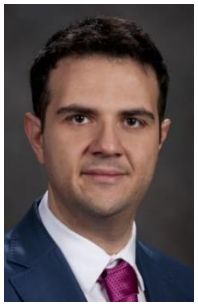
This research was supported by the U.S. National Science Foundation under Grant CNS-1460333.

References

[1] Cisco, "Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are," Cisco white paper, 2015.
[2] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," IEEE Internet of Things Journal, vol. 3, pp. 854-864, Dec. 2016.
[3] M. Peng, S. Yan, K. Zhang and C. Wang, "Fog-computing-based radio access networks: issues and challenges," IEEE Network, vol. 30, pp. 46-53, July 2016.
[4] E. C. Hall and R. M. Willett, "Online Convex Optimization in Dynamic Environments," IEEE Journal of Selected Topics in Signal Processing, vol. 9, pp. 647-662, June 2015.
[5] A. Borodin and R. El-Yaniv, Online computation and competitive analysis, Cambridge University Press, 2005.
[6] G. Lee, W. Saad and M. Bennis, "An Online Secretary Framework for Fog Network Formation with Minimal Latency," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, Paris, 2017.
[7] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog Radio Access Networks," in Proc. IEEE Int. Symp. on Inform. Theory, Barcelona, 2016.
[8] S. M. Azimi, O. Simeone, A. Sengupta and R. Tandon, "Online Edge Caching in Fog-Aided Wireless Network,"

arXiv:1701.06188 [cs], Jan. 2017.
[9] G. Lee, W. Saad and M. Bennis, "Online Optimization for Low-Latency Computational Caching in Fog Networks," 2017.
[10] D. Niyato, D. I. Kim, P. Wang and M. Bennis, "Joint admission control and content caching policy for energy harvesting access points," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, Kuala, 2016.
[11] G. Lee, W. Saad, M. Bennis, A. Mehbodniya and F. Adachi, "Online Ski Rental for ON/OFF Scheduling of Energy Harvesting Base Stations," IEEE Trans. Wireless Commun., vol. 16, pp. 2976-2990, May 2017.

Gilsoo Lee (S'13) received his BS and MS in Electronic Engineering from Sogang University, South Korea. He is currently a PhD student at the Bradley department of Electrical and Computer Engineering at Virginia Tech, VA, USA. His research interests include optimization and game-theoretic approach in energy-harvesting networks, energy-efficient networks, and fog computing.



Walid Saad (S'07, M'10, SM'15) received his Ph.D degree from the University of Oslo in 2010. Currently, he is an Associate Professor at the Department of Electrical and Computer Engineering at Virginia Tech, where he leads the Network Science, Wireless, and Security (NetSciWiS) laboratory, within the Wireless@VT research group. His research interests include wireless networks, game theory, cybersecurity, unmanned aerial vehicles, and cyber-physical systems. Dr. Saad is the recipient of the NSF CAREER award in 2013, the AFOSR summer faculty fellowship in 2014, and the Young Investigator Award from the Office of Naval Research (ONR) in 2015. He was the author/co-author of five conference best paper awards at WiOpt in 2009, ICIMP in 2010, IEEE WCNC in 2012, IEEE PIMRC in 2015, IEEE SmartGridComm in 2015, and EuCNC in 2017. He is the recipient of the 2015 Fred W. Ellersick Prize from the IEEE Communications Society. In 2017, Dr. Saad was named College of Engineering Faculty Fellow at Virginia Tech. From 2015 – 2017, Dr. Saad was named the Steven O. Lane Junior Faculty Fellow at Virginia Tech. He currently serves as an editor for the IEEE Transactions on Wireless Communications, IEEE Transactions on Communications, and IEEE Transactions on Information Forensics and Security.

Mehdi Bennis (SM'15) received his M.Sc. degree in Electrical Engineering jointly from the EPFL, Switzerland and the Eurecom Institute, France in 2002. From 2002 to 2004, he worked as a research engineer at IMRA-EUROPE investigating adaptive equalization algorithms for mobile digital TV. In 2004, he joined the Centre for Wireless Communications (CWC) at the University of Oulu, Finland as a research scientist. In 2008, he was a visiting researcher at the Alcatel-Lucent chair on flexible radio, SUPELEC. He obtained his Ph.D. in December 2009 on spectrum sharing for future mobile cellular systems. Currently Dr. Bennis is an Adjunct Professor at the University of Oulu and Academy of Finland research fellow. His main research interests are in radio resource management, heterogeneous networks, game theory and machine learning in 5G networks and beyond. He has co-authored one book and published more than 100 research papers in international conferences, journals and book chapters. He was the recipient of the prestigious 2015 Fred W. Ellersick Prize from the IEEE Communications Society and the 2016 Best Tutorial Prize from the IEEE Communications Society. Dr. Bennis serves as an editor for the IEEE Transactions on Wireless Communication.

Human-enabled Edge Computing: When Mobile Crowd-Sensing meets Mobile Edge Computing

Luca Foschini[1], Michele Girolami[2]
[1]Dipartimento di Informatica: Scienza e Ingegneria (DISI), University of Bologna, Italy
[2]ISTI-CNR, Pisa, Italy
luca.foschini@unibo.it, michele.girolami@isti.cnr.it

1. Introduction

MEC is an architectural model and specification proposal (i.e., by European Telecommunications Standards Institute - ETSI) that aims at evolving the traditional two-layers cloud-device integration model, where mobile nodes directly communicate with a central cloud through the Internet, with the introduction of a third intermediate middleware layer that executes at so-called network edges. This promotes a new three-layer device-edge-cloud hierarchical architecture, which is recognized as very promising for several application domains [1]. In fact, the new MEC model allows moving and hosting computing/storage resources at network edges close to the targeted mobile devices, thus overcoming the typical limitations of direct cloud-device interactions, such as high uncertainty of available resources, limited bandwidth, unreliability of the wireless network trunk, and rapid deployment needs.

Although various MEC solutions based on fixed edges enable an increase of the quality and performance of several cloud-assisted device services, currently there are still several non-negligible weaknesses that affect this emerging new model. First, the number of edges is generally limited because edges are deployed statically (usually by telco providers) and their configuration and operation introduce additional costs for the supported services, such as deployment, maintenance, and configuration costs. Second, once deployed, edges are rarely re-deployed (due to the high re-configuration cost) in other positions and this might result in high inefficiency, e.g., as service load conditions might significantly change dynamically. Finally, some geographical areas might become interesting hotspots for a service only during specific time slots, such as a square becoming crowded due to an open market taking place only at a specific timeslot and day of the week.

At the same time, the possibility to leverage people roaming though the city with their sensor-rich devices has recently enabled Mobile Crowd-Sensing (MCS). In fact, by installing an MCS application, any smartphone can become part of a (large-scale) mobile sensor network, partially operated by the owners of the phones themselves. However, for some high-demanding MCS applications (e.g., a surveillance service that, for security purposes, monitors an environment with smartphone cameras that capture photos/videos of the surroundings and exploits face recognition to trace suspicious users' movements), regular smartphones often have not enough capabilities to timely perform the requested local tasks, in particular if considering their possible immersion in hostile environments with possible frequent intermittent disconnections from the global cloud.

In other words, we claim that there are several practical cases of large and growing relevance where the joint exploitation of MEC and MCS would bring highly significant benefits in terms of efficient resource usage and perceived service quality. However, notwithstanding recent advances in both MEC and MCS, to the best of our knowledge, only a very limited number of seminal works has explored the mutual advantages in the joint use of these two classes of solutions, and they are mostly focused on pure technical communication aspects without considering the crucial importance of having humans as central contributors in the loop [2, 3, 4].

The paper reports some research ideas and findings in a brand new area that we call Human-driven Edge Computing (HEC) defined as a new model to ease the provisioning and deployment of MEC platforms as well as to enable more powerful MEC-enabled MCS applications. First and foremost, *HEC eases the planning and deployment of the basic MEC model*: it mitigates the potential weaknesses of having only Fixed MEC entities (FMEC) by exploiting MCS to continuously monitor humans and their mobility patterns, as well as to dynamically re-identify hot locations of potential interest for the deployment of new edges. Second, to overcome FMEC limitations, *HEC enables the implementation and dynamic activation of impromptu and temporary Mobile MEC entities (M^2EC)* that leverage resources of locally available mobile devices. Hence, a M^2EC is a local middleware proxy dynamically activated in a logical bounded location where people tend to stay for a while with repetitive and predictive mobility patterns [5], thus realizing a mobile, opportunistic, and participatory edge node. Third, given that M^2EC, differently from FMEC, does not implement powerful backhaul links toward the core cloud, *HEC exploits local one-hop communications*

and the store-and-forward principle by using humans (moving with their devices) *as VM/container couriers* to enable migrations between well-connected FMEC and local M^{2}EC.

2. Boosting Mobile Edge Computing through Human-driven Edge Computing

We refer to the scenario shown in Fig. 1. It extends the usual three-layer device-MEC-cloud hierarchical architecture (based on the interposition of FMEC entities) with the addition of the new M^{2}EC entity. Indeed, the MCS approach combined with the seamless tracking of volunteers (monitoring both their mobility and their performance in terms of completion rates of assigned sensing tasks) allows to: i) identify the optimal locations where people tend to interact. Such locations ease the effective deployment of FMEC and M^{2}EC. Furthermore, it allows to ii) select of those users willing to host M^{2}EC. Such users act as local access points to the hierarchical HEC.

We experienced with the ParticipAct MCS living lab [6] in order to clarify the effectiveness of architecture proposed. We learned from ParticipAct that some locations aggregate people during all the day (such locations are indeed ideal candidates for the FMEC, see E_1, E_2, and E_3 in Fig. 1). At the same time some locations become *active* only during shorter and different timeslots (e.g., P_1 and P_4 from 9:00AM to 10:30AM, while P_2 is frequented only from 4:00PM to 6:00PM). These latest areas, out of the highly frequented people paths, would highly benefit of being served by a local (in time and space) M^{2}EC, while it would be inefficient and overprovisioned to have additional FMEC there (see Fig. 1).

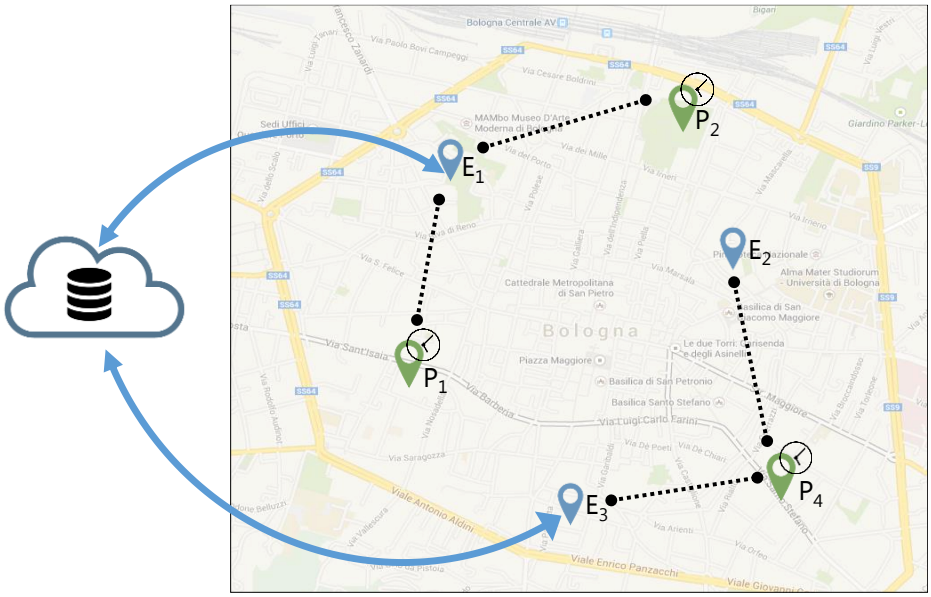


Fig. 1: FMEC, M^{2}EC, and couriers in our HEC model.

Another interesting aspect we learned from the ParticipAct MCS living lab concerns HECs. They exploit opportunistic interactions among devices in order to enable the migration of Virtual Machines (VM)/containers. This feature can be achieved by leveraging human couriers moving from/to different FMECs (see Fig. 1) [7]. In our reference architecture, devices can interact through one-hop *ad-hoc* communications. Such interactions are possible by using short-range network interfaces, such as Bluetooth (i.e., up to 25m), Wi-Fi configured in direct mode (i.e., up to 150m) or the LTE-direct technology (i.e., up to 500m).
Similarly to the paradigm adopted with the MSN, *courier* devices automatically down/upload VM/containers from the FMEC as soon as they are close enough to another device in order to transfer data. In turn, devices can share data gathered from other devices roaming in the same M^{2}EC (see dotted lines in Fig. 1). Refer to Section 3 for the selection criteria of the most suitable human couriers.

Without claiming completeness and due to space limitations, in the following we briefly overview the current state-of-the-art in the main related fields. Focusing on architectural aspects of HEC, the MEC/fog literature has already produced some relevant modeling work and some seminal design/implementation results. Narrowing to efforts close to ours, as reported in [1], some first exploratory research activities have considered cooperation issues between edges

and the core, but only a very few works concentrated on the opportunities of having cooperation between devices and the edges. Considering MCS as application scenario, [2] and [3] propose to enhance the MCS process by leveraging intermediate MEC nodes, namely, FMECs, to boost data upload from mobile nodes to the infrastructure [2] and to provide more computing/storage capabilities closer to end mobile devices [3]. A very recent and interesting work, the closest to our HEC concept for what relates to enabling more collaboration between entities co-located at edges is [4]: it proposes not only to have the traditional "vertical" collaboration between devices, MEC, and cloud level, but also an "horizontal" collaboration between entities at the same level via ad-hoc communications; however, it neglects humans and social/mobility effects, namely, there is no idea to dynamically identify and impromptu form M^{2}ECs as in our novel HEC proposal.

Finally, concerning the system and the implementation aspects, only a very few research activities are focused on the migration of VM/container on MEC middleware for mobile services over hostile environments. It is worth to notice that such activities are relevant aspects of modern CPS. Authors of [8] highlight the limitations of traditional live VM migration based on edge devices. They propose a live migration approach in response to client handoff in cloudlets, with less involvement of the hypervisor and, at the same time, by promoting migration to optimal offload sites. Authors also discuss how to adapt the system to the changing network conditions and processing capacity. The work described in [9] presents the *foglets* programming infrastructure. Such infrastructure handles some mechanisms for quality/workload-sensitive migration of service components among fog nodes. Another interesting work is reported in [10]. It proposes the usage of *cloudlets* to support mobile multimedia services and to adjust the resource allocation triggered by runtime handoffs. Concerning the handoff evaluation, the authors of [11] study the handoff conditions in relation to various aspects such as signal strength, bit rate, number of interactions between cloudlets and associated devices. Finally, [12] proposes a multi-agent-based code offloading mechanism. It adopts a reinforcement learning and code blocks migration in order to reduce both execution time and energy consumption of mobile devices. To the best of our knowledge, these papers explore the integrated management of handover operations with VM/container migration. However, none of them considers the possibility of exploiting peoples' devices as storage/VM/container couriers.

3. Mobile Edge Computing extended through the Crowd

We first overview the HEC architecture as well as its main components and functionalities. Then, we present some guidelines and engineering tradeoffs for the selection of FMEC, M^{2}EC, and of the human couriers.

3.1. The Reference Architecture of the HEC Middleware

HEC extends the emerging MEC three-layer hierarchical architecture. In particular, we consider two types of MECs, namely FMEC and M^{2}EC. By focusing on our HEC middleware at mobile devices, we distinguish between regular mobile devices (capable of working only as service clients) and powerful devices (which may be promoted dynamically to host virtualized functions and to serve as M^{2}EC nodes). In our current implementation, we identify a number of powerful devices based on the hardware and software features (ie. tablets or laptops that are locally paired with smartphones). It is worth to notice that the evolution trend of mobile/embedded devices is such that the potential set of mobile nodes that can be promoted to M^{2}EC at runtime is ever increasing. Under this respect, some interesting benchmarks show that also RaspberryPI boards can adequately run OpenStack++ middleware [13]. We consider that our HEC middleware is already installed on such nodes before starting the provisioning of services, even if more sophisticated dynamic mechanisms for HEC middleware download at runtime can be easily integrated. Our HEC middleware implementation fits a wide spectrum of heterogeneous mobile devices, with the only constraint to run Android (iOS version currently under development).

For what concerns the MCS applications, we consider that only highly demanding or group-oriented locality-based MCS tasks are delegated to FMEC and M^{2}EC nodes, possibly based on dynamic considerations (e.g., residual battery energy). At this stage, the MCS tasks that have been already implemented and experimented for execution at HEC nodes are i) video analysis for face recognition and ii) analytics on all or fused monitoring indicators over geographical areas of highest interest and density such as data fusion, history-based processing of temporal series.

3.2. The selection of FMEC and M^{2}EC and Human-enabled VM/Container Migration

Our architecture is configured with a number of FMEC and M^{2}EC. They are selected by analyzing the human mobility over an observation period. Concerning the FMECs, we consider those locations remaining mostly *active* during the whole day. These are locations not subject of mobility changes. To this purpose, we use the DBSCAN

algorithm in order to detect clusters of users roaming around the same location [6, 14]. DBSCAN returns K distinct clusters, we filter out some of them, in particular we restrict to $k \leq K$ clusters as FMECs.

The M^2EC selection is achieved by spotting those locations of our region becoming *active* only during specific time slots. In fact, our goal is dynamically (re)configure our cloud architecture according to the natural rhythm of a city. To this purpose, we analyze the human mobility only during some temporal slots. In particular, we select those slots characterizing the typical phases of a routinary working day. For each of the slots, we cluster together the positions of the users with a process similarly to the one described for the FMEC. Also for the M^2EC selection, we adopted the DBSCAN algorithm. It results with H clusters of which we keep the top $h \leq H$.

Our process allows setting the parameters *k* and *h* according to the mobility and sociality features of the mobility dataset considered. Specifically, a small crowded region can be provisioned with a low number of FMEC, but with a high number of M^2EC since crowded area change quickly along the day. Conversely, in a wide depopulated area it could be possible to increase the number of FMEC and, at the same time, reducing the number of M^2EC, since mobility changes slowly with the time.

Once FMEC to M^2EC are selected, we then consider how to move that among them. To this purpose, we consider humans (i.e., *couriers*), and their mobile devices provisioned with our HEC middleware, as the primary actors that can be involved into the loop. We assume that mobile devices are equipped with different kinds of network interfaces (short, medium and broadband) and of storage capacity. The storage allows devices to store-carry-and-forward data among FMEC and M^2EC, as well as it allows replicating data across users joining at the same time the same M^2EC For the selection of couriers, we keep track of user mobility and prefers those users that have a more repetitive and predictable behavior: the more a user commutes from a FMES to a M^2EC, the more he/she is a good courier candidate.

Since not all the FMECs are connected to all M^2ECs during the 24 hours, we consider the possibility of reducing the bandwidth in the cloud-to-FMEC direction and consequently the storage resources at FMECs. To this purpose, the HEC implements a *load balancing* policy. Such policy exploits the knowledge of the mobility and of the connectivity between FMECs and M^2EC in order to select which VMs/containers requires to be moved move from the cloud to the FMECs. The load balancing strategy relies on the locality principle according to which VMs/containers are loaded in advance to those FMECs that are more likely to be store-and-forwarded by a courier toward a M^2EC.

Also for the sake of briefness and due to paper length limitations, further design/implementation details about our HEC proposal are not reported here because out of the central scope of this paper, which presented the vision and the main design guidelines of our innovative HEC solution. At the current stage, we are working in order to test these ideas through a set of experiments based on the real-world ParticipAct dataset which reproduces the mobility of about 170 students in the Emilia Romagna region (Italy) about 2 years [6].

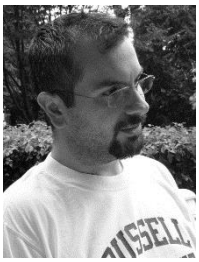
4. Conclusion

This paper presented HEC, a new architecture model to ease the provisioning and to extend the coverage of traditional MEC approaches by bringing together the best of MEC and MCS. The cornerstone of our proposal lies in the ability to dynamically leverage human sociality and mobility effects to broaden the MEC coverage through the impromptu formation of M^2ECs. Those encouraging results are pushing us to further investigate and refine our HEC model and we are currently exploring various related areas. On the one hand, we are working to enable the self-adaptable fine tuning of our HEC middleware to the different dynamics and variations of the city pulse, for instance to the different behaviors that might present along the year, such as working vs. vacation periods, and the week, such as working days vs. weekends. On the other hand, we are investigating innovative techniques in order to reduce the latency of downloading VMs/containers on M^2EC nodes via parallelization of I/O and configuration operations.

References

[1] S. Wang et al., "A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications", IEEE Access, vol. PP, no. 99, pp.1-1
doi:10.1109/ACCESS.2017.2685434.

[2] S. K. Datta, R. P. Ferreira da Costa, C. Bonnet and J. Härri, "oneM2M architecture based IoT framework for mobile crowd sensing in smart cities", in Proceedings of 2016 European Conference on Networks and Communications (EuCNC), 2016, pp. 168-173.
[3] K. M. S. Huq, S. Mumtaz, J. Rodriguez, P. Marques, B. Okyere and V. Frascolla, "Enhanced C-RAN Using D2D Network", IEEE Communications Magazine, vol. 55, no. 3, pp. 100-107, March 2017.
[4] T. X. Tran, A. Hajisami, P. Pandey and D. Pompili, "Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges", IEEE Communications Magazine, vol. 55, no. 4, pp. 54-61, April 2017.
[5] Pappalardo, L.; Simini, F.; Rinzivillo, S.; Pedreschi, D.; Giannotti, F.; Barabási, A.L., "Returners and Explorers dichotomy in Human Mobility", Nature Communications, vol. 6, Article 8166, 2015.
[6] G. Cardone, A. Cirri, A. Corradi, L. Foschini, "The ParticipAct Mobile Crowd Sensing Living Lab: The Testbed for Smart Cities", IEEE Commununications Magazine, vol. 52 , no. 10, pp. 78–85, October 2014.
[7] M. Girolami, S. Chessa, A. Caruso, "On Service Discovery in Mobile Social Networks: Survey and Perspectives", Computer Networks, vol. 88, pp. 51-71, 2015.
[8] K. Ha et al., "Adaptive VM Handoff Across Cloudlets", Technical Report CMU-CS-15-113, CMU School of Computer Science, 2015.
[9] E. Saurez et al., "Incremental Deployment and Migration of Geo-Distributed Situation Awareness Applications in the Fog", in Proceedings of ACM Distributed and Event-based Systems Int. Conf., pp. 258-269, 2016.
[10] M. Felemban, S. Basalamah, A. Ghafoor, "A Distributed Cloud Architecture for Mobile Multimedia Services", 2013.
[11] A. Ravi, S.K. Peddoju, "Handoff Strategy for Improving Energy Efficiency and Cloud Service Availability for Mobile Devices", Wireless Personal Communications, vol. 81, no. 1, pp. 101–132, 2015.
[12] M.G.R. Alam, et al., "Multi-agent and Reinforcement Learning Based Code Offloading in Mobile Fog", in Proceedings of International Conference on Information Networking (ICOIN), 2016.
[13] P. Bellavista, M. Solimando, A. Zanni, "A Migration-enhanced Edge Computing Support for Mobile Services in Hostile Environments - Lessons Learnt from Platform Implementation and Deployment", to appear in Proceedings of International Wireless Communications and Mobile Computing Conference, pp. 1-6, 2017.
[14] M. Ester, H. peter Kriegel, J.S, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", in Proceedings of International Conference on Knowledge Discovery, AAAI Press, pp. 226–231, 1996.



Luca Foschini graduated from the University of Bologna, Italy, where he received a Ph.D. degree in computer engineering in 2007. He is now an assistant professor of computer engineering at the University of Bologna. His interests include distributed systems and solutions for system and service management, management of cloud computing, context data distribution platforms for smart city scenarios, context-aware session control and adaptive mobile services, and mobile crowd sensing and crowdsourcing.

Contact him at luca.foschini@unibo.it.

Michele Girolami graduated from the University of Pisa, where he received a Ph.D. degree in computer science in 2015. Currently, he is a member of research staff at CNR-ISTI with the Wireless Network Laboratory. His research interests include data forwarding and service discovery in mobile social networks, crowd sensing techniques, and context-aware middleware for smart environments.

Contact him at michele.girolami@isti.cnr.it.

Mobile Edge Computing: Recent Efforts and Five Key Research Directions

Tuyen X. Tran, Mohammad-Parsa Hosseini, and Dario Pompili
Department of Electrical and Computer Engineering
Rutgers University–New Brunswick, NJ, USA
{tuyen.tran, parsa, pompili}@cac.rutgers.edu

1. Introduction

In the past decade, we have witnessed Cloud Computing play as significant role for massive data storage, control, and computation offloading. However, the rapid proliferation of mobile applications and the Internet of Things (IoT) over the last few years has posed severe demands on cloud infrastructure and wireless access networks. Stringent requirements such as ultra-low latency, user experience continuity, and high reliability are driving the need for highly localized intelligence in close proximity to the end users. In light of this, Mobile Edge Computing (MEC) has been envisioned as the key technology to assist wireless networks with cloud computing-like capabilities in order to provide low-latency and context-aware services directly from the network edge.

Differently from traditional cloud computing systems where remote public clouds are utilized, the MEC paradigm is realized via the deployment of commodity servers, referred to as the MEC servers, at the edge of the wireless access network. Depending on different functional splitting and density of the Base Stations (BSs), a MEC server can be deployed per BS or at an aggregation point serving several BSs. With the strategic deployment of these computing servers, MEC allows for data transfer and application execution in close proximity to the end users, substantially reducing end-to-end (e2e) delay and releasing the burden on backhaul network [1]. Additionally, MEC has the potential to empower the network with various benefits, including: (i) optimization of mobile resources by hosting compute-intensive applications at the network edge, (ii) pre-processing of large data before sending it (or some extracted features) to the cloud, and (iii) context-aware services with the help of Radio Access Network (RAN) information such as cell load, user locations, and radio resource allocation.

In this letter, as a backdrop to identifying research questions, we briefly review recent research efforts on enabling MEC technologies and then discuss five key research directions. Specifically, the goals of this letter are: (i) to raise awareness of relevant and cutting-edge work being performed from various literature, and (ii) to identify a number of important research needs for future MEC systems.

2. Recent Efforts in Enabling MEC Technologies

Fueled by the promising capabilities and business opportunities, the MEC paradigm has been attracting considerable attention from both academia and industry. A number of deployment scenarios, service use cases, and related algorithms design has been proposed to exploit the potential benefits of MEC and to justify its implementation and deployment from both a technical and business point of view. In this section, we briefly review the recent efforts from both standardization and research perspectives towards enabling MEC technologies in wireless networks.

2.1 Proofs of Concepts and Standardization Efforts

In 2013, Nokia Networks introduced the very first real-world MEC platform [2], in which the computing platform–Radio Applications Cloud Servers (RACS)–is fully integrated with the Flexi Multiradio BS. Saguna also introduced their fully virtualized MEC platform, so called Open-RAN [3], that can provide an open environment for running third-party MEC applications. Besides these solutions, MEC standardization is being specified by the European Telecommunications Standards Institute (ETSI), which recently formed a MEC Industry Specifications Group (ISG) to standardize and moderate the adoption of MEC within the RAN. In the introductory white paper [4], four typical service scenarios and their relationship to MEC have been discussed, ranging from Augmented Reality (AR) and intelligent video acceleration to connected cars and IoT gateway. In the MEC World Congress 2016, ETSI has announced six Proofs of Concept (PoCs) that were accepted by the MEC ISG, including:

- Radio Aware Video Optimization in a Fully Virtualized Network (RAVEN);
- Flexible IP-based Services (FLIPS);
- Enterprise Services;
- Healthcare–Dynamic Hospital User, IoT, and Alert Status Management;
- Multi-Service MEC Platform for Advanced Service Delivery;

- Video Analytics.

These PoCs strengthen the strategic planning and decision making of organizations, helping them identify which MEC solutions may be viable in the network. Also in this Congress, ETSI MEC ISG has renamed Mobile Edge Computing as Multi-access Edge Computing in order to reflect the growing interest in MEC from non-cellular operators, which takes effect starting from 2017 [5]. The technical requirements for MEC are specified in [6] to guarantee interoperability and to promote MEC deployment. These requirements are divided into generic requirements, service requirements, requirements on operation and management, and finally security, regulations and charging requirements. Most recently, the 3GPP has shown a growing interest in incorporating MEC into its 5G standard and has identified functionality supports for edge computing in a recent technical specification contribution [7].

2.2 MEC Architecture and Virtualization

In recent years, the concept of integrating cloud computing-capabilities into the wireless network edge has been considered in the literature under different terminologies, including Small Cell Cloud (SCC), Mobile Micro Cloud (MMC), Follow Me Cloud (FMC), and CONCERT [8]. The basic idea of SCC is to enhance the small cells, such as microcells, picocells or femtocells, with additional computation and storage capabilities so as to support edge computing [9]. By exploiting the Network Function Virtualization (NFV) paradigm, the cloud-enabled small cells can pool their computation power to provide users with services/applications having stringent latency requirements. Similarly, the concept of MMC introduced in [10] allows users to have instantaneous access to the cloud services with low latency. Differently from the SCC where the computation/storage resources are provided by interworking clusters of enhanced small cells, the User Equipment (UE) exploits the computation resources of a single MMC, which is typically connected directly to a BS. The FMC concept [11] proposes to move computing resources a bit further from the UEs, compared to SCC and MMC, to the core network. It aims at having the cloud services running at distributed data centers so as to be able to follow the UEs as they roam throughout the network. In all these described MEC concepts, the computing/storage resources have been fully distributed; conversely, the CONCERT concept proposes hierarchically placement of the resources within the network in order to flexibly and elastically manage the network and cloud services.

2.3 Computation Offloading

The benefits of computation offloading have been investigated widely in conventional Mobile Cloud Computing (MCC) systems. However, a large body of existing works on MCC assumed an infinite amount of computing resources available in a cloudlet, where offloaded tasks can be executed in negligible delay [12], [13]. Recently, several works have focused on exploiting the benefits of computation offloading in MEC network [14]. The problem of offloading scheduling was then reduced to radio resource allocation in [15], where the competition for radio resources is modeled as a congestion game of selfish mobile users. The problem of joint task offloading and resource allocation was studied in a single-user system with energy harvesting devices [16], and in a multi-cell, multi-user systems [17]; however, the congestion of computing resources at the MEC server was not taken into account. A similar problem is studied in [18] for single-server MEC systems, where the limited resources at the MEC server were factored in, and later on extended to multi-server MEC systems in [19].

2.4 Edge Caching

The increasing demand for massive multimedia services over mobile cellular network poses great challenges on network capacity and backhaul links. Distributed edge caching, which can well leverage MEC paradigm, has therefore been recognized as a promising solution to bring popular contents closer to the users, to reduce data traffic going through the backhaul links as well as the time required for content delivery, and to help smoothen/regulate the traffic during peak hours. In general, edge caching in wireless networks has been investigated in a number of works (cf. [20-22] and references therein). Recently, in [23], [24], we have proposed a cooperative hierarchical caching paradigm in a Cloud Radio Access Network (C-RAN) where the cloud-cache is introduced as a bridging layer between the edge-based and core-based caching schemes. Taking into account the heterogeneity of video transmissions in wireless networks in terms of video quality and device capabilities, our previous work in [25] proposes to utilize both caching and processing capabilities at the MEC servers to satisfy users' requests for videos with different bitrates. In this scheme, the collaborative caching paradigm has been extended to a new dimension where the MEC servers can assist each other to not only provide the requested video via backhaul links but also to transcode it to the desired bitrate version.

3. Five Key Research Directions for MEC in Wireless Networks

Research on MEC lies at the intersection of wireless communications and cloud computing, which has resulted in many interesting research opportunities and challenges. The spectrum of research required to achieve the promises of MEC requires significant investigation along many directions. In this section, we highlight and discuss the key open research issues and future directions, which are categorized into five main topics as follows.

3.1 Deployment Scenarios and Resource Management

The key concept of MEC is to shift the cloud computing-capabilities closer to the end users in order to reduce the service latency and to avoid congestion in the core network. However, there has been no formal definition on what the MEC servers would be and where they should be deployed within the network. Such decisions involve investigating the site-selection problem for MEC servers where their optimal placement is coupled with the computational resource provisioning as well as with the deployment budget. In addition, it is critical to determine the required server density to cope with the service demands, which is closely related to the infrastructure deployment cost and marketing strategies. Finally, the deployment of MEC servers also depends on the RAN architecture where different functional splitting options between the BSs and the centralized processing center (such as in C-RAN) are specified, depending on the delay requirement and fronthaul capacity.

3.2 Computation Caching and Offloading

The combination of computation and storage resources at the MEC servers offers unique opportunities for caching of computation tasks. In this technique, the MEC server can cache several application services and their related database, and handle the offloaded computation from multiple users so as to enhance the user experience. Computation caching can help decrease the load on the access link by providing computing results to the end users without the need to fetch their tasks beforehand. Unlike content caching, computation caching presents several new challenges. First, computing tasks can be of diverse types and depend on the computing environment; while some of the content is cacheable for reuse by other devices, personal computing data is not cacheable and must often be executed in real time. Second, it is not practical to build popularity patterns locally at each server; instead, studying popularity distributions over larger sets of servers can provide a broader view on the popularity patterns of computing tasks.

3.3 IoT Applications and Big Data Analytics

The emerging IoT and Big Data services have changed the traditional networking paradigm where the network infrastructure, instead of being the dump pipe, can now process the data and generate insights. MEC resources can be utilized for pre-processing of massive IoT data so as to reduce bandwidth consumption, to provide network scalability, and to ensure a fast response to the user requests. A MEC platform can also encompass a local IoT gateway functionality capable of performing data aggregation and big data analytics for event reporting, smart grid, e-health, and smart cities. For instance, our previous work in [26] describes an autonomic edge-computing platform that supports deep learning for localization of epileptogenicity using multimodal rs-fMRI and EEG big data. To fully exploit the benefits of MEC for IoT, there needs to be significant research on how to efficiently distribute and manage data storage and computing, how to make edge computing collaborate with cloud computing for more scalable services, and how to secure the whole system.

3.4 Mobility Management

Mobility management is an essential feature for MEC to ensure service continuity for highly dynamic mobile users. For vehicular communications and automotive, integrating MEC with mobile cloud computing or vehicular cloud, wherein mobile or vehicle resources are utilized for communication and computation services, is a highly challenging issue from the service orchestration perspective. For many applications, estimating and predicting the movement and trajectory of users as well as personal preference information can help the MEC servers improve the user experience. For example, mobility prediction can be integrated with edge caching to enhance the content migration at the edges and caching efficiency. In addition, to achieve better user computation experience, existing offloading techniques can be jointly considered with mobility-aware scheduling policies at the MEC servers. This approach introduces a set of interesting research problems including mobility-aware online prefetching of user computation data, server scheduling, and fault-tolerance computation. For instance, in our previous works [27], [28], multi-tier distributed computing infrastructures based on MEC and Mobile Device Cloud (MDC) are proposed to link mobility management and pervasive computing with medical applications.

3.5 Security and Privacy

Security issues might hinder the success of the MEC paradigm if not carefully considered. Unlike traditional cloud computing, MEC infrastructure is vulnerable to site attacks due to its distributed deployment. In addition, MEC requires more stringent security policies as third-party stakeholders can gain access to the platform and derive information regarding user proximity and network analytics. Existing centralized authentication protocols might not be applicable for some parts of the infrastructure that have limited connectivity to the central authentication server. It is also important to implement trust-management systems that are able to exchange compatible trust information with each other, even if they belong to different trust domains. Furthermore, as service providers want to acquire user information to tailor their services, there is a great challenge to the development of privacy-protection mechanisms that can efficiently protect users' locations and service usage.

4. Conclusion

Mobile Edge Computing (MEC) is an emerging technology to cope with the unprecedented growth of user demands for access to low-latency computation and content data. This paradigm, which aims at bringing the computing and storage resources to the edge of mobile network, allows for the execution of delay-sensitive and context-aware applications in close proximity to the end users while alleviating backhaul utilization and computation at the core network. While research on MEC has gained its momentum, as reflected in the recent efforts reviewed in this letter, MEC itself is still in its nascent stage and there is a myriad of technical challenges that need to be addressed. In this regard, we discussed five key open research directions that we consider to be among the most important and challenging issues of future MEC systems.

References

[1] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges," IEEE Communications Magazine, vol. 55, no. 4, pp. 54-61, 2017.

[2] Intel and Nokia Siemens Networks, "Increasing mobile operators' value proposition with edge computing," Technical Brief, 2013.

[3] Saguna and Intel, "Using mobile edge computing to improve mobile network performance and profitability," White paper, 2016.

[4] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile Edge Computing — A Key Technology Towards 5G," ETSI white paper, vol. 11, 2015.

[5] N. Sprecher, J. Friis, R. Dolby, and J. Reister, "Edge computing prepares for a multi-access future," Sep. 2016. [Online]. Available: http://www.telecomtv.com/articles/mec/edge-computing-prepares-for-a-multi-access-future-13986/

[6] ETSI GS MEC 002: Mobile Edge Computing (MEC); Technical Requirements V1.1.1, March 2016.

[7] 3GPP, "Technical specification group services and system aspects; system architecture for the 5g systems; stage 2 (release 15)," 3GPP TS 23.501 V0.4.0, Apr. 2017.

[8] J. Liu, T. Zhao, S. Zhou, Y. Cheng, and Z. Niu, "CONCERT: a cloud-based architecture for next-generation cellular systems", IEEE Wireless Communications, vol. 21, no. 6, pp. 14-22, Dec. 2014.

[9] FP7 European Project, "Distributed computing, storage and radio resource allocation over cooperative femtocells (TROPIC)," [Online]. Available: http://www.ict-tropic.eu/, 2012.

[10] S. Wang, et al., "Mobile Micro-Cloud: Application Classification, Mapping, and Deployment", Annual Fall Meeting of ITA (AMITA), 2013.

[11] T. Taleb, A. Ksentini, and P. A. Frangoudis, "Follow-Me Cloud: When Cloud Services Follow Mobile Users," IEEE Transactions on Cloud Computing, vol PP, no. 99, pp. 1-1, May 2017.

[12] W. Zhang, Y. Wen, and D. O. Wu, "Collaborative task execution in mobile cloud computing under a stochastic wireless channel," IEEE Trans. Wireless Commun., vol. 14, no. 1, pp. 81–93, 2015.

[13] Z. Cheng, P. Li, J. Wang, and S. Guo, "Just-in-time code offloading for wearable computing," IEEE Trans. Emerg. Topics Comput., vol. 3, no. 1, pp. 74–83, 2015.

[14] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading." IEEE Communications Surveys & Tutorials, Mar. 2017.

[15] X. Chen, "Decentralized computation offloading game for mobile cloud computing," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 4, pp. 974–983, 2015.

[16] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," IEEE J. Sel. Areas in Commun., vol. 34, no. 12, pp. 3590–3605, 2016.

[17] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-

edge computing," IEEE Trans. Signal Inf. Process. Over Netw., vol. 1, no. 2, pp. 89–103, 2015.
[18] X. Lyu, H. Tian, P. Zhang, and C. Sengul, "Multi-user joint task offloading and resources optimization in proximate clouds," IEEE Trans. Veh. Technol., vol. 66, no. 4, pp. 3435-3447, April 2017.
[19] T. X. Tran and D. Pompili. "Joint Task Offloading and Resource Allocation for Multi-Server Mobile-Edge Computing Networks." arXiv preprint arXiv:1705.00704, 2017.
[20] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," IEEE Communications Magazine, vol. 52, no. 8, pp. 82-89, 2014.
[21] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in Proc. IEEE INFOCOM, pp. 1107-1115, 2012.
[22] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," IEEE/ACM Transactions on Networking, vol. 22, no. 5, pp. 1444-1462, 2014.
[23] T. X. Tran and D. Pompili, "Octopus: A Cooperative Hierarchical Caching Strategy for Cloud Radio Access Networks," in Proc. IEEE Int. Conf. on Mobile Ad hoc and Sensor Systems (MASS), pp. 154-162, Oct. 2016.
[24] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative Hierarchical Caching in 5G Cloud Radio Access Networks (C-RANs)," IEEE Network, July 2017.
[25] T. X. Tran, P. Pandey, A. Hajisami, and D. Pompili, "Collaborative Multi-Bitrate Video Caching and Processing in Mobile-Edge Computing Networks," in Proc. IEEE/IFIP Conference on Wireless On-demand Network Systems and Services (WONS), pp.165-172, 2017.
[26] M. P. Hosseini, T. X. Tran, D. Pompili, K. Elisevich, and H. Soltanian-Zadeh, "Deep Learning with Edge Computing for Localization of Epileptogenicity using Multimodal rs-fMRI and EEG Big Data," in Proc. IEEE Int'l Conf. Autonomic Computing (ICAC), July 2017, to appear.
[27] M. P. Hosseini, A. Hajisami, and D. Pompili, "Real-time Epileptic Seizure Detection from EEG Signals via Random Subspace Ensemble Learning," in Proc. IEEE Int'l Conf. Autonomic Computing (ICAC), Wurzburg, Germany, Jul. 2016.
[28] M. P. Hosseini, H. Soltanian-Zadeh, K. Elisevich, and D. Pompili, "Cloud-based Deep Learning of Big EEG Data for Epileptic Seizure Prediction," IEEE Global Conference on Signal and Information Processing (GlobalSIP), Washington, D.C., Dec. 2016.

Tuyen X. Tran is working towards his PhD degree in Electrical & Computer Engineering (ECE) at Rutgers University, NJ. He is pursuing research in the fields of wireless communications and mobile cloud computing, with emphasis on Cloud Radio Access Networks and Mobile-Edge Computing. He received the MSc degree in ECE from the University of Akron, USA, in 2013, and the BEng degree (Honors Program) in Electronics and Telecommunications from Hanoi University of Technology, Vietnam, in 2011. He was a research intern at Huawei Technologies R&D Center, Bridgewater, NJ, during the summers of 2015 and 2016. He was the recipient of the Best Paper Award at the IEEE/IFIP Wireless On-demand Network systems and Services Conference (WONS) 2017.

Mohammad-Parsa Hosseini is a Senior Member of IEEE. He is a PhD candidate and a research assistant in the Dept. of ECE at Rutgers University, NJ, USA and a member of the CPS Lab under the guidance of Prof. Pompili. He is collaborating with Medical Image Analysis Lab at Henry Ford Health System, MI, under the guidance of Prof. Soltanian-Zadeh specifically in neuroimaging and data science. He is collaborating with the Clinical Neurosciences Dept. Spectrum Health, MI, under the guidance of Prof. Elisevich in computational neuroscience. His research focuses on deep/machine learning, signal/image processing, big data, and cloud computing with future applications in the field of health care. He was a research intern at Apple Inc., Silicon Valley, CA, during the summer of 2017. Previously, he was a PhD student in the ECE Dept. of Wayne State University, MI, in 2013 and he has been teaching as an adjunct professor at several universities since 2009.

Dr. Dario Pompili is an Assoc. Prof. with the Dept. of ECE at Rutgers U. He is the director of the Cyber-Physical Systems Laboratory (CPS Lab), which focuses on mobile computing, wireless communications and networking, acoustic communications, sensor networks, and datacenter management. He received his PhD in ECE from the Georgia Institute of Technology in June 2007. He had previously received his 'Laurea' (combined BS and MS) and Doctorate degrees in Telecommunications and System Engineering from the U. of Rome "La Sapienza," Italy, in 2001 and 2004, respectively. He is a recipient of the NSF CAREER'11, ONR Young Investigator

Program'12, and DARPA Young Faculty'12 awards. In 2015 he was nominated Rutgers Chancellor's Scholar. He published more than a hundred refereed scholar publications: with about 7,000 citations, Dr. Pompili has an h-index of 29 and a i10-index of 53 (Google Scholar, May'17). Since 2014 he is a Senior Member of both the IEEE Communications Society and the ACM.

SPECIAL ISSUE ON Advances in Light Field Image Processing & Applications

Guest Editors: Erhan Ekmekcioglu[1] *and Pedro A. A. Assunção* [2, 3]

[1]*Loughborough University London, United Kingdom*
[2] *Instituto de Telecomunicações, Portugal*
[3] *Instituto Politécnico de Leiria, Leiria, Portugal*

E.Ekmekcioglu@lboro.ac.uk; amado@co.it.pt

Being a novel imaging technique, Light Field (LF) imaging not only conveys the intensity of the light directed from a scene, but also its directionality. Thus, two layers of information are presented in LF images, one being the angularity of approaching light rays and the other being the usual colour and luminance information. As a result, LF images enable several possibilities, such as image re-focussing, viewpoint adjustment, and scene depth extraction, which would otherwise not be possible with conventional camera shots. LF image acquisition deploys a different optics setup compared to conventional imaging systems, consisting of an array of multiple lenses and image sensors (i.e., multi-camera arrays) or special micro-lens arrays placed in compact camera systems. Owing to its complexity and large data volume, LF image processing and communication has attracted attention from research communities in computer vision and signal processing. With the more common availability of LF displays and off-the-shelf LF camera products, the research efforts in that area have further gone up recently. The LF technology has been in use for emerging applications in various domains, such as computational photography, augmented reality, light-field microscopy, and 3D display of objects and visual scenes in the whole field-of-view.

Light Field communications cover an end-to-end cycle, i.e., starting from acquisition to post-processing and compression, then transmission and finally display rendering. While the acquisition and displaying of the LF images require specific hardware, mostly the modified versions of existing techniques, which are originally developed for conventional images, have been used for the processing and compression of LF images. Nevertheless, LF images/videos differ fundamentally from their traditional counterparts, thus making those techniques sub-optimal in most cases. In this Special Issue, authors highlight their research findings and perspectives on the different aspects of the LF imaging technology.

The first contribution by Christine Guillemot and Reuben Farrugia, titled "*Light field image processing: overview and research issues*", briefs the readers about the fundamentals of the plenoptic function, the varieties and the design trade-offs of the LF capturing systems. Furthermore, the authors provide us with a survey of the past and ongoing research works addressing various aspects, such as the compression of vast LF data, the inherent spatial and angular resolution trade-off, and user interactivity issues. They outline the major research questions in the field open to further investigation.

The second contribution by Joao M. Santos, et. al., titled "*Performance evaluation of light field pre-processing methods for lossless standard coding*", provides a detailed performance comparison of various raw data pre-processing and coding standards in the context of LF images. In an exploratory study, the authors test the lossless compression gain of deploying different data arrangement models and colour transforms with the well-known standard compression standards, such as JPEG2000, JPEG-LS, and HEVC.

In their paper titled "*Towards Adaptive Light Field Video Streaming*", Peter A. Kara, et. al., give an overview of the issue of Quality of Experience measurement in the context of LF visualisation. Based on the authors' research works in the evaluation of the LF video quality on a holographic display in relation with the field-of-view, spatial and angular resolutions, a new adaptive LF video streaming approach is proposed.

Finally, in his paper titled "*Light Fields for Near-eye Displays*" Fu-Chung Huang sheds light on the design aspects of special near-eye displays for LF visualisation. Near-eye displays are of particular importance given the surging popularity of Virtual Reality head mounted displays. Near-eye LF displays have several advantages over traditional head-mounted displays, such as providing superior depth of field, more comfortable and natural viewing free from vergence - accommodation conflict. The author outlines several different near-eye light display technologies with a comparison in terms of multiple aspects including resolution, field-of-view and the form factor.

With this Special Issue we have no intent to present a complete picture on the state of the LF technology and the applications it powers. However, we hope that the presented papers provide the audience with a brief tutorial and valuable insight into the persisting challenges in the area, and predictions for the future research.

Our special thanks go to all authors for their precious contributions to this Special Issue. We would also like to acknowledge the gracious support from the Board of MMTC Communications - Frontiers.

Erhan Ekmekcioglu received his Ph.D. degree from University of Surrey, UK, in 2010, where he worked as a post-doctoral researcher in multimedia communication systems until 2014. Since October 2014 he is with the Institute for Digital Technologies at Loughborough University London as a senior researcher. His research interests include 2D/3D and multi-view video processing, coding, and transport, quality of experience, immersive and interactive multimedia. He is the co-author of around 50 peer-reviewed research articles, book chapters, and a book on 3D-TV systems.

Pedro A. A. Assunção received the Licenciado and M.Sc. degrees from the University of Coimbra, in 1988 and 1993, respectively, and the Ph.D. in Electronic Systems Engineering from the University of Essex, in 1998. He is currently professor of Electrical Engineering and Multimedia Communication Systems at the Polytechnic Institute of Leiria and a senior researcher at the Institute for Telecommunications, Portugal. His current research interests include high efficiency and 360-degree, multi-view video and light field coding, multiple description and robust coding, error concealment and quality evaluation. He is a senior member of the IEEE.

Light field image processing: overview and research issues

Christine Guillemot[1], *Reuben Farrugia*[2],
[1]*INRIA, Rennes, FRANCE*
[2]*University of Malta, MALTA*
Christine.Guillemot@inria.fr; reuben.farrugia@um.edu.mt

1. Introduction

Light field (LF) imaging first appeared in the computer graphics community with the goal of photorealistic 3D rendering [1]. Motivated by a variety of potential applications in various domains (e.g., computational photography, augmented reality, light field microscopy, medical imaging, 3D robotic, particle image velocimetry), imaging from real light fields has recently gained in popularity, both at the research and industrial level.

Research effort has been dedicated to the practical design of systems for capturing real-world light fields which go from cameras arrays [2-5] to single cameras mounted on moving gantries and plenoptic cameras [6,7] based on the principle of integral imaging first introduced by Lipman in [8]. The commercial availability of plenoptic cameras and the equipment of recent smart phones with several cameras, with a single specialized sensor, or with a wafer-level-optics camera array [9], which can, to some extent, capture light fields, even if they are not as angularly dense as those captured by plenoptic cameras, has given a novel momentum to light field research. The flow of rays captured by light field acquisition devices is in the form of large volumes of data retaining both spatial and angular information of a scene, which enables a variety of post-capture processing capabilities, such as re-focusing, extended focus, different viewpoint rendering and depth estimation, from a single exposure.

While offering unprecedented opportunities for advanced image analysis, creation and editing features, real light fields capture poses a number of challenging problems. The data captured by light fields cameras is not only big in volume and of high dimension, which is an issue for storage and communication, but also overwhelming in several other aspects such as the need for high processing power and the angular-spatial resolution trade-off inherent to light field capture devices. The volume of data inherent to light fields is an issue for user interaction which requires near real-time processing, potentially on devices having limited computational power. Editing with a tractable complexity and in a consistent manner the large number of views cannot be solved with a straightforward application of now well-known 2D images editing algorithms. After a brief recall of the plenoptic function and of light fields capturing devices, this paper gives an overview of the main research directions addressing the above challenging problems.

2. Plenoptic function and real light fields capturing devices

Light field capturing is about sampling the plenoptic function which is a 7D function $L(x,y,z,\phi,\theta,\Lambda,t)$ describing the light rays emitted by a scene and received by an observer at a particular point (x,y,z) in space, following an orientation defined by the angles (ϕ,θ), with a wavelength Λ, at a given time instant t. For a static light field, the 7D plenoptic function can be simplified into a 4D representation called 4D light field in [10] and Lumigraph in [11], describing the radiance along rays by a function L(x; y; u; v) of 4 parameters at the intersection of the light rays with 2 parallel planes, as shown in Fig.1.left. This simplification is done assuming constant radiance of a light ray from point to point, and given that an RGB sampling of the wavelength is performed by the color filters coupled with the CCD sensors.

Fig. 1: (left) *Illustration of the two planes parameterization of the 4D static (one time instant) light field; (right) Rendered image at one focus; and epipolar image corresponding to horizontal red line in left image.*

The light field can be seen as capturing an array of viewpoints (called sub-aperture images in particular in the case of micro-lens based capturing devices) of the imaged scene with varying angular coordinates u and v. A photograph at a particular focus is computed from the 4D data by integrating the light across epipolar images. An epipolar image represents a 2D slice of the 4D light field (e.g. a (x,u) slice in Fig. 1.right). The epipolar image shown in Fig.1.right gives an observation of the light field at a constant y-value corresponding to the red line in the picture. Each vertical line in the epipolar image represents the light field observed at varying sub-apertures (u) of the main lens and at a given pixel location x.

Camera arrays have thus been naturally designed to capture the set of views, offering a high spatial resolution for each view but a low angular resolution (limited set of views) hence a large baseline. Targeted applications include long range depth estimation, change of viewpoint and view synthesis, such as AR content capture or movie post production. Camera gantries have also been built in which a single camera moves along a plane and takes captures at regular time intervals.

While camera arrays capture the scene from different viewpoints, hence with a large baseline, plenoptic cameras use an array of micro-lenses placed in front of the photosensor to separate the light rays striking each microlens into a small image on the photosensors pixels, and this way capture dense angular information with a small baseline. Plenoptic cameras are now available on the market with first generation Lytro cameras that typically target consumer photography via their refocusing feature, and the Raytrix cameras that instead target the industrial market with accurate, monocular depth estimation. Two optical designs have been considered for plenoptic cameras, the so-called "plenoptic 1.0" design, called unfocused plenoptic camera, in which the main lens focuses the subject on the lenslet array [6], and the "plenoptic 2.0" design [7], also called focused plenoptic camera, in which the image plane of the main lens is the object plane of the lenslet array (see Fig. 2).

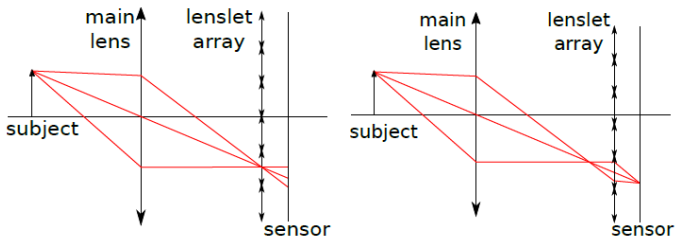


Fig. 2: *(left) plenoptic 1.0 optical design; (right) plenoptic 2.0 optical design.*

In the unfocused camera, the light emitted by one point in the 3D scene is spread over several pixel sensors of the raw lenslet data. Every pixel behind a lenslet corresponds to a different angle. Extracting views from the raw lenslet data captured by plenoptic cameras involves several processing steps [12]: devignetting which, with white images, aims at compensating for the loss of illumination at the periphery of the micro-lenses, color demosaicing, alignment of the sensor data with the micro-lens array, and converting the hexagonal sampling grid into a rectangular sampling grid.

3. Compression of the large volumes of light field data

Given their significant demand in terms of storage capacity, the problem of light field compression rapidly appeared as quite critical in the computer graphics community using light fields for image rendering. Early solutions considered for synthetic light fields were based on classical coding tools, JPEG-coding schemes [11], vector quantization [13], or wavelet coding [14] applied on each view of the 2D array separately. While the separate encoding of each view naturally allows random access to any sample of the light field, the compression factor of these solutions is however hardly exceeding 20. Predictive coding inspired from video compression techniques have then been considered for further increasing the compression factor [15], in which a few views are encoded in Intra while the other views are encoded as P-images where each block can be predicted from one of the neighboring Intra views with or without disparity compensation. Motivated by the objective of random access and progressive decoding, which is not enabled by predictive schemes, the authors in [16] consider instead a wavelet transform applied in the 4 dimensions of the light field, while a Principal Component Analysis (PCA) is used in [17]. Solutions for light field compression have then evolved following advances in mono-view and multi-view video compression (e.g. using MVC) [18].

The emerging devices for capturing real light fields also record very large volumes of data. To take only a few

examples, a Lytro Illum camera captures 61 Mpixels (15×15×625×434 pixels) while a camera array, as for example a rig of 4×4 cameras of spatial resolution 2048×1088px captures 35 Mpixels. Each pixel of course has three color components represented on 8 bits. Research for compressing real light field data has evolved along two main directions. The first type of approaches consists in directly compressing the raw lenslet data after de-vignetting and demosaicing (e.g., [19-26]) and the second type of approaches compresses the views extracted from the lenslet data (e.g., [27-29]) or captured by a camera array.

Most solutions proposed for directly encoding the lenslet data aim at exploiting spatial redundancy or self-similarity between the micro-images. The micro-image is the set of pixels behind each micro-lens and is also sometimes called elemental image. Spatial prediction modes have thus been proposed for unfocused cameras in [19] based on a concept of self-similarity compensated prediction [23][24], or using locally linear embedding techniques in [20]. Bi-directional spatial prediction modes have also been added in HEVC for encoding elemental images captured by a focused 2.0 camera [25] and by an unfocused camera [26]. The authors in [22] instead partition the raw data into tiles which are then encoded as a pseudo-video sequence using HEVC.

While the first category of methods only applies to light fields captured by micro-lens based plenoptic cameras, a second category of methods consists in encoding the set of sub-aperture images (or views) extracted from the lenslet images or captured by camera rigs. The authors in [27] form a pseudo-sequence by using a lozenge scanning order and encode this pseudo-sequence using HEVC inter-coding, while in [28] a coding order and a prediction structure inspired from those used in the multi-view coding (MVC) coding standard is proposed, showing significant performance gains compared with HEVC-Intra. In [27] and [28], inter-view correlation is exploited via motion estimation and compensation methods, as in video coding, whereas the authors in [29,30] use homographies and 2D warping instead of classical predictive coding to remove inter-view redundancy. The approach in [30] actually aims at reducing the dimension of the captured data via a low rank approximation of views aligned by homographies which are jointly optimized with the low rank model. This approach called HLRMA [30] yields, with the data set of the ICME 2016 challenge, an average PSNR gain of 2.24 dB compared with a direct encoding of the views as a video sequence scanned following a lozenge scan order starting from the central view, while the pseudo-sequence approach of [28] yields an average gain of 1.78 dB.

The need for efficient compression solutions has motivated the JPEG-Pleno group to launch an initiative for defining a light field compression standard [31], for both data captured by plenoptic cameras and by camera arrays. The standardization phase is on-going with the goal of having an international standard in January 2019.

4. Handling the spatial and angular resolution trade-off

Plenoptic cameras and smart phones equipped with multiple camera sensors can capture both spatial and angular information of light rays from a single capture [6]. However, since the sensor is limited, it is difficult to have both a dense angular and spatial light field sampling. The angular sampling is related to the number of sensor pixels located behind each microlens for Plenoptic cameras while it corresponds to the number of cameras on the wafer for mobile devices. This trade-off between angular and spatial resolution leads to a significantly lower spatial resolution compared to traditional 2D cameras [6].

A first category of approaches consists in resampling the light field to reconstruct sub-aperture images with resolutions higher than the number of micro-lenses. The authors in [32] use the super-resolution discrete focal stack transform to super-resolve the focal stack of a light field. In the same vein, a light field reconstruction approach is proposed in [33], where the de-multiplexed sub-aperture images are first interpolated with barycentric interpolation to adapt to the hexagonal layout of the micro-lenses, and then refined using pixels of neighboring views using ray interpolation. The resulting light field still contains aliasing, which is unnatural. They hence use the dictionary-based single-image super-resolution method proposed in [34] to restore each sub-aperture image separately.

A second category of methods exploits the depth information to increase both spatial and angular super-resolution. Based on the image formation model of the plenoptic camera, the authors in [35] us a depth map to estimate the reflectance in a Bayesian framework where a Markov Random Field prior was used to regularize the solution. The authors in [36] estimate disparity maps locally using epipolar plane image analysis and then use a variational model for the synthesis of super-resolved novel views. A patch-based approach was proposed in [37] where they model the light field patches using a Gaussian mixture model using disparity as prior. The spatial resolution of the 4D light field is then restored using linear minimum mean square estimation (LMMSE). Nevertheless, these methods rely on the accuracy of the disparity estimation algorithm used, which generally fail to restore reliable disparity maps in real-world light fields. Moreover, a significant number of occluded regions make their restoration difficult and

generally lead to blur artefacts in regions with large parallax.

Machine learning is used in [38 - 40] to super-resolve real-world light fields of higher quality. Deep convolutional neural networks (DCNN) were used in [38] for both spatial and angular super-resolution. This method first employs a spatial DCNN to restore each sub-aperture image separately followed by another DCNN to synthesize novel views. The resulting sub-aperture images are incoherent across sub-aperture images since they are restored separately. Deep learning was used in [39] to synthesize new views from a sparse set of input views. More specifically, a cascade of two DCNNs is used where the first one learns the disparity while the second one learns the synthesis of novel views. The authors in [40] have used principal component analysis (PCA) and ridge regression (RR) to learn a linear mapping between low- and high-resolution patch-volumes, which are a stack of collocated 2D patches from each sub-aperture image. This method exploits the light field structure and restores sub-aperture images that are more coherent.

5. Light field user interaction and editing

User interaction and light field editing (e.g., segmentation, object removal and inpainting, colorization) now common with 2D images are made difficult for light fields due to the big volume of data to be processed. Besides the computing complexity, one difficulty resides in the fact that the edits on one view must in addition be consistent across views.

Graph-cut used with Random Markov Fields (RMF) is a well-known tool for 2D image segmentation. It has thus been naturally considered for co-segmentation of multiple views using different models such as an appearance model based on color in [41] or on other cues in [42]. However, its complexity quite rapidly increases with the volume of data (number of views and dimension of each view). This is the reason why multi-view co-segmentation methods usually consider a limited number of views. In addition, with dense light fields, the baseline being much smaller, the views are much more correlated. Hence, label consistency can be more strongly enforced.

The problem of dense light field segmentation has been addressed in a semi-supervised manner allowing the user to enter scribbles on the central view. These scribbles are used in [43] to learn a joint color and depth classifier with a random forest technique. The result of the classification is then regularized using a variational approach to segment each ray using its 4D spatial and angular neighborhood. The segmentation of 9x9 views of size 768x768 however takes over 5 minutes. The authors in [44] use the same structure with an anisotropic 4D neighborhood and a SVM classifier to learn the color model, further increasing the computational load. A graph structure merging several rays coming from the same scene point is proposed in [45] which allows dividing the number of nodes by around 50, hence significantly decreasing the computational load of the regularization (9x9 views of size 768x768 are segmented in 4 to 6 sec.).

Light field editing has been essentially tackled from the angle of edits propagation in a consistent manner from one view to the other ones. The goal is to enable user interaction with the whole light field while entering inputs on one view only. A 3D voxel-based model of the scene with an associated radiance function is proposed in [46] to propagate pixel edits and illumination changes. Stroke-based editing is also described in [47] where the edits are propagated in a downsampled version of the light field to reduce the computational load.

Object removal is a complex editing task requiring the development of inpainting techniques. While 2D image inpainting has been widely addressed in the literature, there are few works on light fields inpainting. A first category of approaches inpaint one view of the light field using a 2D method and then propagates the inpainting to the other views in a consistent manner. One example of such approach is described in [48] where an exemplar patch-based method is used for the central view. For the other views, instead of searching a best matching patch in the known region of the view to inpaint, the patch is searched in the first inpainted view in order to ensure a better consistency across views. Instead of propagating the inpainting from one view to the others, the authors in [49] describe a 4D patch-based method where the consistency is ensured by minimizing a 4D patch bi-directional similarity measure. All these methods progress patch per patch in a greedy fashion and suffer from a high computational complexity. In addition, they may not ensure a global coherence on the entire light field. The authors in [50] suggest instead using a variationnal framework to define constraints on the epipolar images with the help of disparity information raising other questions related to the estimation of disparity for the region to be inpainted. Despite these preliminary works, the problem of fast light field inpainting which would enable user interaction, with a global coherence on the entire light field remains a difficult problem.

6. Conclusion

This paper gave a quick overview of main research trends in relation to a few critical problems in light field image processing. Given the very big volume of highly redundant data, even with static light fields, it became rapidly evident that progress in this area requires developments which go beyond a straightforward application or extension of well-known 2D imaging techniques. Even if most works focused on static light fields, the volume of high dimensional data becomes even more critical with video light fields for which the above problems remain largely open.

Acknowledgement

This work has been supported by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM).

References

[1] M. Levoy, Light fields and computational imaging, IEEE computer, 39(8), pp. 46-55, Aug. 2006.
[2] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array," in CVPR, vol. 2. IEEE, 2004, pp. II–294.
[3] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," ACM Trans. Graph., vol. 24, no. 3, pp.765–776, Jul. 2005.
[4] C. Zhang and T. Chen, "A self-reconfigurable camera array," in SIGGRAPH Sketches. ACM, 2004, p. 151.
[5] K. Venkataraman, D. Lelescu, J. Duparr´e, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "Picam: An ultra-thin high performance monolithic camera array," ACM Trans. on Graphics (TOG), vol. 32, no. 6, p. 166, 2013.
[6] R. Ng, "Light field photography," Ph.D. dissertation, Stanford University, 2006.
[7] T. Georgiev, G. Chunev, and A. Lumsdaine, "Super-resolution with the focused plenoptic camera," Proc. of SPIE - The International, Society for Optical Engineering, vol. 7873, 2011.
[8] G. Lippmann, La photographie intégrale. Comptes-Rendus, Académie des Sciences 146, 446-551, 1908.
[9] C.-T. Huang, J. Chin, H.-H. Chen, Y.-W. Wang, and L.-G. Chen, "Fast realistic refocusing for sparse light fields," in 2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 1176–1180
[10] M. Levoy and P. Hanrahan, Light Field Rendering, Proc. ACM Siggraph, pp. 31-42,1996.
[11] S. J. Gortler, R. Grzeszczuk, R. Szeliski, M. Cohen, The lumigraph, SIGGRAPH,pp. 43–54, 1996.
[12] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun. 2013.
[13] M. Levoy and P. Hanrahan, "Light field rendering," in Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, ser. SIGGRAPH '96. New York, NY, USA: ACM, 1996, pp. 31–42.
[14] P. Lalonde and A. Fournier, "Interactive rendering of wavelet projected light fields," in Proceedings of the 1999 Conference on Graphics Interface '99. Morgan Kaufmann Publishers Inc., 1999, pp. 107–114.
[15] M. Magnor and B. Girod, "Data compression for light-field rendering," IEEE Trans. Cir. and Sys. for Video Technol., vol. 10, no. 3, pp. 338–343, Apr. 2000.
[16] M. Magnor, A. Endmann, B. Girod, "Progressive compression and rendering of light fields," in Vision, Modelling and Visualization, 2000, pp. 199– 203.
[17] D. Lelescu and F. Bossen, "Representation and coding of light field data," Graph. Models, vol. 66, no. 4, pp. 203–225, Jul. 2004.
[18] M. Magnor and B. Girod, "Data compression for light-field rendering," IEEE Trans. Cir. and Sys. for Video Technol., vol. 10, no. 3, pp. 338–343, Apr. 2000.
[19] C. Conti, P. Nunes, and L. D. Soares, "New HEVC prediction modes for 3D holoscopic video coding," International Conf. on Image Processing, ICIP, pp. 1325–1328, 2012.
[20] L. Lucas, C. Conti, P. Nunes, L. Ducla Soares, N. Rodrigues, C. Pagliari, E. Da Silva, S. De faria, "Locally linear embedding-based prediction for 3D holoscopic image coding using HEVC," EUSIPCO, pp. 11–15, Sept. 2014.
[21] Y. Li, M. Sjostrom, R. Olsson, U. Jennehag, "Efficient intra prediction scheme for light field image compression," ICASSP, pp. 539–543, 2014.
[22] C. Perra and P.Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," Int. Conf. on Multimedia and Expo, ICME, July 2016.
[23] C. Conti, P. Nunes, and L. D. Soares, "HEVC-based light field image coding with bi-predicted self-similarity compensation ," Int. Conf. on Multimedia and Expo, ICME, July 2016.
[24] R. Monteiro, L. Lucas, C. Conti, P. Nunes, N. Rodrigues, S. Faria, C. Pagliari, E. Da Silva, and L. Suares, "Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction," Int. Conf. on Multimedia and Expo, ICME, July 2016.
[25] Y. Li, R. Olsson, and M. Sjostrom, "Compression of unfocused plenoptic images using a displacement Intra prediction ," Int. Conf. on Multimedia and Expo, ICME, July 2016.
[26] Y. Li, M. Sjostrom, R. Olsson, and U. Jennehag, "Efficient intra prediction scheme for light field image compression," ICASSP, pp. 539–543, 2014.

[27] M. Rizkallah, T. Maugey, C. Yaacoub, and C. Guillemot, "Impact of light field compression on focus stack and extended focus images," EUSIPCO, pp. 898–902, 2016.
[28] D. Liu, L. Wang, L. Li, X. Z., W. F., and Z. W, "Pseudo-sequence based light field image compression," Int. Conf. on Multimedia and Expo, ICME, July 2016.
[29] S. Kundu, "Light field compression using homography and 2d warping," in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, March 2012, pp. 1349–1352
[30] X. Jiang, M. Le Pendu, R. A. Farrugia, and C. Guillemot, "Homography-based low rank approximation of light fields for compression," IEEE J. on Selected Topics in Signal Processing, special issue on light field image processing, submitted, 2017. (https://www.irisa.fr/temics/demos/lightField/LowRank2/LRcompression.html)
[31] "JPEG Pleno Call for Proposals on Light Field Coding", ISO/IEC JTC 1/SC29/WG1N74014 74th Meeting, Geneva, Jan. 2017
[32] F. Perez Nava, J. P. Like, "Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera," *IEEE Proc. Of 3DTV Conference*, May 2009.
[33] D. Cho, M. Lee, S. Kim, Y-W. Tai, "Modeling the calibration pipeline of the Lytro camera for high quality light-field image reconstruction," in *Int. Conf. on Computer Vision*, 2013.
[34] J. Yang, J. Wright, T. Huang, Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2008.
[35] T.E. Bishop, S. Zanetti, P. Favaro, "Light field superresolution," in *IEEE ICCP*, 2009
[36] S. Wanner, B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606-619, 2014
[37] K. Mitra, A. Veeraraghavan, "Light field denoising, light field, superresolution and stereo camera based refocusing using a gmm light field patch prior," in *IEEE Proc. Int. Cong. On Computer Vision and Pattern Recognition*, 2012.
[38] Y. Yoon, H-G. Jeon, D. Yoo, J-Y. Lee, I.S. Kweon, "Light-Field Image Super-Resolution using Convolutional Neural Networks," in *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 848-852, Jun. 2017.
[39] N.K. Kalatari, T-C. Wang, R. Ramamoorth, "Learning-based view synthesis for light field cameras," in *ACM Trans. On Graphics*, vol. 35, no 6, Nov. 2016.
[40] R. Farrugia, C. Galea, C. Guillemot, Super Resolution of Light Field Images using Linear Subspace Projection of Patch-Volumes, IEEE J. on Selected Topics in Signal Processing, special issue on light field image processing, submitted, 2017 (https://www.irisa.fr/temics/demos/lightField/LFSR/LSsuperresolution.htm)
[41] C. Rother, T. Minka, A. Blake, V. Kolmogorov, Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2006.
[42] L. Mukherjee, V. Singh, J. Peng, Scale invariant cosegmentation for image groups. In: IEEE Conference on c Computer Vision and Pattern Recognition (CVPR), 2011.
[43] S. Wanner, C. Straehle, B. Goldlucke, Globally consistent multi-label assignment on the ray space of 4D light fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[44] H. Mihara, T. Funatomi, K. Tanaka, H. Kubo, H. Nagahara, and Y. Mukaigawa, "4d light-field segmentation with spatial and angular consistencies," in Proceedings of IEEE Int. Conf. on Computational Photography (ICCP), 2016.
[45] M. Hog, N. Sabater, and C. Guillemot, "Light field segmentation using a ray-based graph structure," in European Conference on Computer Vision. Springer, 2016, pp. 35–50.
[46] S. M Seitz and K. N. Kutulakos. Plenoptic Image Editing. International Journal of Computer Vision, 48(2):115–129, 2002.
[47] A. Jarabo, B. Masia, and D. Gutierrez. Efficient propagation of light field edits. In Proc. of the V Ibero-American Symposium in Computer Graphics, SIACG 2011, pages 75–80, 2011.
[48] Williem, K. W. Shon, and I. K. Park. Spatio-angular consistent editing framework for 4D light field images. Multimedia Tools and Applications, 75(23):16615–16631, 2016.
[49] Ke-Wei Chen, Ming-Hsu Chang, and Yung-Yu Chuang. Light field image editing by 4D patch synthesis. In 2015 IEEE International Conference on Multimedia and Expo, ICME 2015, Turin, Italy, June 29 - July 3, 2015, pages 1–6.
[50] B. Goldluecke and S. Wanner, The variational structure of disparity and regularization of 4d light fields. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Portland, pages 1003–1010, June 23-28, 2013.

Christine Guillemot holds a PhD degree from ENST - Paris. From Nov. 1985 to Oct. 1997, she has been with FRANCE TELECOM. From Jan.90 to mid 91, she has worked at Bellcore, NJ, USA, as a visiting scientist. Since Nov. 1997, she is 'Director of Research' at Inria, head of a team dedicated to the design of algorithms for the image and video processing chain, with a focus on analysis, representation, compression, and editing, including for emerging modalities such as high dynamic range imaging and light fields. She has served as Associate Editor for several IEEE journals (IEEE Trans. on Image Processing, IEEE Trans. on Circuits and Systems for Video Technology, IEEE Trans. on Signal Processing, …). She is currently Senior Area Editor of IEEE Trans. On Image Processing (2016-2017) and member of the IEEE- trans. On Multimedia steering committee. She has been a member of the IEEE IMDSP (2002-2007) and IEEE MMSP (2005-2008) technical committees. Christine

Guillemot is an IEEE fellow.

Reuben Farrugia received the first degree in Electrical Engineering from the University of Malta, Malta, in 2004, and the Ph.D. degree from the University of Malta, Malta, in 2009. In Jan. 2008 he was appointed Assistant Lecturer with the same department and is now a Senior Lecturer. He has been in technical and organizational committees of several national and international conferences. In particular, he served as General-Chair on the IEEE Int. Workshop on Biometrics and Forensics (IWBF) and as Technical Programme Co-Chair on the IEEE Visual Communications and Image Processing (VCIP) in 2014. He has been contributing as a reviewer of several journals and conferences, including IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video and Technology and IEEE Transactions on Multimedia. On Sept. 2013 he was appointed as National Contact Point of the European Association of Biometrics (EAB).

Performance evaluation of light field pre-processing methods for lossless standard coding

João M. Santos[†], Pedro A. A. Assuncao[*‡], Luís A. da Silva Cruz[*†],*
Luís Távora[‡], Rui Fonseca-Pinto[‡] and Sérgio M. M. Faria[*‡]*
[*]Instituto de Telecomunicações, Portugal
[†]University of Coimbra, Coimbra, Portugal
[‡]Instituto Politécnico de Leiria, Leiria, Portugal
e-mails: {joao.santos, amado, luis.cruz, sergio.faria}@co.it.pt,
{luis.tavora,rui.pinto}@ipleiria.pt

1. Introduction

The concept of Light Field (LF) was first introduced by Lippman in 1908 [1], expressing the idea of capturing all the information conveyed by the light rays. Since those early days, worldwide research and new technological developments led to the production of consumer-grade LF cameras that are now available for researchers and consumers alike. These are mainly characterised by their ability to record not only light intensity but also the directionality of light-rays that reach the camera. This technology is based on a specially designed array of micro-lenses (MLA) that is placed in front of the camera sensor [2]. As shown in Figure 4, light rays from the same point in the scene, but reaching the camera from different directions, correspond to different samples in the camera sensor, thus sampling the light field. While these sample values represent the light intensity, their spatial positions within the micro-image created by each micro-lens are associated with the direction of the corresponding light ray. Therefore, the sample array (i.e., light field image) captured by the camera encodes the light intensity reaching the sensor from slightly different perspectives multiplexed into different spatial positions. More details about LF acquisition systems can be found in [3,4,5].

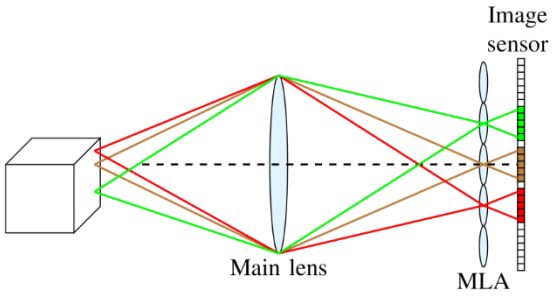


Figure 4. Light-field imaging acquisition system

Different applications of LF are mostly supported by computational methods that can be implemented as post-processing operations such as image rendering with different focal planes, depth-of-field or viewing perspectives [2,6,7] and extraction of depth maps [8]. Since LF cameras capture a much larger amount of data in comparison to conventional cameras, the need for compression is of utmost importance. To this aim, standard image and video encoders have been used, but these usually fail to optimally exploit the intrinsic redundancy of LF data. Alternatively, in recent years, several approaches for lossy compression of LFs have been proposed, trying to exploit the specific characteristics of LF representation data, namely the correlation between neighbouring micro-images [9] and the correlation in sub-aperture images, using three-dimensional transforms [10]. For other applications, where the full accuracy of the originally captured LF needs to be preserved, lossless encoding must be used for the entire representation data. Some LF lossless coding methods have been reported in the literature. For instance, in [11], Perra encodes the non-rectified lenslet image by exploiting the correlation between micro-images, like Henlin *et al*, in [12], where the proposed method encodes the sub-aperture images extracted from the rectified lenslet data, exploiting inter-image correlations by applying different predictors to regions of the same depth.

Commonly, LF cameras generate as output an RGB lenslet image. However, this image format may not be the most adequate format for an efficient compression. Therefore, pre-processing techniques may be used to convert the data to a format that enables higher compression performance of current standard encoders. In this context, this paper presents and analyses the use of two types of pre-processing techniques that increase the compression efficiency of standard lossless encoders, namely lenslet data rearrangement and colour transformation.

The paper is organised as follows: Section 2 describes some image representation formats for encoding LF image

data and the use of reversible colour transformations. Section 3 presents experimental results and Section 4 concludes the paper.

2. Pre-processing of LF images

2.1 LF data arrangements

The underlying idea behind pre-processing LFs is to obtain different representation formats, which may have an impact on the coding efficiency. Since standard image and video encoders exploit the spatial and temporal correlations of neighbouring pixels, it is expected that different spatio-temporal arrangements and colour formats influence the performance of encoding algorithms. As mentioned above, the common output of LF cameras is formatted as a matrix of lenslet images, which present a regular alveolar structure of micro-images, which in turn form a rectangular grid of micro-images after rectification. The micro-image structure of a lenslet LF is shown in Figure 5.

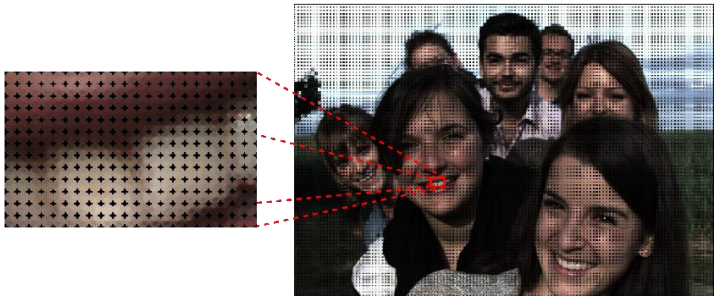


Figure 5. Rectified lenslet image and a zoomed detail from light field *Friends*_1 [13].

Since LF images include directional information about light rays, it is possible to extract different perspectives or views from a rectified lenslet, by selecting pixels located at the same spatial position in all micro-images. The black pixels located at the corners of the micro-images correspond to the black images shown in the corners of the light field image with all views (Figure 3). The whole set of different views form a stack of sub-aperture images. Then, rather than encoding the light field image as a still picture, such stack of images can be arranged as a pseudo - video sequence (PVS) and used as the input video of a standard video encoder.

The coding performance is studied for three different rearrangement methods used to create pseudo-temporal sequences from sub-aperture images: Raster, Spiral [14] and New Spiral, as shown in Figure 6. In the New Spiral method, the black images are put sequentially at the end of the pseudo video sequence by following the yellow scan and then the red one (Figure 3c). Another possible LF representation is the Epipolar format. Epipolar Plane Images (EPI) [15] are defined by the cross-section of a given PVS stack. Figure 7a) represents a cross section of a PVS stack and Figure 7b) the corresponding epipolar image.

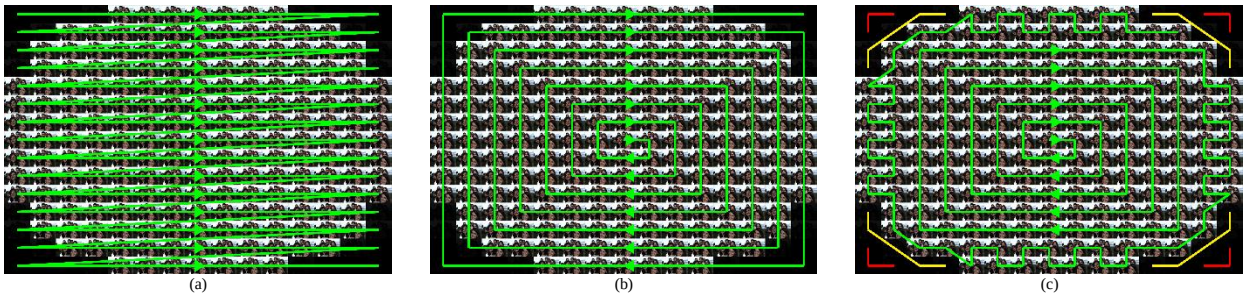
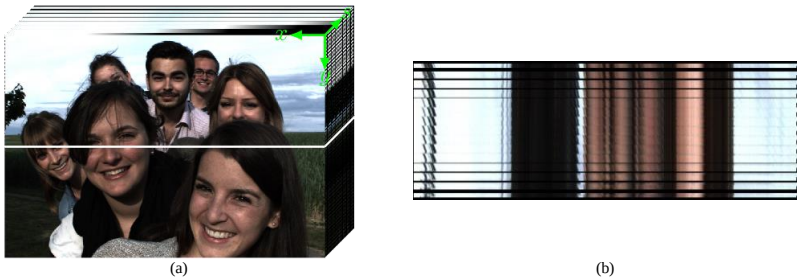


Figure 6. Representation of the used scans: (a): Raster, (b): Spiral and (c) New Spiral.

2.2. Colour transforms

Since natural images in RGB format present significant correlation between the three components [16], in general this is not the most efficient format for compression, due to the cross-component redundancy. Therefore applying decorrelation methods, *i.e.* reversible colour transforms, prior to the encoding process can improve the coding efficiency. There are several lossless colour transforms designed for natural images that can also be used with LF images. In this study, the following transforms are studied: A2 [17], LDgDb [18], LDgEb [18], RCT [19] (used in the

JPEG 2000 standard), RDgDb [18] and YCoCg [20] (used in the JPEG XR standard). Despite the high decorrelation capability of these techniques reported in the literature [18], their use in combination with standard encoders does not always result in higher compression efficiency.



(a) (b)

Figure 7. Example of a subset of EPIs extracted from the pseudo-video sequence *Friends*_1 using the *Raster* scan [13].

3. Experimental Results and Performance Analysis

The impact of the reversible colour transformations and LF data arrangements on the compression efficiency of standard lossless encoders is presented in this section. In these experiments, three lossless encoders were used, JPEG2000 [19], JPEG-LS [21] HEVC [22]. The ICME 2016 Grand Challenge dataset [13] was used along with Dansereau's LF image processing toolbox version 0.4 (LF Toolbox v0.4) [23]. The results obtained are shown in Table 1, where the lower values of bpp for each *encoder - data arrangement* pair are underlined, and the best result for each data arrangement is in bold.

Table 1. Average bitrates (bpp) for all combinations of colour transformations, data arrangements and scan pattern (for HEVC).

Encoder	Arrangement	RGB	A2	LDgDb	LDgEb	RCT	RDgDb	YCoCg
JPEG2000	Lenslet	13.21	10.31	9.95	10.08	9.92	9.99	9.97
	PVS	10.67	9.73	9.46	9.49	9.41	9.60	9.41
	EPI	10.28	9.12	8.87	8.91	**8.84**	8.98	8.86
JPEG-LS	Lenslet	10.14	8.83	8.54	8.65	8.52	8.60	8.55
	PVS	10.21	9.85	9.59	9.62	9.55	9.74	9.57
	EPI	8.87	8.39	8.16	8.20	**8.13**	8.27	8.17
HEVC	Lenslet	10.04	9.64	9.39	9.45	9.37	9.45	9.41
	PVS Raster	7.63	7.62	7.43	7.43	7.41	7.47	7.44
	PVS Spiral	7.55	7.54	7.36	7.35	7.34	7.39	7.37
	PVS New-Spiral	7.53	7.52	7.34	7.33	**7.32**	7.37	7.34
	EPI Raster	8.67	8.59	8.39	8.38	8.36	8.46	8.39
	EPI Spiral	8.44	8.37	8.17	8.16	8.14	8.23	8.17
	EPI New-Spiral	8.08	8.02	7.81	7.81	7.79	7.88	7.83

For intra encoders, such as JPEG2000 and JPEG-LS, the EPI data format presents the highest compression efficiency, for all colour transforms. Note that in this case inter-view (inter-frame) correlation is not exploited, thus only intra-subaperture frame correlations are used. For the video/inter-frame encoder, the best result is obtained for the sub-aperture images' stack. Regarding the colour transforms, all encoders present the highest compression efficiency when using the RCT transform, regardless of the data format. The results obtained show that organizing the data in the form of PVS produces higher compression ratios due to the higher degree of spatial correlation between consecutive views, which positively impacts the coding performance. It is also worth noting that the PVS New-Spiral scanning order might further improve the coding performance, which is justified by better data compaction due to aggregation of the black images at the end of the PVS. The relative gain in compression efficiency, when comparing the Spiral and New-Spiral scanning orders, is greater for EPI than PVS. This is due to the higher impact of the New Spiral scanning order on EPI coding, where the improved spatio-temporal correlation is better exploited. In the work presented in [18], the authors describe a similar study for natural images, which shows that, in general, the RDgDb transform has higher impact in the compression performance. In the same work, the compression efficiency of RCT is slightly lower than the RDgDb transform, i.e., a difference lower than 0.06bpp.

However, in the case of LFs, the results in Table 1 show that RCT yields better results than RDgDb, up to 0.19bpp. This suggests that, even for different data arrangements, LF lossless coding does not present the same behaviour as classic 2D images. Thus, this may be an open topic for further research in order to find better reversible colour transformations for LF coding.

4. Conclusion

The efficient compression of Light Field information is currently under active research by the image and video coding and communications community. This paper provided a review and exploratory study on the use of pre-processing techniques for LF lossless encoding, namely data arrangement and colour transformations and their impact on the compression efficiency of standard encoders. The results presented in this paper clearly show that the commonly used RGB format does not result in the best compression performance for the standard encoder. It is also shown that lossless coding using a single light field image (i.e., lenslet format) yields worse performance than coding the same data arranged as a PVS of sub-aperture images (i.e., video encoding). Considering both types of pre-processing and the different standard encoders, the best results were obtained using an HEVC encoder together with the RCT transform and the LF data arranged as a PVS of sub-aperture images.

Acknowledgement

This work was funded by DERMOPLENO Project in the scope of R&D Unit 50008, financed by FCT/MEC through national funds, co-funded by FEDER – PT2020 partnership agreement and PhD Grant SFRH/BD/114894/2016.

References

[1] G. Lippmann, "Epreuves reversibles donnant la sensation du relief," *J. Phys. Theor. Appl.*, vol. 7, pp. 821-825, 1908.

[2] M. Harris, "Focusing on everything," *IEEE Spectrum*, vol. 49, pp. 44-50, May 2012.

[3] M. Levoy and P. Hanrahan, "Light Field Rendering," in *Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, New York, NY, USA, 1996, pp. 31-42.

[4] C. Zhou and S. K. Nayar, "Computational Cameras: Convergence of Optics and Processing," *IEEE Trans. Image Process.*, vol. 20, pp. 3322-3340, Dec 2011.

[5] M. Levoy, "Light Fields and Computational Imaging," *Computer*, vol. 39, pp. 46-55, Aug 2006.

[6] D. Donatsch, S. A. Bigdeli, P. Robert, and M. Zwicker, "Hand-held 3D light field photography and applications," *The Visual Computer*, vol. 30, p. 897, Jun 2014.

[7] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Linear volumetric focus for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 34, pp. 1-20, Mar 2015.

[8] H. G. Jeon et al., "Accurate depth map estimation from a lenslet light field camera," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, june, 2015, pp. 1547-1555.

[9] A. Aggoun, "A 3D DCT Compression Algorithm For Omnidirectional Integral Images," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 2, may 2006, p. II.

[10] A. Aggoun, "Compression of 3D Integral Images Using 3D Wavelet Transform," *J. Display Technol.*, vol. 7, pp. 586-592, Nov 2011.

[11] C. Perra, "Lossless plenoptic image compression using adaptive block differential prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, 2005, pp. 1231-1234.

[12] P. Helin, P. Astola, B. Rao, and I. Tabus, "Sparse modelling and predictive coding of subaperture images for lossless plenoptic image compression," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Hamburg, june, 2016, pp. 1-4.

[13] M. Rerabek and T. Ebrahimi, "New Light Field Image Dataset," in *International Conference on Quality of Multimedia Experience QoMEX*, Lisbon, Portugal, 2016.

[14] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data formats for high efficiency coding of Lytro-Illum light fields," in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, nov 2015, pp. 494-497.

[15] S. Wanner, J. Fehr, and B. Jähne, "Generating EPI Representations of 4D Light Fields with a Single Lens Focused Plenoptic Camera," in *Advances in Visual Computing: 7th International Symposium, ISVC 2011, Las Vegas, NV, USA, September 26-28, 2011. Proceedings, Part I*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 90-101.

[16] W. Pratt, "Spatial Transform Coding of Colour Images," *IEEE Trans. Commun. Technol.*, vol. 19, pp. 980-992, Dec 1971.

[17] T. Strutz, "Adaptive selection of colour transformations for reversible image compression," in *Proc. 20th European Signal*

Processing Conf. (EUSIPCO) 2012, aug 2012, pp. 1204-1208.

[18] R. Starosolski, "New simple and efficient colour space transformations for lossless image compression," *Journal of Visual Communication and Image Representation*, vol. 25, pp. 1056-1063, 2014.

[19] ITU, "ITU-T Recommendation T.812: Information technology – JPEG 2000 image coding system: An entry level JPEG 2000 encoder," International Telecommunication Union, Geneva, Switzerland, ITU 2007.

[20] ITU, "ITU-T Recommendation T.832: Information technology – JPEG XR image coding system - Image coding specification," International Telecommunication Union, Geneva, Switzerland, ITU 2016.

[21] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS," *IEEE Trans. Image Process.*, vol. 9, pp. 1309-1324, Aug 2000.

[22] G. J. Sullivan et al., "Standardized Extensions of High Efficiency Video Coding (HEVC)," *IEEE J. Sel. Topics Signal Process.*, vol. 7, pp. 1001-1016, Dec 2013.

[23] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, Calibration and Rectification for lenselet-Based Plenoptic Cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2013.

João M. Santos received the B.Sc. and M.Sc. degres in Electrical and Electronic Engineering from Instituto Politécnico de Leiria in 2013 and 2016, respectively. Since 2016, he is a doctoral student in Electrical and Computer Engineering at University of Coimbra. His research interests include digital signal and image processing, namely image and video compression. He is currently working in the lossless compression of Light Fields. He has published four conferences papers and a journal paper in this area. He is a student member of IEEE.

Pedro A. A. Assunção received the Licenciado and M.Sc. degrees from the University of Coimbra, in 1988 and 1993, respectively, and the Ph.D. in Electronic Systems Engineering from the University of Essex, in 1998. He is currently professor of Electrical Engineering and Multimedia Communication Systems at the Polytechnic Institute of Leiria and a senior researcher at the Institute for Telecommunications, Portugal. His current research interests include high efficiency and 360-degree, multiview video and light field coding, multiple description and robust coding, error concealment and quality evaluation. He is a senior member of the IEEE.

Luís A. da Silva Cruz received the Licenciado and M.Sc. degrees from the University of Coimbra, Portugal, in 1989 and 1993 respectively. He also holds an MSc degree in Mathematics and a Ph.D. degree in Electrical Computer and Systems Eng. from Rensselaer Polytechnic Institute (RPI), US, granted in 1997 and 2000 respectively. He is Assistant Professor at the Dep. of Electrical and Computer Engineering of the University of Coimbra, Portugal and researcher of the Institute for Telecommunications of Coimbra where he works on video processing and coding, medical image processing and wireless communications. He is a member of the EURASIP, SPIE and IEEE.

Luís Távora, received the Licenciado degree in Physics with Engineering from the University of Coimbra in 1993, and a PhD Physics from the University of Surrey (UK) in 1998. He is currently a professor of Physics at the Polytechnic Institute of Leiria. His research interests include modelling stochastic and deterministic imaging systems, in particularly x-ray based devices. He is current research activities involve Light Fields acquisition and processing.

Rui Fonseca-Pinto received the BSc degree in Mathematics from the University of Évora in 2000, and the M.Sc. in Applied Mathematics (Medical Physics) in 2004. He received the PhD in Biomedical Engineering and Biophysics in 2010 from the University of Lisbon, Portugal. He holds the BSc degree in Medicine from the University of Lisbon. His current research interests include interdisciplinary areas in Medicine and Engineering, in the field of the Autonomic Nervous System and Dermoscopic Image Processing of skin melanocytic lesions. Currently he is Professor at the Polytechnic Institute of Leiria, and Researcher at the Instituto de Telecomunicações.

Sérgio M. M. Faria (S'92–M'96–SM'08) received the E.E. and M.Sc. degrees in electrical engineering from the Universidade de Coimbra, Portugal, in 1988 and 1992, respectively, the Ph.D. degree in Electronics and Telecommunications from the University of Essex, U.K., in 1996, and the title of "Agregado" from the University of Lisbon, Portugal, in 2014. He is Professor of Instituto Politécnico de Leiria and Senior researcher at Instituto de Telecomunicações, Portugal. His current research interests include 2D/3D and Light Field image and video processing and coding, motion representation, and medical imaging. He is Area Editor of Signal Processing: Image Communication. He has been Scientific and Program Committee Member of many international conferences.

Towards Adaptive Light Field Video Streaming

Peter A. Kara[1], *Aron Cserkaszky*[2, 3], *Attila Barsi*[2], *Maria G. Martini*[1], *Tibor Balogh*[2]
[1]*WMN Research Group, Kingston University, London, UK*
[2]*Holografika, Budapest, Hungary*
[3]*Pazmany Peter Catholic University, Budapest, Hungary*
{p.kara, m.martini}@kingston.ac.uk, {a.cserkaszky, a.barsi, t.balogh}@holografika.com

1. Introduction

Light field technology at the time of this paper can be considered to be one of the hottest topics in the research area of future 3D visualization. We emphasize the term "future"; while virtual reality (VR) has already entered the consumer market and it is widely commercially available with booming content and services, light field displays are still quite far from becoming widespread in our everyday lives. It needs to be noted that, although they are not yet present in the consumer market, they are indeed commercially available, as certain horizontal-only parallax (HOP) displays can be purchased, and their usage is spreading in the industry.

The sheer fact that no additional viewing equipment is required in order to experience the immersive visual content in 3D makes it desirable from the perspective of the users. At the end of the day, it is the Quality of Experience (QoE) that determines the true "value" of the systems, products and services, and the failure to satisfy the users can result in the termination of entire technologies (i.e., the impending fate of stereoscopic 3D televisions). The QoE of the glasses-free, naked-eye light field visualization is not only a challenge to assess due to specific visual phenomena that do not apply to other forms of visualization, but also because the corresponding subjective quality evaluation techniques are not yet standardized.

However, QoE research is already active in the field; QoE studies are being published, some of which one day might be looked back as pioneering work for the perceived quality of light field visualization. The majority of these studies has been and is being carried out on Holografika's HoloVizio displays, as these are currently the only commercially available light field displays. For example, the work of Tamboli *et al.* [1] and Adhikarla *et al.* [2] used the HoloVizio HV721RC light field display [3] for their studies, while the tests of Dricot *et al.* [4] and Subbareddy *et al.* [5] were carried out on the HoloVizio C80 cinema system [6]. It is common in most studies that visual stimuli are created by the researchers themselves – particularly designed for the selected display – either by capturing a scene via camera(s) or by generating virtual stimuli via rendering [7].

Even though we are far from the entry of light field displays to the consumer market, there are already efforts towards light field streaming. It is of course important to differentiate between the general streaming of a static scene [8] [9] and actual video streaming [10] [11]. Solutions which allow the user to access a portion of a light field along a chosen trajectory is known as interactive light field streaming [12] [13], which are motivated by the massive data requirements of light field visualization and intend to minimize the necessary transmission rate. Indeed, a single static scene can already reach a data size of several hundred MB before conversion, depending vastly on the Field of View (FOV). As a static scene can be practically considered to be a video frame, it is not difficult to calculate that the size of a 90-minute video with 60 frames per second can be far over a hundred PB.

Compression can most certainly reduce the data size; however, even after compressing the light field, the data to be transmitted is simply immense. Besides compression, data could be reduced by the degradation of certain parameters of the light field at hand. It is sufficient to think about conventional 2D adaptive video streaming, where a lower bandwidth is compensated by video frames in smaller spatial resolutions in order to avoid serious interruptions in video playback, such as rebuffering events.

In this paper, we give an overview of the QoE researches we have performed in light field visualization, and based on our prior and current findings, we propose the fundamentals of a novel protocol for adaptive light field streaming. The aim of the protocol is to enable QoE-centric playback with fewer interruptions, while taking into consideration the aspects of perceived visualization quality, in order to construct efficient and user-friendly future light field video streaming services.

The remainder of the paper is structured as follows: Section 2 briefly describes our related research in the perceived quality of light field visualization. Section 3 introduces our proposal for adaptive light field video streaming. The paper is concluded by Section 4.

2. Research on the QoE of Light Field Visualization

In case of light field visualization, the angle measured from the display's perspective in which the content can be observed is the FOV, which is not to be confused with the FOV of other visualization technologies, where is it observer-centric. Evidently, the bigger the FOV, the more data is required, if all other parameters – such as the density of source views – are fixed. Light field displays vary in FOV; while some only support 45 degrees [6] [14] or 70 degrees [3], having a full-horizontal 180-degree FOV is also practically achievable [15]. For this wide-FOV display, the HoloVizio 80WLT [15], we carried out a subjective assessment research [16], where test participants had to evaluate different FOVs. As the display had a fixed FOV and was not recalibrated between test cases, we used the rendered test stimuli to implement the different FOV values, ranging from 15 degrees to 180 degrees. For example, the test case with a 90-degree FOV had its corresponding source views outside the FOV (45 degrees from the left and the right) replaced with the background color of the stimuli, thus the stimuli were only visible inside the FOV. The 12 selected FOV values (15, 30, 45 etc.) were investigated for general acceptance, willingness to use and willingness to pay (WTP). Our results indicate that no significant differentiation can be made over 135 degrees, thus a larger FOV does not come with added value. However, this does not imply that light field visualization above this extent is pointless; e.g., in an exhibition use case scenario, where a multitude of human observers wish to view the 3D content simultaneously, having a larger FOV is perfectly valid. For a home entertainment use case scenario, where the content is streamed over the network, providing a 135-degree FOV instead of a 180-degree one induces 25% of data reduction, while still satisfying the user's needs.

Similarly to 2D visualization, the spatial resolution of the source data set fundamentally affects the size of the data to be transmitted. However, during light field visualization, light rays hit irregular positions on the holographic screen, eliminating the concept of pixels, yet this concept still applies to the source data, i.e., the discrete views of a rendered scene that are to be converted. Due to the properties of light propagation, where light rays are emitted from the optical engine array of either a back-projection or front-projection system, lower source spatial resolution manifests in blur instead of pixelation. In our research [17], we compared spatial resolutions up to 4K Ultra HD, by directly rendering stimuli in the given resolutions. The paired comparisons were made using a 5-point Degradation Category Rating (DCR) [18] scale, which collected subjective data on the perceivable differences and also on the dissatisfaction evoked by the quality implications of lower resolutions. We found that test participants were unable to distinguish the highest resolutions, and more importantly, that even very low resolutions could be acceptable, in the sense that their quality degradations were deemed only "slightly annoying" compared to the highest available resolutions.

Unlike multi-view autostereoscopic 3D displays, where the content horizontally repeats itself in a small angle inside the FOV, light field displays utilize the entire FOV, which means that the content can be seen from genuinely different angles inside the FOV, depending where the observer is located, and thus the number of simultaneous viewers is not limited by the number of so-called "sweet points". However, it is not enough to have a sufficiently large FOV in which the content can be observed in an angular-dependent manner. The immersive 3D experience comes from the continuous horizontal motion parallax. This means that during the sideways transition of the observer, the parallax effect is smooth, and there are no discrete views visible. This requires a certain light ray density, which is referred to as angular resolution. It is important to differentiate the angular resolution of the display and the content. The display's angular resolution is a given fix value, determined by the layout and parameters of the optical engines of the system [19]. The angular resolution of the content is calculated from the number of source views (that are to be converted) over the size of the FOV. If the source content angular resolution is not high enough, light field visualization will suffer the crosstalk effect and discrete image borders might appear as well. The higher the angular resolution is, the smoother the horizontal parallax is, but also the higher the transmission rate requirement is; more source views mean more data. Therefore, it is important to provide a sufficiently high angular resolution in order to have an excellent user experience, while maintaining a supportable total data size. To investigate the thresholds of parallax perception, we conducted a series of measurements that aimed at the reduction of angular resolution. We rendered stimuli in different angular resolutions, and displayed them on the HoloVizio C80 cinema system [6] during quality assessment [20]. The display was calibrated to a 45-degree FOV, and the number of source views ranged from 15 to 150. As angular resolution is the ratio of source views and FOV, e.g., the test condition with 90 views corresponded to 2 views per degree; in the literature, referring to this extent as an angular resolution of 0.5 is also common. The findings show the strong correlation between the perceived visual quality and angular resolution, and point out that an angular resolution of 1 view per degree or lower is not acceptable.

Horizontal motion parallax refers to the sideways motion of the observer; however, no actual physical motion is required to experience the parallax effect. If the observer is in a fixed position – either standing or sitting – there is always a natural movement of the head. If the head of the observer is somehow perfectly fixed, the parallax effect still applies because of the two eyes; in fact, even in case of a single eye, the movement of the eye alone is enough to perceive the horizontal parallax of light field displays. Yet it needs to be noted that the smooth, continuous horizontal motion parallax during the sideways movement of the observer comes with a different visual experience, compared to the scenario of a fixed-position observer. Our QoE research that involved observers with static viewing locations supports the hypothesis that the lack of observer movement can increase tolerance towards angular resolution reduction [21]. This is particularly relevant if we consider a use case scenario which does not enable user movement, such as a cinema [22]. Exhaustive comparisons are being carried out in order to reinforce our findings and conclusions, and to determine precise threshold level differences.

The angular resolution of the source content can be increased in certain ways. One approach is to apply light field reconstruction to the data set, but depending on the algorithm used and the input of the method, the introduced visual artifacts can impact perceived quality more than the low angular resolution, although it may also improve contrast, from which QoE can benefit [23]. Interpolation techniques create intermediate views between the existing ones, through which the total number of views and thus angular resolution can be increased. We conducted a subjective quality assessment test [24] where participants had to compare interpolated data sets (interpolation based on disparity and sweeping planes) with their inputs (data sets with low angular resolution) and the corresponding ground truths (directly rendered in high angular resolution), using a 7-point comparison scale [25]. With an input of 1 view per degree, both techniques performed notably better than their inputs, boosting QoE through a significantly higher angular resolution. For lower inputs (e.g., 10 source views), interpolation based on disparity could not benefit the perceived quality, unlike the sweeping planes approach.

3. Proposed Novel Protocol

According to the best knowledge of the authors, this paper is the first contribution in the literature on the adaptive streaming of light field video content. The proposed novel protocol is described on the level of fundamental operation, but precise parametrization is not given as the corresponding researches are currently being carried out. The full protocol with detailed synchronization of parameters (matching spatial and angular resolution values) is yet to be published.

The core of the proposed adaptive light field streaming solution is to store different quality representations of the content and provide what is suitable for the available bandwidth, just as in case of conventional 2D streaming [26]. However, the main difference here is that not only spatial resolution, but angular resolution is considered as well; insufficient bandwidth would result in the reduction of angular resolution to a tolerable extent. What is particularly beneficial regarding the protocol is how spatial and angular resolutions affect each other; content with a given lower angular resolution can be just as well or even better tolerated when the spatial resolution is also lower. This hypothesis originates from the perceptual phenomenon of blurred light field visualization at low spatial resolution; such blur can reduce the visual degradations of low angular resolution, particularly the discrete image borders. The results of the related subjective quality assessments are yet to be disseminated.

The protocol is designed for unconverted light field data. It does not apply to converted content, as the data has fixed spatial and angular resolution after conversion, which is always the same for a given light field display regardless of content parameters. Conversion is performed real-time, thus it is feasible to send unconverted data over the network for streaming purposes. In case the server knows the parameters of the display, converted data can be transmitted and conversion at the client side can be skipped. However, if the parameters of the unconverted data – e.g., spatial resolution – are lower than the capabilities of the display, the converted data is likely to be larger in size, thus it is more cost-effective to transmit the unconverted content.

The light field display's FOV and interpolation techniques are not considered by the proposed protocol. Sending light field data for only a portion of the FOV that is being utilized can significantly reduce the transmission rate, but it requires real-time information on the observer's (or observers') location. Such systems are feasible; however, the initial protocol is dedicated to regular display solutions and does not rely on user tracking. Interpolation techniques could greatly benefit transmission solutions, as sparse data sets could be interpolated into content with high angular resolution. Directly involving interpolation in adaptive or any kind of streaming is unfeasible at the time of this paper, as the computational requirements of such techniques are far too high to enable a run time that is suitable for real-time solutions. As offline-only techniques, they can improve the QoE, and could actually be used on the server side when preparing the different quality representations.

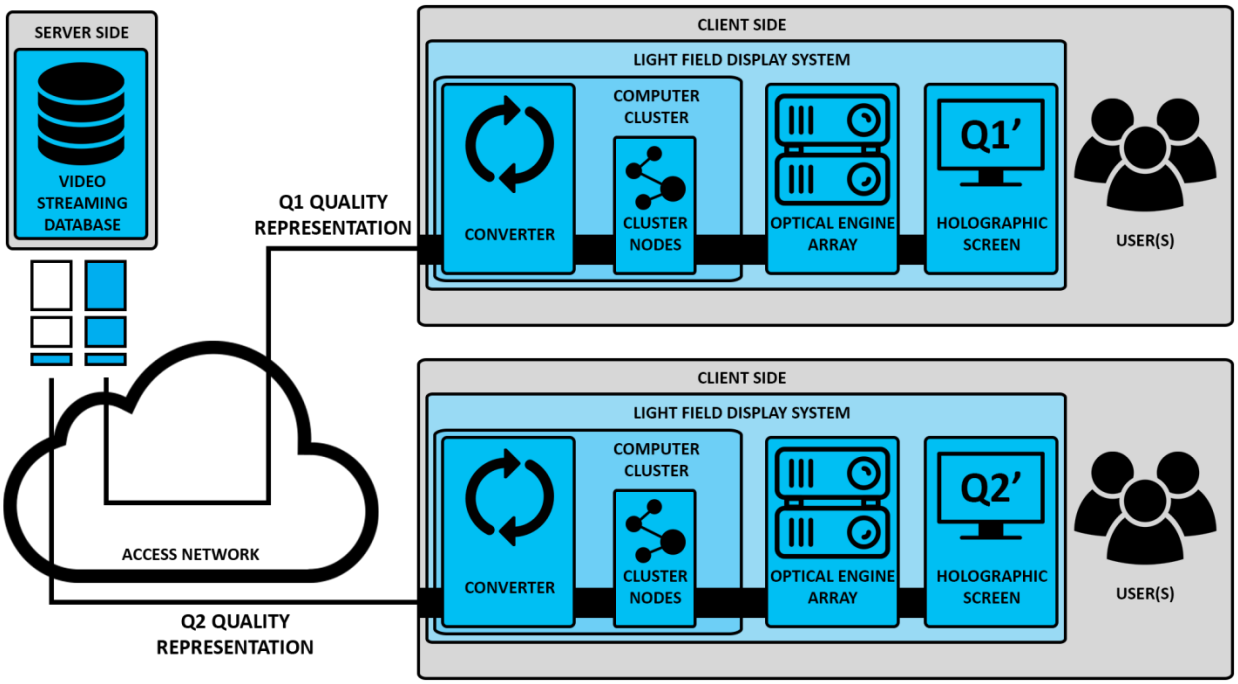


Figure 1. Dynamic adaptive light field streaming over the network.

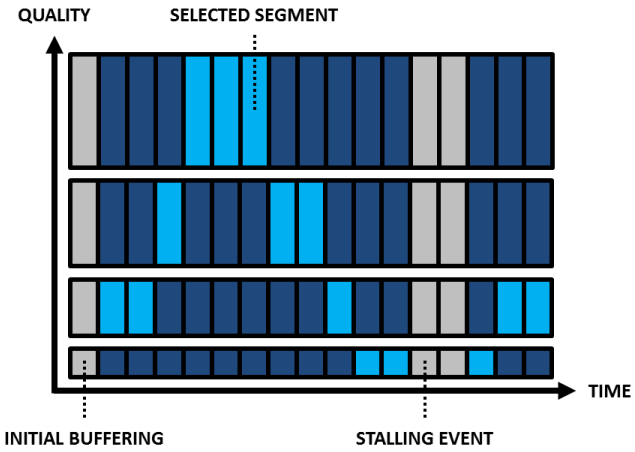


Figure 2. Quality switching between segments with different quality representations.

Figure 1 shows an example of streaming different quality representations. In this scenario, *Q1* is a high-quality segment, while *Q2* represents low quality parameters (spatial and angular resolution). They are requested according to the available bandwidth. Let us assume that both clients have the same light field display system. In this case, both representations get converted on the cluster nodes to the same spatial and angular resolutions (after which they are no longer sets of discrete images), based on the capabilities and number of the optical engines, respectively. However, the outputs of conversion will differ, according to their inputs; *Q1'* will have a higher visual quality than *Q2'*. Again, the outputs of the converters will be identical in data size regardless of the quality of the inputs, but the video streaming segments, which are to be transmitted over the access network, are different in size. Streaming over time can be performed similarly to conventional 2D streaming, as shown in Figure 2. Yet there are still studies to be carried out regarding the impact of quality switching parameters (number of switching events in a given period, quality level durations, switching frequency etc.) on the QoE, in order for the users to actually benefit from adaptive light field video streaming.

4. Conclusion

The paper presented an overview of the researches we performed on the perceived visual quality of light field displays, and based on our findings, we proposed a novel protocol for adaptive light field streaming. By dynamically switching between different representations of quality (composed of combinations of different spatial and angular resolution values), based on the available bandwidth, the number and duration of interruptions in light field streaming could be decreased. Our current and future works in the topic include the effect of spatial and angular resolution reduction on perceived quality, tolerable rebuffering events and quality switching.

Acknowledgement

The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 643072, Network QoE-Net, and also from the Marie Sklodowska-Curie grant agreement No 676401, Network ETN-FPI.

References

[1] R. R. Tamboli, B. Appina, S. Channappayya, S. Jana, "Supermultiview content with high angular resolution: 3D quality assessment on horizontal-parallax lightfield display," *Signal Processing: Image Communication*, 2016.

[2] V. K. Adhikarla, F. Marton, T. Balogh, E. Gobbetti, "Real-time adaptive content retargeting for live multi-view capture and light field display," *The Visual Computer*, 2015.

[3] HoloVizio 721RC light field display, www.holografika.com/Documents/Holovizio_721RC-emailsizenomovie.pdf (retrieved May 2017).

[4] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, F. Dufaux, P. T. Kovacs, V. K. Adhikarla, "Subjective evaluation of super Multiview compressed contents on high-end light-field 3D displays," *Signal Processing: Image Communication*, 2015.

[5] S. Darukumalli, P. A. Kara, A. Barsi, M. G. Martini, T. Balogh, "Subjective Quality Assessment of Zooming Levels and Image Reconstructions based on Region of Interest for Light Field Displays," *International Conference on 3D Imaging (IC3D)*, 2016.

[6] HoloVizio C80 light field cinema system, www.holografika.com/Documents/HoloVizio_C80.pdf (retrieved May 2017).

[7] M. Levoy, P. Hanrahan, "Light field rendering," *Computer Graphics (SIGGRAPH96)*, 1996.

[8] C.–L. Chang, B. Girod, "Receiver-based rate-distortion optimized interactive streaming for scalable bitstreams of light fields," *IEEE International Conference on Multimedia and Expo (ICME'04),* 2004.

[9] P. Ramanathan, B. Girod, "Random access for compressed light fields using multiple representations," *IEEE International Workshop on Multimedia Signal Processing*, 2004.

[10] G. Cheung, A. Ortega, N.–M. Cheung, "Interactive streaming of stored multiview video using redundant frame structures," *IEEE Transactions on Image Processing*, 2011.

[11] T. Balogh, P. T. Kovacs, "Real-time 3D light field transmission," *SPIE Photonics Europe*, *International Society for Optics and Photonics*, 2010.

[12] P. Ramanathan, M. Kalman, B. Girod, "Rate-distortion optimized interactive light field streaming," *IEEE Transactions on Multimedia*, 2007.

[13] W. Cai, G. Cheung, T. Kwon, S. J. Lee, "Optimized frame structure for interactive light field streaming with cooperative caching," *IEEE International Conference on Multimedia and Expo (ICME'11)*, 2011.

[14] HoloVizio 361P light field display, www.holografika.com/Documents/HoloVizio_360P_emailsize.pdf (retrieved May 2017).

[15] HoloVizio 80WLT light field display, www.holografika.com/Documents/HoloVizio_80WLT-emailsize.pdf (retrieved May 2017).

[16] P. A. Kara, P. T. Kovacs, M. G. Martini, A. Barsi, K. Lackner, T. Balogh, "From a Different Point of View: How the Field of View of Light Field Displays affects the Willingness to Pay and to Use," *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[17] P. A. Kara, P. T. Kovacs, M. G. Martini, A. Barsi, K. Lackner, T. Balogh, "Viva la Resolution: The Perceivable Differences between Image Resolutions for Light Field Displays," *5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS)*, 2016.

[18] ITU-T Rec. "P.910: Subjective video quality assessment methods for multimedia applications," 2008.

[19] P. T. Kovacs, A. Boev, R. Bregovic, A. Gotchev, "Quality measurements of 3D light-field displays," *Eighth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2014.

[20] P. A. Kara, M. G. Martini, P. T. Kovacs, S. Imre, A. Barsi, K. Lackner, T. Balogh, "Perceived Quality of Angular Resolution for Light Field Displays and the Validity of Subjective Assessment," *International Conference on 3D Imaging (IC3D)*, 2016.

[21] P. A. Kara, A. Cserkaszky, S. Darukumalli, A. Barsi, M. G. Martini, "On the Edge of the Seat: Reduced Angular Resolution of a Light Field Cinema with Fixed Observer Positions," *9th International Conference on Quality of Multimedia Experience (QoMEX)*, 2017.

[22] P. A. Kara, A. Cserkaszky, A. Barsi, M. G. Martini, "The Couch, the Sofa, and Everything in between: Discussion on the Use Case Scenarios for Light Field Video Streaming Services," *International Young Researcher Summit on Quality of Experience in Emerging Multimedia Services (QEEMS)*, 2017.

[23] P. A. Kara, P. T. Kovacs, S. Vagharshakyan, M. G. Martini, A. Barsi, T. Balogh, A. Chuchvara, A. Chehaibi, "The Effect of Light Field Reconstruction and Angular Resolution Reduction on the Quality of Experience," *12th International Conference on Signal Image Technology & Internet Based Systems (SITIS) 3rd International Workshop on Quality of Multimedia Services (QUAMUS)*, 2016.

[24] A. Cserkaszky, P. A. Kara, A. Barsi, M. G. Martini, "To Interpolate or not to Interpolate: Subjective Assessment of Interpolation Performance on a Light Field Display," *IEEE International Conference on Multimedia and Expo (ICME) 8th Workshop on Hot Topics in 3D Multimedia (Hot3D)*, 2017.

[25] ITU-R Rec., "BT.500-13: Methodology for the subjective assessment of the quality of television pictures," 2012.

[26] M. G. Michalos, S. P. Kessanidis, S. L. Nalmpantis. "Dynamic adaptive streaming over HTTP," *Journal of Engineering Science and Technology Review*, 2012.

Peter A. Kara received his M.Sc. degree in Computer Engineering from the Budapest University of Technology and Economics in Hungary in 2013. He participated in the CONCERTO project of the European Union's 7th Framework Programme, and he is currently a research associate at the Wireless Multimedia Networking Research Group of Kingston University and a fellow of the H2020 QoE-Net project of the EU. His research interests include multimedia streaming, quality assessment of services, light field technology and cognitive bias in perceived quality.

Aron Cserkaszky received his M.Sc. degree in physics from the Budapest University of Technology and Economics in Hungary in 2011. He is currently a researcher at Holografika Ltd. in the area of full parallax imaging. He previously worked on GPU photon simulations and image reconstructions of PET and SPECT systems.

Attila Barsi received the M.Sc. degree in Informatics Engineering from Budapest University of Technology and Economics in 2004. From 2004 to 2007, he was attending the Ph.D. program of Budapest University of Technology and Economics, researching real-time graphics. Since 2006, he is in employment at the light field display manufacturer company, Holografika Ltd., where he is researching real-time light field capture and rendering. Since 2009, he is employed as the lead software developer for the company. He was working on several EU research projects and is currently participating as a supervisor in the QoE-NET and ETN-FPI European training networks. His research interest includes real-time rendering, light fields, multi-camera systems and video streaming on high speed networks.

Maria G. Martini is (full) Professor in the Faculty of Science, Engineering and Computing in Kingston University, London, where she also leads the Wireless Multimedia Networking Research Group and she is Course Director for the MSc in "Networking and Data Communications". She received the Laurea in electronic engineering (summa cum laude) from the University of Perugia (Italy) in 1998 and the Ph.D. in Electronics and Computer Science from the University of Bologna (Italy) in 2002. She is a Fellow of The Higher Education Academy (HEA). She has led the KU team in a number of national and international research projects, funded by the European Commission (e.g., OPTIMIX, CONCERTO, QoE-NET, Qualinet), UK research councils, UK Technology Strategy Board / InnovateUK, and international industries. Her research interests include wireless multimedia networks, cross-layer design, joint source and channel coding, 2D/3D error resilient video, 2D/3D video quality assessment, and medical applications. She is the author of over 100 international scientific articles and book chapters, and the inventor of several patents on wireless video.

Tibor Balogh CEO and Founder of Holografika, has extensive experience in the field of holography and optical engineering. He graduated as an electric engineer from the Technical University of Budapest and worked for SZKI (one of the major software houses in Hungary). Later, he was an assistant professor at the Eotvos Lorand University. Today he is responsible for the overall management of his company, supervises the work of the research and development team and forms the overall business strategy. He was awarded the Joseph Petzval medal, the Kalmar prize and the Dennis Gabor Award for his work, was World Technology Award finalist in 2006. He has several patents and publications and actively follows the developments of 3D display technologies around the world.

Light Fields for Near-eye Displays

Fu-Chung Huang
NVIDIA, CA, USA
fuchungh@nvidia.com

1. Introduction

The most important requirement to make any near-eye display successful is to provide a comfortable visual experience. This requirement has many boxes to check: having high resolution and wide field of view, being lightweight, having small form factor, and supporting focus cue. Like 3D TVs and movies, near-eye displays also need to solve the vergence and accommodation conflicts. In current Virtual Reality (VR) displays, the user fixates his focus on the fixed focal plane, and the disparity in the pre-processed content drives the eye to verge and creates a 3D sensation. This is particularly challenging for Augmented Reality (AR) as the virtual content needs to match the real-world object at arbitrary depth, and the eye needs to constantly switch its focus between the real and the virtual, and oftentimes causes a fatigue viewing. Using a light field display provides an opportunity to fix the problem; the extended Depth of Field (DoF) enables the virtual content to be displayed at the correct depth, allowing for a comfortable viewing experience. Additionally, unlike the conventional 2D displays, light field displays have the ability to check many boxes of these abovementioned requirements, and we will describe how to navigate the design space that uses light fields.

2. Related Work

A light field $l(\mathbf{x}, \mathbf{u})$ describes the 4-dimensional arrangement and the distribution of light in free space using geometric optics, and is expressed in both space $\mathbf{x} \in R^2$ and angle $\mathbf{u} \in R^2$; a light field display is responsible for bringing such arrangement of light to the eye as if the light field is emitted from the real-world object. Light field can be captured using light field camera or can be rendered and synthesized using Computer Generated Imagery. While there have been many far-field TV-like light field displays, putting it to near-eye creates a new era of research.

There are many different implementations of light field displays, e.g. multilayer display (gaze contingent varifocal displays and multifocal displays), holographic displays, MicroLens Array (MLA)-based light field, and compressive attenuation-based light field. A multilayer-based multi-focal display (Lee et al. [10]) approximates the light field by only presenting the light at discrete locations and depths. The Fourier analysis (Narain et al. [9]) shows that the configuration is equivalent to a sparsely sampled Radon Transform in Computational Tomography. With dense enough spacing, the display is capable of fooling the eye into believing true 3D. On the other hand, a gaze contingent display (Padmanaban et al. [12]) adjusts the optical virtual plane of the display with respect to the user's gaze and focusing. Although the display allows the eye to truly focus onto the correct depth, objects' depth in the periphery is not faithfully preserved. A holographic display emits a high-quality light field truly representing the real-world, but it is expensive to compute and is limited in field of view, the eye box, and eye-relief, making it unsuitable for near-eye display. Both MLA-based and compressive light field offer great accuracy to approximate the near-eye light field, however, there are many constraints making one favorable to the other, and generally trade-offs are made to satisfy certain design needs. In the following sections, we will describe three configurations to approximate the near eye light field.

2.1 MLA-based Near-eye Light Field Display

Integral imaging based on Microlens Arrays (MLA) (Lanman and Luebke [3]) and Parallax Barrier based on pinhole array (Aksit et al. [1]) are effective and simple methods to create a near-eye light field display. Similar to light field camera, this type of light field display trades the spatial resolution for the angular resolution. Taking the integral imaging as an example, each microlens magnifies its underlying elemental image at the virtual plane, as shown in Figure 8. Multiple magnified virtual elemental images coming from many microlenses overlap on top of each other, and the optical setup creates a magnified virtual light field display. Since each point on the virtual image maps to multiple points on different elemental images, several rays are created connecting the pixel on the virtual image to its corresponding pixels on the elemental images. The number of rays determines the degrees of freedom to control the angular variations on the virtual image, and this capability is commonly referred as the depth of field of the light field display (Wetzstein et al. [7]).

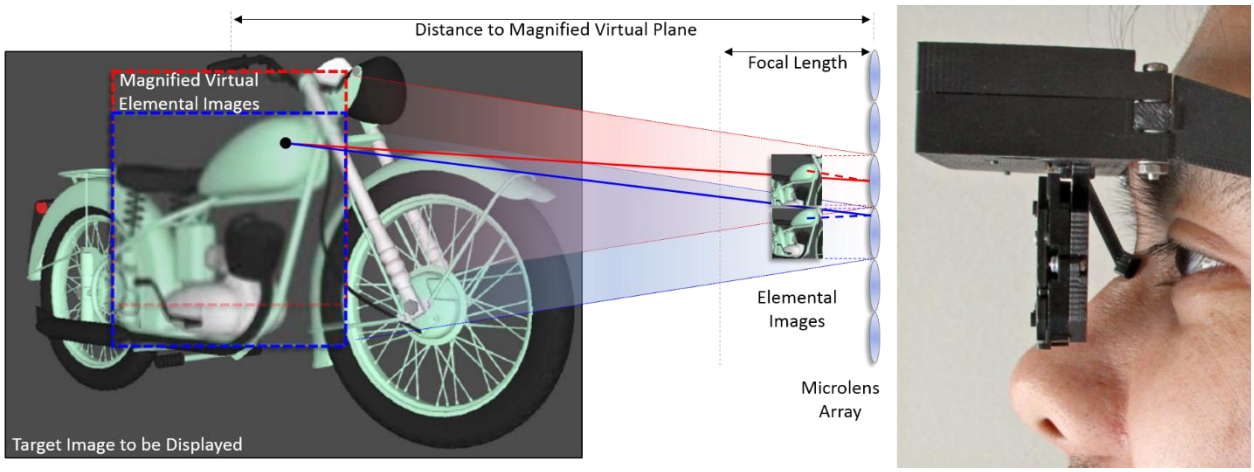


Figure 8: Near-eye light field display using Integral Imaging. (Left): Overlapped virtual elemental images create light fields at the magnified virtual image plane. The extended depth of field from this virtual light field allows for a continuous focus cue and avoids vergence-accommodation conflict. However, significant spatial resolution is traded for angular resolution. (Right): Short focal length of the microlens also allows for a thin and lightweight near-eye display, and the programmable remapping of light rays allows the user to see the near-eye display without wearing another corrective eyeglasses.

Viewing within the depth of field of the display allows for a continuous focus cue and supports accommodation; vergence-accommodation conflict is avoided with this setup. The short focal length of the microlens also enables a thin and lightweight near-eye display.

The second advantage of light-field near-eye displays is to provide a personalized vision correction (Huang et al. [5]). Since the light field allows a programmable remapping of the rays, inversely mapping the individualized aberration to pre-distort the target light field allows the user to wear the near-eye display without additional corrective eyeglasses. All these capabilities are achieved via software rather than optics.

The near-eye light field display, like many light field cameras, has one serious drawback that the display achieves the angular manipulation by sacrificing the spatial resolution to a direction against where the display industry heads to. In the near-eye light field, a 10^2: 1 resolution reduction trade-off is made.

2.2. Pin Light Display

An integral imaging based near eye light field display also makes the application to augmented reality challenging; using a beam splitter expands the form factor significantly to relay the optics for see-through capability. Maimone et al. [4] utilize a defocused diffused pin-point light source modulated by a transmissive LCD to "paint" the content onto the retina; the optical setup is equivalent to have a mini laser scanning projector in front of the eye.

To minimize the light engine, Maimone et al. etch a sparse set of diffusive pin-points on an acrylic glass, as shown in Figure 9 (top-right). The entire glasses is only a few millimeters thick, which is ideal for augmented reality glasses. The display has wide field of view and its projected image is invariant to accommodation. The wide field of view ($\geq 100°$) is easily achieved by tiling the diffusive pin-points painting to a wider extent of the retina. Since each pixel on the retina is painted by only a single ray, the focusing of light and the retinal blur is non-existent: the target image remains sharp in all accommodation states. Although changing the focal length of the eye also changes the refractive power and thus the optical paths, Maimone et al. show that the magnification only changes 3% of the original size.

The pin light display assumes a precise knowledge of the eye location, thus allowing for a very small eye box. The authors propose two solutions to the inconvenience: using eye tracking or light field. Eye tracking has been shown to be practical ($250Hz$) to reduce the computation cost by foveating a high-resolution rendering to the fovea and a low-resolution content to the periphery. A precise eye tracking also helps to reduce the complexity in the optical setup of the near-eye display. When eye tracking is not available, the rendering requires a wider eye box within which the eye can move freely, as shown in Figure 10 (left), and the optical setup needs modification. Again, significant

resolution is sacrificed to enable a near-eye light field with wider eye box.

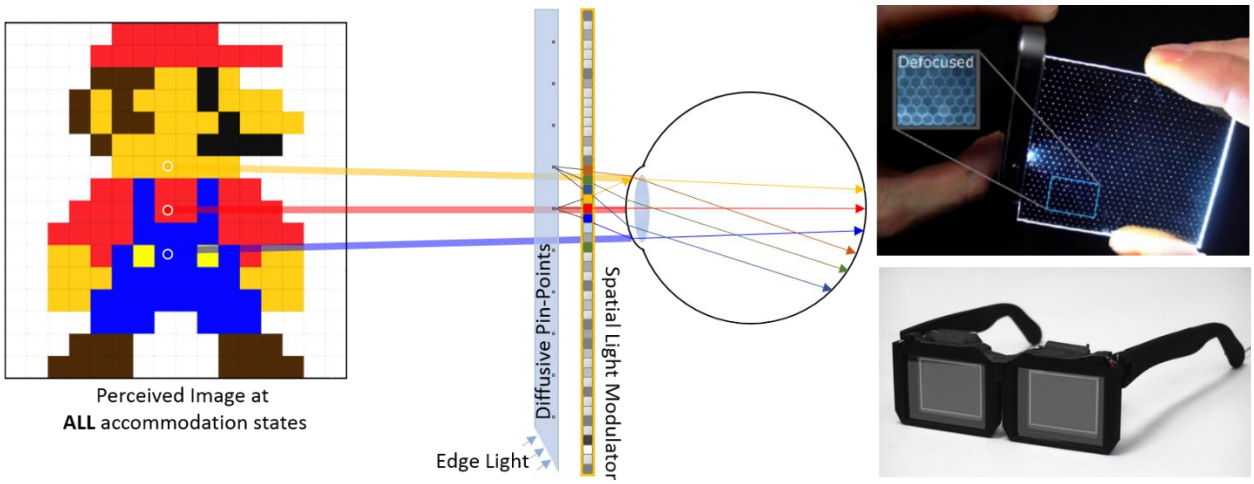


Figure 9: Pin light displays. (Left): The incoming edge light is injected to the wave guide and is diffused by the pin-point and exits toward the eye, seen as a defocused circle (top-right). When the circle passed through a spatial light modulator, e.g. a transmissive LCD, color intensities are attached to the ray and paint the retina. The system allows for thin and lightweight AR glasses (bottom-right). The image is sharp across all focusing distances and is invariant to accommodation (< 3% variations) due to the retinal painting nature.

To support a near-eye light field, many trade-offs need to be made, but the spatial resolution is an indispensable requirement for any display. In the next section, we show another dimension where a large form factor and content fidelity is traded for higher spatial resolution and focus cue to avoid vergence-accommodation conflict.

2.3. Attenuation-based Multilayer Light Field Stereoscope

Only considering the near-eye case, the light field in front of the pupil is highly compressible. The angular variations mainly from the intra-ocular occlusion are critical in monocular depth perception (Zannoli et al. [6]). Optically compressing light field has been shown by Wetzstein et al. [7] and Maimone and Fuchs [11] with stacked layers of attenuating transmissive LCDs that form a multiplicative tensor field. Huang et al. [5] show that two layers of LCDs are sufficient enough to approximate a near-eye light field with only rank-1 reconstruction without temporal multiplexing, which is also critical for near eye displays to reduce the motion blur.

To compress the light field using two-layer optical setup, we consider the optimal reconstruction:

$$\mathbf{argmin}_{\{t_1,t_2\}} = \left\| l(\mathbf{x},\mathbf{u}) - t_1(\mathbf{x}) t_2\left(\mathbf{x} - \frac{\mathbf{u}-\mathbf{x}}{d_e}\right)\right\|^2, \qquad (1)$$

where d_e denotes the eye relief from the eye to the display, and we assume the distance between the layers t_1 and t_2 is 1. Detailed derivation and solution can be found in Huang et al. [5] and Wetzstein et al. [7].

Although the display still requires a pair of magnifying glasses, it also allows for a wide field of view just like any traditional virtual reality stereoscope and supports a large eye box; it requires little modification to the traditional head-mounted display by only adding a second LCD panel. However, the display form factor remains large and the spatial resolution is also subject to the diffraction limit beyond 1080p resolution with 50 μm due to the pixelated structure in the front modulation panel.

Multiplicative multilayer light field displays offer a few advantages over additive multilayer displays (Narain et. al. [9], Lee et. al. [10]). First, intra-ocular occlusion is better preserved, as analyzed by Huang et al. [5], and this monocular intra-ocular presentation is critical to depth perception (Zannoli et al. [6]). Second, additive multilayer displays require temporal multiplexing between layers, and modern displays are not fast enough to support more than three depths. However, additive multilayer displays are not constrained by the diffraction limits found in multiplicative displays. To the manufacturers and content developers, these constraints and trade-offs need to be

considered carefully to allow for a comfortable visual experience. We summarize the trade-offs in the next section.

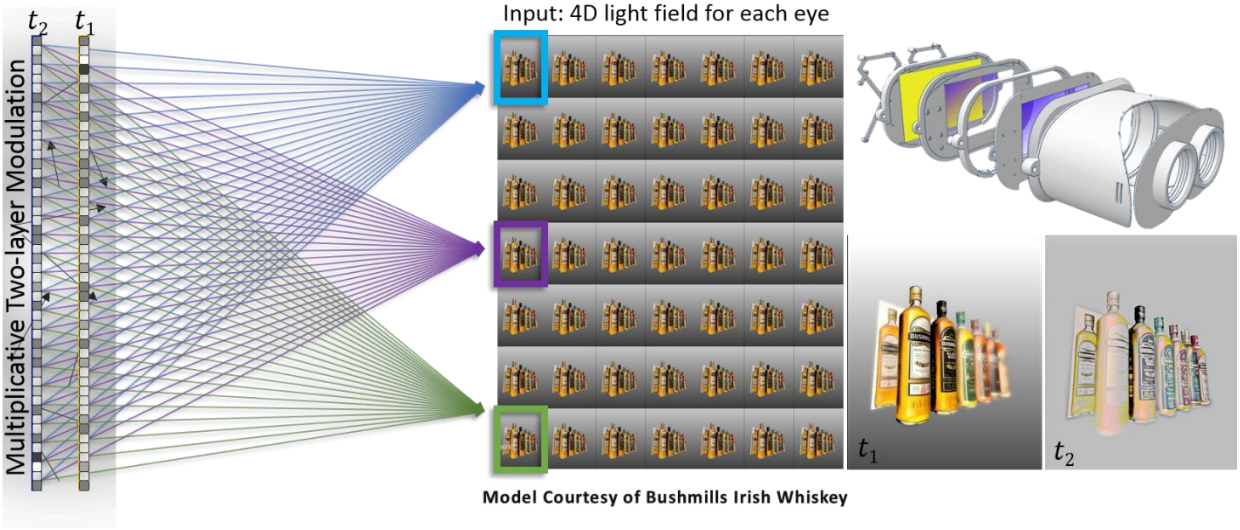


Figure 10: Light Field Stereoscope. Multiplicative two-layer transmissive LCD displays generate a compressed light field. Each shifted multiplication in perspective reconstructs a given input view across the eye box. The optimized transmissive layers are calculated and shown in bottom right. Although the display reconstructs a well-approximated light field, the hardware does not support a lightweight and thin form factor, as shown on the top right.

3. Design Constraints and Trade-off Analysis

Near-eye displays are typically constrained by many factors and usually trade some features for others. In Table 1, we compare many different types of near eye displays with different levels of light field approximation. Specifically, we compare their spatial resolution, Field of View (FoV), eye box, Depth of Field (DoF), requirement for eye-tracking, exactness of 3D representation, and display form factor. We note that making display form factor thin and lightweight is a fundamental challenge that typically requires great sacrifices in other dimensions, and among which the spatial resolution is the most important to preserve in the trade-off space. An extended DoF enables continuous focus cue and allows the eye to accommodate to the desired depth, avoiding the vergence-accommodation conflict problem and improve the visual experience.

	Resolution	FoV	Eye Box	DoF	Eye Tracking	3D	Form
Traditional 2D	Very High	Wide	Wide	No	No	No	Large
Integral Imaging Light	Low	Narrow	Moderate	Yes	No	Real	Thin
Pin Light	Low	Wide	Small	No	Yes	No	Thin
Multiplicative Multilayer	High	Wide	Moderate	Yes	No	Approx.	Large
Additive Multilayer	Very High	Wide	Small	Moderate	Yes	2.5D	Large

Table 1: Design constraints and trade-off analysis

4. Conclusion

In this paper, we show a few light field implementations for near eye displays. In particular, two of the described methods sacrifice the resolution for a lightweight and thin form factor, and a method exploits the compressive nature of near-eye light field to provide extended depth of field.

There are still emerging technologies like Computer Generated Hologram or holographic light field displays in the horizon and this could potentially break more design constraints like the resolution limit, form factor, and provide

extended depth of field by making diffraction our friend. To conclude, it is an exciting time for near-eye displays and also to witness many different technologies converging together to solve hard problems.

Acknowledgement

This research review is based on the results from many collaborators in NVIDIA Research: Douglas Lanman (now at Oculus Research), Andrew Maimone (now at Microsoft Research), David Luebke, and Gordon Wetzstein from Stanford University.

References

[1] Kaan Aksit, Jan Kautz, and David Luebke, 2015. Slim near eye display using pinhole aperture arrays. Applied Optics, March, 2015

[2] Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. 2015. The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues. ACM Trans. Graph 34, 4, Article 60 (July 2015), 12 pages.

[3] Douglas Lanman and David Luebke. 2013. Near-eye light field displays. ACM Trans. Graph. 32, 6, Article 220 (November 2013), 10 pages.

[4] Andrew Maimone, Douglas Lanman, Kishore Rathinavel, Kurtis Keller, David Luebke, and Henry Fuchs. 2014. Pinlight displays: wide field of view augmented reality eyeglasses using defocused point light sources. ACM Trans. Graph. 33, 4, Article 89 (July 2014), 11 pages.

[5] Fu-Chung Huang, Gordon Wetzstein, Brian A. Barsky, and Ramesh Raskar. 2014. Eyeglasses-free display: towards correcting visual aberrations with computational light field displays. ACM Trans. Graph. 33, 4, Article 59 (July 2014), 12 pages.

[6] Marina Zannoli, Gordon D. Love, Rahul Narain, and Martin S Banks, 2016. Blur and the perception of depth at occlusions. Journal of Vision 16(6):17 (April 2016).

[7] Gordon Wetzstein, Douglas Lanman, Matthew Hirsch, and Ramesh Raskar. 2012. Tensor displays: compressive light field synthesis using multilayer displays with directional backlighting. ACM Trans. Graph31, 4, Article 80 (July 2012), 11 pages.

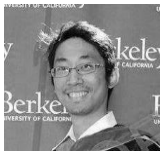
[8] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards foveated rendering for gaze-tracked virtual reality. ACM Trans. Graph. 35, 6, Article 179 (November 2016), 12 pages.

[9] Rahul Narain, Rachel A. Albert, Abdullah Bulbul, Gregory J. Ward, Martin S. Banks, and James F. O'Brien. 2015. Optimal presentation of imagery with focus cues on multi-plane displays. ACM Trans. Graph. 34, 4, Article 59 (July 2015), 12 pages.

[10] Seungjae Lee, Changwon Jang, Seokil Moon, Jaebum Cho, and Byoungho Lee. 2016. Additive light field displays: realization of augmented reality with holographic optical elements. ACM Trans. Graph 35, 4, Article 60 (July 2016), 13 pages.

[11] Andrew Maimone and Henry Fuchs, 2013. Computational augmented reality eyeglasses. *ISMAR* (2013).

[12] Nitish Padmanaban, Robert Konrad, Tal Stramer, Emily A. Cooper, and Gordon Wetzstein. 2017. Optimizing virtual reality for all users through gaze-contigent and adaptive focus displays. PNAS. 2017 144(9) 2183-2188.



Fu-Chung Huang is Scientist at NVIDIA Research. Before joining NVIDIA, he was a visiting scientist at Stanford University. He obtained his Ph.D. degree in Computer Sciences from University of California at Berkeley, CA, USA in 2013, and was also a visiting research student at MIT Media Lab. His work focuses on applying computation and human perception to display technology and optics.

MMTC OFFICERS (Term 2016 — 2018)

CHAIR

Shiwen Mao
Auburn University
USA

STEERING COMMITTEE CHAIR

Zhu Li
University of Missouri
USA

VICE CHAIRS

Sanjeev Mehrotra (North America)
Microsoft
USA

Fen Hou (Asia)
University of Macau
China

Christian Timmerer (Europe)
Alpen-Adria-Universität Klagenfurt
Austria

Honggang Wang (Letters&Member Communications)
UMass Dartmouth
USA

SECRETARY

Wanqing Li
University of Wollongong
Australia

STANDARDS LIAISON

Liang Zhou
Nanjing Univ. of Posts & Telecommunications
China

MMTC Communication-Frontier BOARD MEMBERS (Term 2016—2018)

Guosen Yue	Director	Huawei R&D USA	USA
Danda Rawat	Co-Director	Howard University	USA
Hantao Liu	Co-Director	Cardiff University	UK
Dalei Wu	Co-Director	University of Tennessee	USA
Zheng Chang	Editor	University of Jyväskylä	Finland
Lei Chen	Editor	Georgia Southern University	USA
Tasos Dagiuklas	Editor	London South Bank University	UK
Melike Erol-Kantarci	Editor	University of Ottawa	Canada
Kejie Lu	Editor	University of Puerto Rico at Mayagüez	Puerto Rico
Nathalie Mitton	Editor	Inria Lille-Nord Europe	France
Shaoen Wu	Editor	Ball State University	USA
Kan Zheng	Editor	Beijing University of Posts & Telecommunications	China