Computer Communications 131 (2018) 77–82

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Reprint of: From cloud-based communications to cognition-based communications: A computing perspective

Min Chen[a,b,*], Victor C.M. Leung[c,*]

[a] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[b] Wuhan National Laboratory for Optoelectronics, Wuhan, China
[c] Department of Electrical and Computer Engineering, the University of British Columbia, Vancouver, Canada

ARTICLE INFO

Keywords:
Cognitive computing
Cognitive communications
Cloud-based communications
Edge computing

ABSTRACT

Traditional cloud-based communications provide powerful cloud computing services. However, simply supporting intensive data processing is not sufficient, especially when capacity is limited and ultra-low latency is required. Thus, it is critical to propose a new Artificial Intelligence (AI)-enabled heterogeneous networks, including various terminal networks, fogs and clouds. Derived from cognitive science and data analytics, cognitive computing can mimic or augment human intelligence. When such cognitive intelligence is integrated with communications, traditional services will be renovated with higher accuracy and lower latency. In this paper, we propose cognition-based communications, which originates from both AI-based intelligent computing and advances in communications. Then, we introduce two applications of cognition-based communications, including user-centric cognitive communications, and cognitive internet of vehicles. Through cognition-based communications, we can better meet users' needs, provide them with a better Quality of Experience (QoE), and achieve a higher energy efficiency.

1. Introduction

Novel information services and applications are expanding globally with the rapid development of wireless communication and networking technologies. Advanced networks and communications can greatly enhance users' experience and have made a huge impact in all aspects of people's lifestyles at home, at work, in social exchanges, and economically. Although these advanced techniques have extensively improved users' Quality of Experience (QoE) [1], they are not adequate to meet various requirements such as seamless wide-area coverage, high-capacity hot-spot, low-power massive-connections, low latency high-reliability, and other challenging scenarios. Therefore, it is critical to develop smart wireless communication and networking technologies to support optimized management, dynamic configuration, and fast service composition. Recent year have witnessed that the fusion of computing and communications exhibits a trend to reach such a goal. Cognitive computing, which is derived from cognitive science and data analytics, can mimic or augment human intelligence [2]. In addition, cognitive computing exhibits great potentials to power smart wireless communications, e.g., in self-driving. An intelligent network can be viewed as an existing network integrated with cognitive and cooperative mechanisms to promote performance and achieve intelligence. Under the new service paradigm, there are various technical challenges and problems that need to be addressed to extensively improve the user's QoE, such as complicated decision making for routing, dynamic and context-aware network management, resource optimization, and in-depth knowledge discovery in complex environments. Artificial Intelligence (AI) plays an important role in the ability of cognitive wireless communications to meet many of these technical challenges. Furthermore, wireless communication and network ecosystems must be upgraded with new capabilities, such as the provisioning of personalized and smart Fifth Generation (5G) network services that are assisted by data cognitive intelligence, advanced wireless signal processing based on deep learning, optimized wireless communication physical layer design based on reinforcement learning, adaptive wireless resource management based on cognitive intelligence, etc.

Though various previous works used the similar terminology of "cognitive communications", they focus on the research related with cognitive radio. For example, Green Cognitive Communications in [3], and Cognitive Device-to-Device (D2D) Communications in [4]. In order to avoid the ambiguity, our proposed architecture is named as "cognition-based communications". In this paper, we first introduce cloud-based communications [6], including Cloud Radio Access Network (C-RAN) [7] and Name Data Network (NDN). Then, we present the architecture

[*] Corresponding authors.
E-mail addresses: minchen2012@hust.edu.cn (M. Chen), vleung@ece.ubc.ca (V.C.M. Leung).

https://doi.org/10.1016/j.comcom.2018.09.006
Available online 4 October 2018

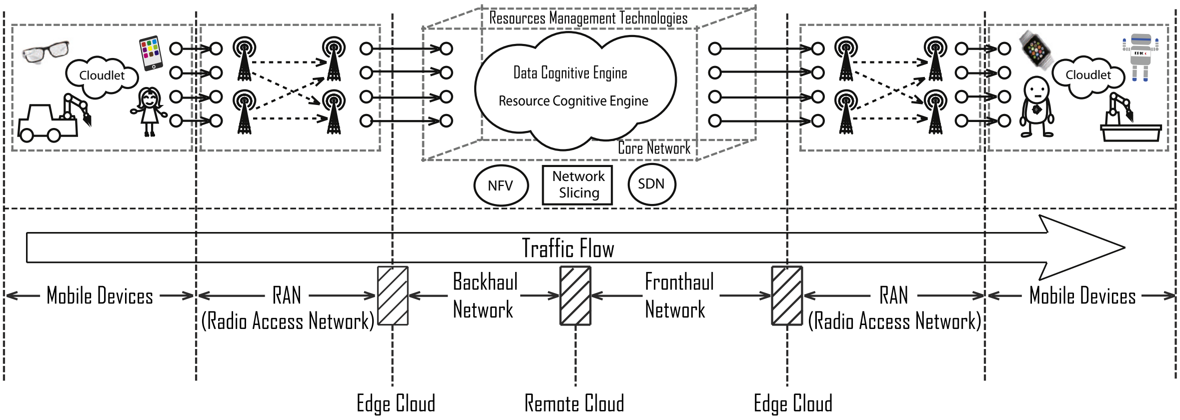


Fig. 1. Cloud-based communications.

and applications of cognition-based communications. In summary, the main contributions of this paper include:

1) We introduce two representative paradigms of cloud-based communications, i.e., computing-centric C-RAN, and cache-enabled NDN.
2) We propose cognition-based communications with a new architecture that includes two layers, i.e., the communication layer and cognition layer. Specifically, the cognition layer consists of two core cognitive engines, i.e., the resource cognitive engine and data cognitive engine.
3) The benefits of applying the proposed cognition-based communications are illustrated by two archetypal wireless networking applications, i.e., user-centric cognitive communications and cognitive internet of vehicles [5].

The remainder of this paper is organized as follows: Section 2 presents the architecture of conventional cloud-based communications, and give its pros and cons. In Section 3, the evolution from cloud-based communications to cognition-based communications is identified from a computer communications' point of view. Section 4 concludes this paper.

2. Cloud-based communications

Cloud-based communications include computation-oriented C-RAN and cache-centric NDN. As shown in Fig. 1, C-RANs offload communication-related computing onto the cloud for centralized processing and management. In NDNs, to decrease the delay of content retrieval, contents are cached in various network nodes in a distributed fashion [2].

2.1. Cloud radio-access networks

Radio Access Networks (RANs) are the foundation of today's mobile cellular networks. Traditional RANs built in second to fourth generation mobile core networks are characterized with the following features: (1) Many Base Stations (BS) are deployed with fixed sector antennas. Each BS covers a small area and only handles phone signals in its coverage range. (2) The service capacity is limited by interference and handover losses. (3) The BSs are built with high infrastructure costs. These features have resulted in low spectral efficiency, non-negligible air loss, limited Quality of Service (QoS), as well a high Capital Expenditure (CAPEX) and Operational Expenditure (OPEX). To address these problems, C-RAN was suggested to upgrade or replace current generation RANs in building 5G or future mobile networks. Furthermore, the C-RAN approach is used to reduce the total cost of CAPEX and OPEX, in order to achieve lower energy consumption and higher spectral efficiency. In C-RAN, the bulky antenna towers used in conventional BS are replaced by many small Remote Radio Heads (RRH). These RRHs operate with little power (solar energy is sufficient) and are easily distributed at a high density in populated user areas. The control and processing in the physical BSs are replaced by using Virtual Base Station (VBS) pools housed in a hierarchy of cloud-based switching centers. A balanced traffic load between the RRHs and the VBS pools is enabled by using a high-speed optical transport network and switches with fiber cable and microwave links. The advantages of using C-RAN are summarized in four aspects: (1) A centralized processing resource pool can support between 10 and 1000 cells with high efficiency. (2) Collaborative communications are used in multi-cell joint scheduling and processing, which solves the air loss and handover problems. (3) C-RAN offers real-time services by targeting the open Information Technology (IT) platform, resource consolidation, and flexible multi-standard operation and migration. And (4) a green-energy and clear mobile telecommunication is realized with much less power consumption, lower operating expenses, and a fast system roll out. Many companies are building C-RAN systems including CISCO and Korean Telecommunication.

2.2. Cache-enabled named data networking

To overcome the limited capacity of backhaul links, paradigms called NDN, "Content-Centric Networking" (CCN), and "Information-Centric Networking" (ICN) have been proposed to handle content-dominated Internet traffic for the RANs and the core networks. The problem of content caching is generally divided into two steps: content placement and content delivery. Content placement includes determining what content to be cached and where to cache it [8]. Content placement also considers how to download the content to the cache node. Content delivery is concerned with how the content is transferred to the requesting users. In general, when the network traffic is low, network resources are cheap and rich (e.g., in the early morning or midnight), and content placement should be conducted. Accordingly, when the network traffic is high, network resources are expensive and scarce, and content delivery should be conducted. Consider a heterogeneous RAN, which consists of a Macro-cell Base Station (MBS), Small cell Base Stations (SBS), and user terminals. We can summarize four basic caching placement strategies: local caching, D2D caching, Small cell Base Station (SBS) caching, and Macro-cell Base Station (MBS) caching. The mobility of users is an important influencing factor for the cache of a wireless access network. By utilizing the mobility of users, the hit rate of content cache can be increased. For caching in the core network, the content can be cached on the routers. The main issues of cache management are content placement and content replacement. For CCNs, content caching is an important research direction. It has found that popular content is often requested (such as popular videos). Therefore, the first important concern for content caching is the popularity of the cache file. The other important factor of content caching is the user's preference for specific content. The user preference for content can be predicted through recommendation algorithms such as collaborative filtering, or through the user's historical requested data.

3. Cognition-based communications

In this section, we will first introduce how to leverage cognition for the optimization of communications in the wireless network. Then, we give the architecture of cognitive-based communications. Finally, we provide two applications of cognition-based communications.

3.1. When AI and cognitive computing meet communications and networking

5G wireless communications are expected to satisfy the diverse service requirements in various aspects of our daily life. Thus, the design and optimization of 5G networks become very challenging. The future 5G network will require robust intelligent algorithms to adapt to network protocols and the management of resources that are required for different services in diverse scenarios. Recent advances in deep learning, convolutional neural networks and reinforcement learning hold significant promise for solving very complex problems that have been considered intractable until now.

It is now appropriate to apply AI technology to 5G wireless communications to tackle optimized physical layer design, complicated decision making, network management, and resource optimization tasks in such networks [9]. Moreover, emerging big data technology has brought us an excellent opportunity to study the essential characteristics of wireless networks, and help us to obtain more clear and in-depth knowledge of the behavior of 5G wireless networks.

AI approaches, well known from the IT disciplines, are beginning to emerge in the networking domain. These approaches can be clustered into AI techniques for network management, network design for AI applications, and system aspects. AI techniques for network management, operations, and automation address the design and application of AI techniques to improve the way we address networking today. Networking has recently become the focus of a huge transformation enabled by new models that have resulted from virtualization and cloud computing. This has led to several novel architectures supported by emerging technologies such as Software-Defined Networking (SDN), Network Function Virtualization (NFV), and more recently edge cloud and fog computing. This development towards enhanced design opportunities, along with increased complexity in networking as well as in networked applications, has increased the need for improved network automation in agile infrastructures. As exemplified by recent initiatives to set-up network automation platforms, this new networking environment calls for even more automation. This can be combined with AI techniques to execute efficient, rapid, and trustworthy management operations. Network design and optimization for AI applications also addresses a complementing topic: the support of AI-based systems through novel networking techniques, including new architectures as well as new performance models.

Cognitive intelligence can be used for optimizing wireless communication technologies, wireless transmission technologies, and wireless system applications and services. Wireless communication technologies include channel modeling, channel state estimation, beamforming, code book design, and signal processing. Wireless transmission technologies include coordinated multiple points transmission, large scale antenna array, and multi-hop relay. Wireless system applications and services include mobility management (e.g., user association and handoff strategy), resource management (e.g., spectrum resource, energy resource, computing resource, and communication resource), multi-media traffic load, network overhead, and network collision.

Cognitive computing can be used for network analytics, network applications, and network automation. Network analytics include cognitive analytics in networking and network problem diagnosis through machine learning. Network application includes resource allocation for virtualized networks using machine learnings, energy-efficient network operations via machine learning algorithms, and machine learning algorithms for network security. Network automation includes deep learning and reinforcement learning in network control and management, and predictive and self-aware networking maintenance.

3.2. Architecture of cognition-based communications

Fig. 2 shows the system architecture of cognition-based communications, which include two layers, i.e., the communication layer and cognition layer. The communication layer mainly consists of various terminal devices (e.g., mobile devices, vehicles, smart phones, etc.) and wireless access points (e.g., base station, etc.). The cognition layer mainly consists of two core cognitive engines, i.e., the resource cognitive engine and data cognitive engine. The interaction between data cognitive engine and resource cognitive engine is the key design issue which is also shown at the top of Fig. 2. The cognition layer mainly processes service data and network data.

Resource cognitive engine: This engine can learn the characteristics of computing resources, communication resources, and network resources (e.g., network type, service data flow, communication quality, and other dynamic environmental parameters, etc.) by cognitive computing. Then, the integrated resource data can be fed back to the data cognitive engine in real time. In addition, the resource cognitive engine can receive the analysis result of the data cognitive engine and realize the real-time dynamic optimization and distribution of resources. As shown in Fig. 2, it mainly includes the resource data pool, network softwarization technologies and resource management technologies. More specifically, the resource cognitive engine includes the following: (1) Resource Data Pool: realize the massive, heterogeneous and real-time connections between terminals (such as smart clothing, intelligent robot, intelligent traffic car, and other cognitive devices), ensure the security, reliability and interoperability of connections, and constitute the resource data pool as a basic architecture for data transmissions. (2) Network Softwarization: utilize network softwarization technologies involving NFV, SDN, self-organized network (SON), and network slicing to realize high reliability and flexibility, ultra-low latency, and extendibility of the edge cognitive system. (3) Resource Management: utilize the resource-management technologies involving compulation off-loading, handover strategy, caching and delivery, and intelligent algorithms to build a cognitive engine with resource optimization and energy saving to enhance QoE and meet the different demands of various heterogeneous applications.

Data cognitive engine: This engine deals with real-time data flows in the network environment and introduces data analysis and automatic service processing capabilities to the edge network. Furthermore, it realizes cognition to the service data and resource data using various cognitive computing methods including data mining, machine learning, deep learning, and AI. The data cognitive engine guides the network resource distribution dynamically and provides the cognitive services. The main data sources are obtained as follows: (1) it can collect external data from the data sources in the application environment, such as physical signs, real-time disease risk level under cognitive health surveillance, or real-time behavior information on mobile user; and (2), it can collect dynamically the internal data on computing resources, communication resources, and network resources of the edge cloud (network type, service data flow, communication quality, and other dynamic environmental parameters).

The interaction between data cognitive engine and resource cognitive engine: The key design issue of cognition-based communications is the interaction between the data cognitive engine and resource

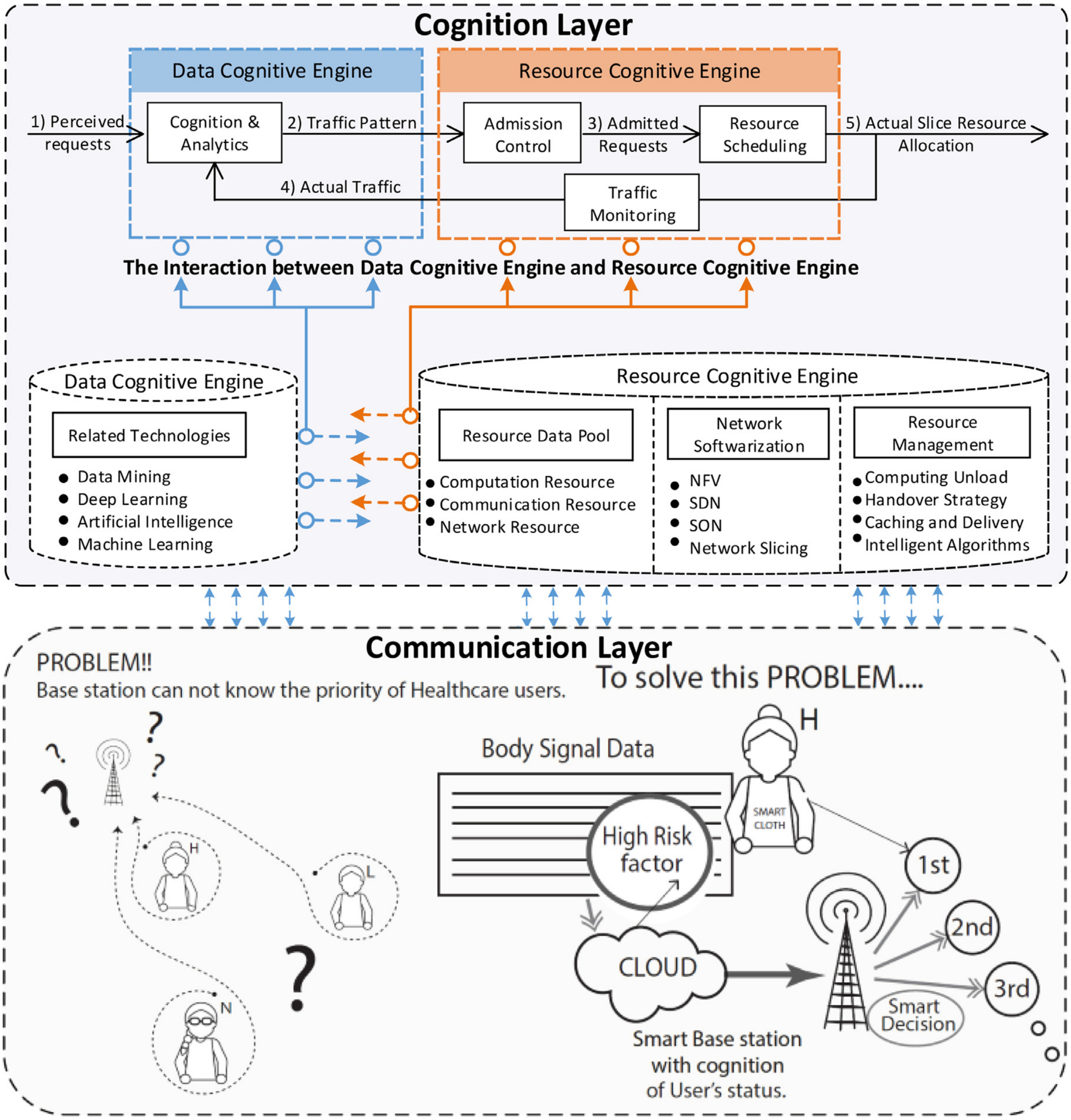


Fig. 2. The architecture of cognition-based communications.

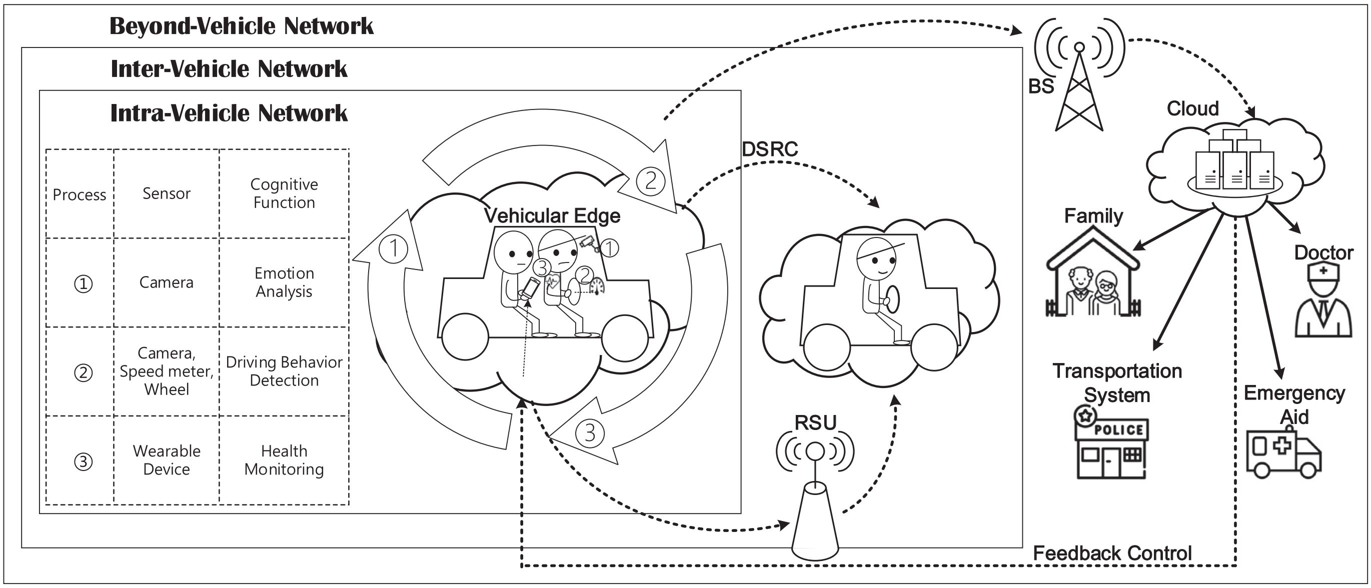
cognitive engine. Here we take cognitive network slicing as an example to illustrate how to fuse the related technologies in cognitive computing and edge computing. As shown in Fig. 2, the data cognitive engine first perceives many requests. The request types of the network-slice service differ from one another according to different demands (latency, reliability and flexibility) of various cognitive applications. Then the data cognitive engine conducts fusion cognitive analysis of the heterogeneous data based on the current resource distribution situation and real-time requests of the tenant with methods of machine learning and deep learning. Next the data cognitive engine reports the analyzed dynamic traffic patterns to the resource cognitive engine. In the resource cognitive engine, there is a joint optimization of the comprehensive benefits and the resource efficiency. Firstly, it conducts admission control to the perceived requests, then conducts dynamic resource scheduling and distribution based on the cognition of network resources, and feeds the scheduling results back to the data cognitive engine, to realize the cognition of the network-slice resources.

3.3. *Applications of cognition-based communications*

Where does the cognition come from? First, based on the data cognitive engine, a user's immediate situation can be obtained. By the cognition of user's status, intelligent communication services can be achieved. In the following section, we give an example of this type of cognition-based communications, i.e., user-centric cognitive communications. Second, through the interactions between the data cognitive engine and resource cognitive engine, cognitive intelligence is realized and embedded into a communication system. As a typical application of this type of cognition-based communications, the design of a cognitive internet of vehicles is also introduced in this section.

3.3.1. *User-centric cognitive communications*

User-centric cognitive communications include two elements, i.e., cognition of users' situations and user-centric resource allocation.



Process	Sensor	Cognitive Function
①	Camera	Emotion Analysis
②	Camera, Speed meter, Wheel	Driving Behavior Detection
③	Wearable Device	Health Monitoring

Fig. 3. Cognitive internet of vehicles: a scenario of driver's health monitoring and emotion care.

- Cognition of a user's situation: Fig. 2 shows an example of a healthcare application. The user's information (i.e., body signals) is first collected. By the use of AI algorithms deployed in the cloud, the user is classified into one of three categories, including a high risk user with the highest priority, a low risk user with medium priority, and a healthy user with the lowest priority.
- User-centric resource allocation: with the cognition of a user's situation, BS can apply some AI algorithm for resource allocation. For example, the user in Fig. 2 has a high risk factor, so BS assigns resources to her with the highest priority. Since the user's life is endangered, her doctor may immediately have video conferencing with her for remote diagnosis, which requires ultra-low communication delay with enough resources in terms of computing and bandwidth.

3.3.2. *Cognitive internet of vehicles*

In [5], an innovative paradigm called Cognitive Internet of Vehicles (CIoV), is proposed. CIoV enhances transportation safety and network security by mining effective information from both physical and network data spaces. As shown in Fig. 3, we explain the mobile cognitive application based on multi-intelligent device interactions with the example of mobile health surveillance. The health status of the driver not only influences his/her own safety, but also influences the safety of passengers in the vehicle, safety of other drivers, and even the traffic system security. In the case of poor health or fatigue of the driver, the attention of the driver may be significantly reduced, and the response time is increased, which may often result in the occurrence of traffic accidents. Therefore, it is very important to monitor the physical health of the driver during the driving process. Under the traditional driving environment, the passenger and driver fail to understand the mutual healthy conditions before. Due to the weakened state of consciousness, the tired driver may even fail to know his/her own status but select to drive continuously, which may greatly threaten the safety of the personnel in the vehicle and other road-users outside the vehicle. To remedy such a situation, the cognitive intra-vehicle network carries out the emotion analysis, driving behavior surveillance, and physical health surveillance. The camera of the intra-vehicle network can entrust the facial expression data of the driver to the vehicle-mounted edge device for analysis. As for the driving behavior detection, the camera detects the eyelid state and micro-nod of the driver, to discover the micro-sleep behavior effectively, analyze in combination with the data collected by the devices such as steering wheel and intelligent odometer embedding in the sensor, remind and give an early warning to the driver, and prevent the occurrence of traffic accidents. In addition, the healthy and physiological index data of each passenger and driver can be collected by the smart clothing and other wearable devices, and uploaded to the vehicle mounted edge for real-time analysis. D2D can be applied in this situation. The vehicle-mounted edge assesses the health status of each user by the data cognitive engine, and reports the analysis result to user's smart phone. Users in the same intra-vehicle network can select the visible window sharing the health status. If the driver suddenly feels unwell (such as a burst of acute disease) during the driving activity, the vehicle-mounted edge will perceive the critically ill condition of the driver in a timely manner from the data collected by the smart clothing, activate the safety automatic driving mode, and give an alarm to the nearby vehicles and cloud. The cloud will dispatch more resources (communication resources of cellular mobile network, and computing resources of remote data center, nearby vehicles and roadside access points) to carry out deeper and more comprehensive condition analysis for the ill driver. At the same time, the cloud rapidly contacts the ambulance, doctor and driver's home. The analysis result will also be delivered to the doctor, so as to make the diagnostic analysis for patients by the time of the ambulance is on the way and enhance the survival rate of the ill driver.

4. Conclusion

In this paper, we have summarized the status and characteristics of the traditional cloud-based communications. Then, by combining traditional cloud-based communications with current popular cognitive computing technology, we have analyzed the feasibility of evolution to cognition-based communications in wireless communication networks. We have proposed the architecture of cognition-based communications, and described the two important modules, i.e., resource cognitive engine and data cognitive engine. They aim to better serve users, provide improved user experience, dynamically allocate communication resources, save computing resources, and achieve energy efficiency. Finally, we have given two representative applications of cognition-based communications, i.e., user-centric cognitive communications, and cognitive internet of vehicles.

Acknowledgement

This work was supported by the National Key R&D Program of China (2018YFC1314600), the National Natural Science Foundation of China (under Grant No. U1705261, 61671088, 61572220), Director Fund of WNLO, the Fundamental Research Funds for the Central Universities (HUST: 2018KFYXKJC045), and the Canadian Natural Sciences and Engineering Research Council.

References

[1] L. Zhou, QoE-driven delay announcement for cloud mobile media, IEEE Trans. Circuits Sys. Video Tech. 27 (1, 2017) (2017) 84–94.

[2] K. Hwang, M. Chen, Big Data Analytics for Cloud/IoT and Cognitive Computing, Wiley, U.K., 2017. ISBN: 9781119247029.

[3] P. Karunakaran, W. Gerstacker, Sensing algorithms and protocol for simultaneous sensing and reception-based cognitive D2D communications in LTE-a systems, IEEE Trans. Cogn. Commun. Netw. 4 (1) (2018) 93–107.

[4] Y. Luo, L. Gao, J. Huang, An integrated spectrum and informati on market for green cognitive communications, IEEE J. Sel. Areas Commun. 34 (12) (2016) 3326–3338.

[5] M. Chen, Y. Tian, G. Fortino, J. Zhang, I. Humar, Cognitive internet of vehicles, Comput. Commun. 120 (2018) 58–70.

[6] T. Higuchi, F. Dressler, O. Altintas, How to keep a vehicular micro cloud intact, in: Proceeding of 87th IEEE Vehicular Technology Conference (VTC Spring 2018), Porto, Portugal, 2018.

[7] K. Sundaresan, M.Y. Arslan, S. Singh, S. Rangarajan, S.V. Krishnamurthy, Fluidnet: a fexible cloud-based radio access network for small cells, IEEE/ACM Trans. Netw. 24 (2) (2016) 915–928.

[8] M. Chen, Y. Hao, M. Qiu, J. Song, D. Wu, I. Humar, Mobility-aware caching and computation offloading in 5G ultradense cellular networks, Sensors 16 (7) (2016) 974–987.

[9] M. Chen, Y. Hao, Task offloading for mobile edge computing in software defined ultra-dense network, IEEE J. Sel. Areas Commun. 36 (3) (2018) 587–597.