中国科学: 信息科学 2018 年 第 48 卷 第 8 期: 1076–1082
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
观点与争鸣

数字视网膜: 智慧城市系统演进的关键环节

高文[1], 田永鸿[1*], 王坚[2]

1. 北京大学信息科学技术学院, 北京 100871
2. 阿里巴巴集团, 杭州 311121
* 通信作者. E-mail: yhtian@pku.edu.cn

收稿日期: 2018–01–31; 接受日期: 2018–03–03; 网络出版日期: 2018–05–21

国家重点研发计划 “云计算与大数据” 重点专项 (批准号: 2017YFB1002400)、国家重点基础研究发展计划 (973) (批准号: 2015CB351800) 和国家自然科学基金大数据科学中心项目 (批准号: U1611461) 资助

摘要 本文阐述了作者对智慧城市建设和发展的主要观点: (1) 如何实时聚合各类城市大数据, 特别是来自视频监控网络的图像视频数据, 并通过构建基于云计算的 “城市大脑” 来分析和挖掘大数据价值并服务于城市运营与管理, 是智慧城市发展中亟待解决的一个关键问题. (2) 现阶段智慧城市建设的现状是 “有眼、有脑”, 但作为 “眼睛” 的摄像头功能过于单一使得 “脑强眼弱”, 其根源在于传统监控摄像机网络所采用的技术体系是为存储而不是分析设计的. 尽管近期有些智能摄像头具有车牌或人脸识别功能, 但是这种单纯强调 “边缘计算” 的方案仍然无法解决 “眼脑合一” 的问题. (3) 为了解决目前阻碍智慧城市系统功能快速演进的难题, 我们应借鉴人类进化了数十万年的视觉系统之 “人类视网膜同时具有影像编码与特征编码功能” 这一特性, 研究与设计数字视网膜, 使之具有统一时间戳和精确地理位置, 能同时进行高效视频编码和紧凑特征表达的联合优化, 并有效支持云端大规模监控视频分析与快速视觉搜索等功能. (4) 为利用数字视网膜来构筑智慧城市的 “慧眼”, 应积极布局与推进相关标准制定、芯片与硬件实现、支撑软件开发与软硬件开源社区, 并开展大规模测试与应用.

关键词 智慧城市, 城市大脑, 数字视网膜

1 现有智慧城市系统存在的问题

智慧城市是把云计算、大数据和人工智能等信息技术应用在城市管理系统中, 通过对各类城市大数据的有效聚合、分析与挖掘, 实现信息化、智能化与城镇化深度融合, 从而有助于缓解 “大城市病”, 实现城市的精细化运营和动态管理. 因此, 如何集合各类城市大数据, 充分分析和挖掘大数据价值, 是智慧城市发展中亟待解决的一个关键问题.

引用格式: 高文, 田永鸿, 王坚. 数字视网膜: 智慧城市系统演进的关键环节. 中国科学: 信息科学, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025
Gao W, Tian Y H, Wang J. Digital retina: revolutionizing camera systems for the smart city (in Chinese). Sci Sin Inform, 2018, 48: 1076–1082, doi: 10.1360/N112018-00025

自 2017 年起, 阿里云在杭州开始建设城市大脑. 作为智慧城市的数据处理和决策中枢, 城市大脑是利用丰富的城市数据资源 (特别是来自视频监控网络的大规模图像视频数据) 和互联网技术来推动经济社会发展和完善社会治理的前瞻性实践. 城市大脑的一个初步应用范例是用于解决城市交通问题: 通过监控摄像头实时监测路面的车流量, 计算路面的车辆占有率, 判断每条道路的饱和程度, 在此基础上对交通信号灯进行优化, 从而在没有对城市路网进行调整的情况下将杭州市部分区域的车辆速度平均提升了 15%.

目前, 阿里城市大脑已经实时接入处理了杭州城区的几百路监控视频. 然而, 对一个大中城市动辄上万路甚至几十万路监控摄像头来说, 这一数字可能还不及百分之一. 若按每路高清视频平均 4 Mbps 码率计算, 实时传输十万路监控视频数据就需要 400 Gbps 带宽. 同样, 按每个配备多 GPU 卡的高性能服务器可以实时处理 (视频解码、特征抽取与分析) 十路高清视频流计算, 十万路监控视频的分析处理至少需要一万台服务器的云计算能力. 无法实时汇聚和处理大规模监控视频流数据, 就意味着城市大脑无法实时感知和分析当下城市正在发生的事态, 更无法根据实时态势做出及时的预测与决策支持. “数据大” 变不成 “大数据”. 更重要的是, 现在的监控系统从设计的时候就是为了存储而做的, 即没有统一时间戳, 也没有准确地理位置信息, 人可以实时监视, 但让计算机代替人完成全部的自动识别还做不到. 因此, 追本溯源, 造成目前这种 “有眼 (摄像头)、有脑 (城市大脑)” 但 “眼弱脑强” 现状的原因, 本质上是传统监控摄像头所采用的技术体系架构问题.

在现有视频监控系统中, 采用的是长期自然形成的 1-1 模式监控技术架构, 即一个摄像机输出一个视频流, 面向一种功能或用途: 有的摄像头负责大屏监视, 有的摄像头负责抓拍人脸, 有的摄像头负责车牌识别, 等. 在技术上, 1-1 模式采用的是 “源端图像视频压缩 $\rightarrow$ 传输 $\rightarrow$ 后端特征提取与分析识别” 的框架, 其中前端设备的任务是视频采集、压缩和传输, 云端服务器的任务是处理和分析, 包括视频解压缩、人工校验、对象检测、模式识别、事件分析等. 这种模式的好处是设备的安装调试比较简单. 然而, 由于特征提取与分析识别需要在解码重构后的图像视频上完成, 压缩将必定影响其性能 [1]. 为了减少传输带宽和节省存储, 部分视频监控系统甚至过度压缩, 从而造成图像视频质量过低, 视觉特征受损, 严重影响分析识别精度.

为验证这一论断, 进行一个简单的实验分析. 使用最新 AVS2 编码器 [2] 获得具有不同量化参数 (QP) 的重构视频, 再提取视频帧并测试不同任务的性能, 如视觉搜索、人脸识别和行人再识别. 如果编码参数 QP 变大, 图像视频的码率将相应减小, 但同时将降低重构的图像视频质量. 如图 1 所示, 各分析识别任务的性能都对 QP 变化敏感. 其中 QP 为 38 是临界值, 即当 QP 大于 38 时, 分析识别的效果将会急剧下降.

因此, 架构于现有视频监控系统基础上的图像视频分析识别性能总体比较低, 应用效率不高. 其根源在于, 现有视频监控系统是为存储数据并再由人工离线检查而设计, 从而造成大部分数据在其生存期内始终没有分析和利用, 也无法有效支持城市级视频大数据分析与挖掘. 这些问题迫使我们需要重新反思在智慧城市的应用背景下, 现有的 1-1 模式视频监控技术架构是否依然合理可行.

一种可能的解决方案是将多种不同的分析识别算法同时嵌入到摄像头端, 例如一些智慧停车摄像头集成了车牌识别、车辆跟踪与违章检测等功能, 一些人脸抓拍摄像头集成了 “人脸检测与识别” 功能. 从 1-1 模式扩展到了 1-m 模式, 是近期海康威视等厂商研制智能摄像头时所采取的主要技术路线. 1-m 模式大大提升了摄像头的利用效率与智能化水平, 但是这种单纯强调 “边缘计算” 的方案仍然无法解决 “眼脑合一” 的问题, 无法高效支撑云端的大规模监控视频分析与视觉搜索. 因此, 为了有效支持城市大脑, 必须颠覆现有 “单摄像头单输入流” 的监控技术架构.

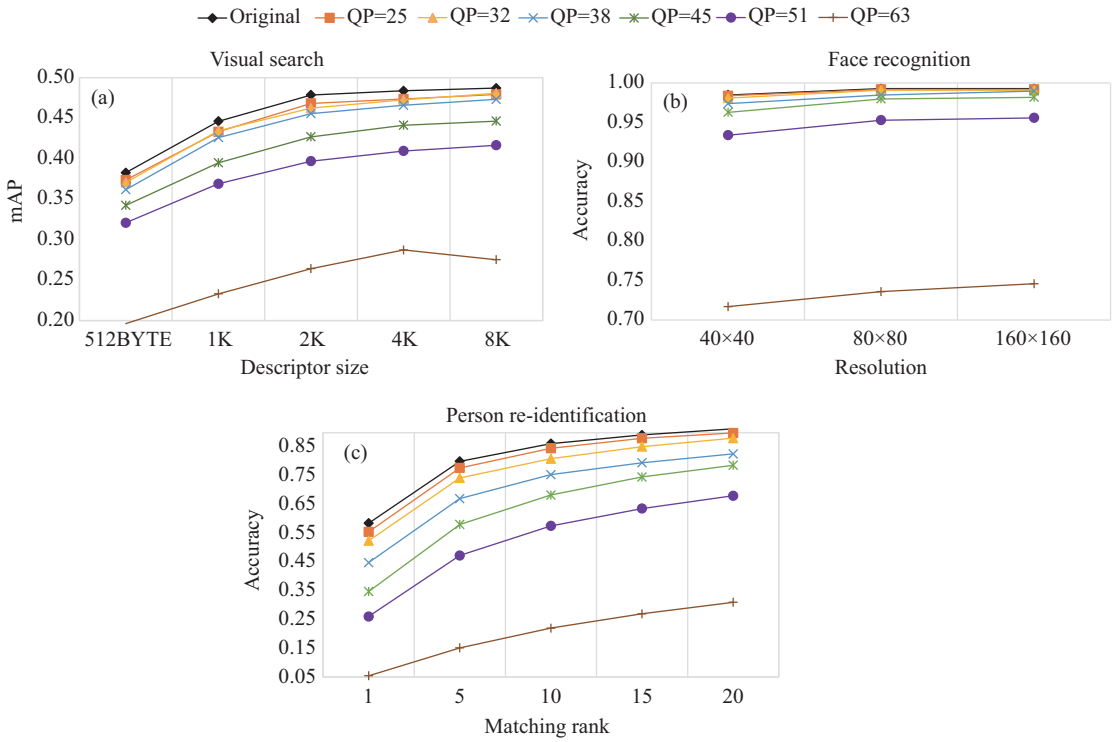

图 1　(网络版彩图) 视频压缩对不同分析检索任务的影响, 包括 (a) 视觉搜索, (b) 人脸识别, (c) 行人再识别. 实验中每个任务选择一个标准数据集, 利用最新的 AVS2 编码器并采用不同的量化参数 (QP) 来压缩图像与视频, 再利用不同重构图像视频来评测相应任务性能

Figure 1　(Color online) The effect of video compression on different analysis and retrieval tasks, including (a) visual search, (b) face recognition, and (c) person re-identification. In the experiments, we selected one benchmark dataset for each task, and utilized the state-of-the-art AVS2 codec to obtain the reconstructed images and videos with different quantization parameters (QPs). Then the reconstructed images and videos were used to evaluate the performance of different tasks

2　解决上述问题的对策: 数字视网膜

众所周知, 数码相机的生物学原型是人类的视网膜. 就像数码相机中能采集 “像素” 一样, 视网膜能获取并编码大量的视觉数据. 然而, 生物学研究表明 [3], 除了能编码像素之外, 视网膜还可以提取并编码场景或物体的特征, 如纹理、轮廓等. 因此, 视网膜可以看作是一个并行图像处理器: 利用感光器阵列获取图像或视频, 使用其内部电路来计算场景的神经表征, 再通过神经节细胞的轴突将其传送到更高层的视觉系统. 从这个角度来看, 传统的数码相机仅仅只模拟视网膜的一部分功能. 因此, 一个自然的问题就是, 如何借鉴 “人类视网膜同时具有影像编码与特征编码功能” 这一生物特性来研究和设计一种更高效的摄像头. 我们称之为数字视网膜摄像头 (retina-like camera), 简称为数字视网膜 (digital retina).

本文所定义的数字视网膜, 必须满足如下条件: (a) 使用全网统一的时间; (b) 提供精确地理位置; (c) 提供视频数据的高效编码功能; (d) 提供视频数据的紧凑特征表达; (e) 支持视频编码与特征表达的联合优化. 与以往传统摄像头相比, 数字视网膜的核心在于 “单摄像机双数据流”, 其中压缩视频流是为了存储和离线观看, 而紧凑特征流则是为了大数据分析与搜索. 图 2 描述了数字视网膜的概念架构, 其中数字视网膜成为联接城市大脑的 “慧眼”. 支撑数字视网膜的核心技术包括如下.

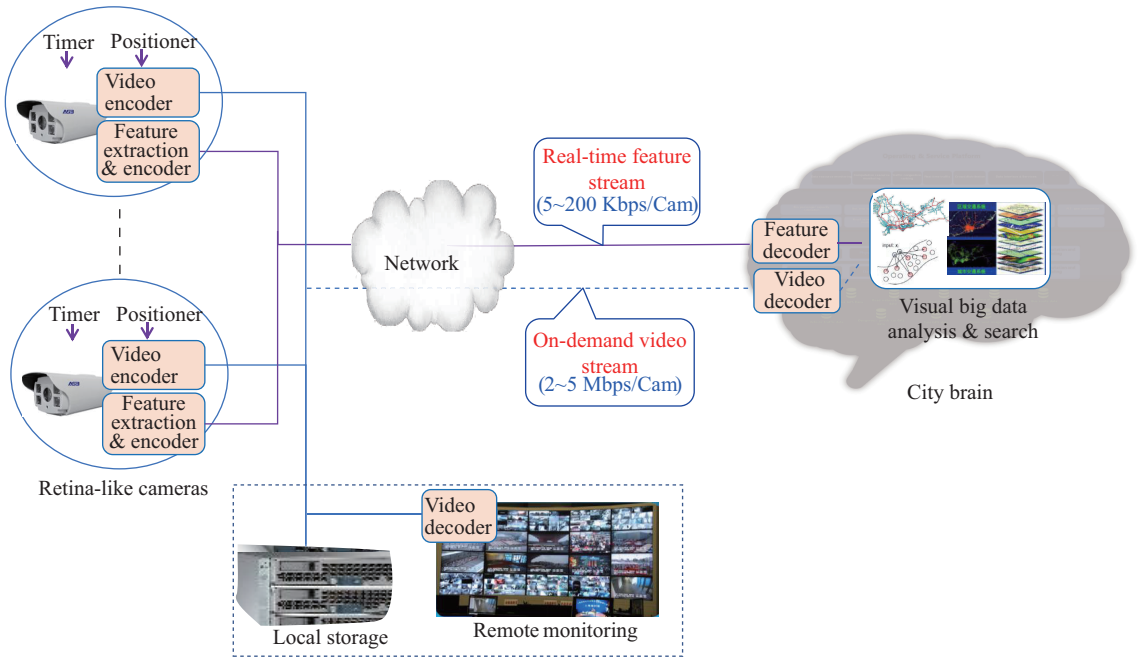

图 2 (网络版彩图) 数字视网膜成为联接城市大脑的 “慧眼”. 通过 “特征实时汇聚 + 视频按需调取” 来支撑城市视频大数据分析与搜索

Figure 2 (Color onine) The compound-eye camera system for the smart city by connecting a large number of digital retinas. In this system, feature streams can be realtime aggregated into the city brain, while video streams are saved in the local storages and pulled to the city brain only on demand

(1) 基于背景模型的场景视频编码. 现有监控摄像头通常采用通用的视频编码技术标准. 总体来看, 视频编码技术标准主要针对广播电视视频来制定, 采用混合编码框架, 压缩比每十年翻一番. HEVC/H.265 及我国 AVS2 属于目前最新的第 3 代标准, 压缩比达 300:1. 近年来, 中国学术界和产业界技术专家制定了专门针对监控视频的编码标准, 即 AVS2 场景编码标准. 该标准针对大多数监控视频场景相对固定的特点, 将背景建模融入视频压缩技术框架, 利用背景预测来消除监控视频中的场景数据冗余, 实现了监控视频压缩效率翻番 [4].

(2) 视频特征的紧凑表达. 视觉特征表征是图像视频分析处理的基础. 面向视觉搜索的紧凑视觉特征描述子国际标准 —— MPEG CDVS (ISO/IEC 15938-13) 是多媒体领域我国主导的首项 ISO 国际标准 [5]. 最近国内外相关技术专家正将其扩展到面向视频分析的特征描述子 MPEG CDVA. 深度学习近年来在图像分类、语音识别等领域取得巨大成功. 为此, 我们建立了深度特征的帧内帧间压缩框架, 利用 Hash 网络将浮点型深度特征进行量化, 并根据不同的内容设计了不同的帧间编码结构与模式, 例如当场景变化较大时需进行独立编码的 I-Feature, 当场景变化较小时可用前一帧对应特征表示的 S-Feature, 以及当场景有一定程度变化时仅需编码残差的 P-Feature, 在此基础上建立了特征码率与检索准确率间优化模型 (RLO), 从而获得了紧凑的深度描述子. 实验结果表明, 针对视觉搜索任务, 平均每对象仅需 100 bit/帧, 即可达到与无压缩深度特征相当的检索性能 [6].

(3) 视频编码与紧凑表达的联合优化. 数字视网膜同时输出压缩视频流和紧凑特征流, 因此应根据码流的大小, 设计联合优化函数来计算如何分配各自的码率, 使得压缩视频流能高效传输和可靠恢复, 而紧凑特征流则能满足分析识别需求. 联合优化涉及面向视频编码的率失真优化 (RDO) 和面向特征编码的率失准优化 (RLO). 初步实验结果表明 [7], 利用编码信息能使得特征编码获得更紧凑的描

述子, 而利用特征描述子则能为视频编码导出准确的仿射运动模型, 从而提升视频压缩效率.

硬件实现是数字视网膜大规模应用的前提. 近期的硬件实现方案可采取监控 SOC+FPGA 方案, 在中长期则可采用芯片实现方案. 在近期监控 SOC+FPGA 方案中, 可采用已有的智能监控 SOC (如 Hi3519) 来支持图像信号处理、AVS2 视频编码与简单 CNN 模型, 而 FPGA 支持多对象检测、快速跟踪、特征提取, 并实现特征的压缩编码. 在远期方案中, 需要设计一颗内嵌 CNN 处理器的多功能芯片, 同时支持图像信号处理、AVS2 视频编码、对象检测、特征提取与压缩、双流同步封装等.

与传统监控视频技术体系相比, 数字视网膜具有如下特点:

- 高性能. 特征提取直接可在未压缩视频帧上提取, 避免视频压缩使得特征受损, 从而影响分析识别性能.
- 高效率. 云端系统直接在解码后特征上进行分析, 避免云端在大规模图像视频数据上进行计算密集的特征提取运算.
- 可伸缩. 云端系统不依赖图像视频数据本身进行分析识别, 但可根据需要调取图像视频流, 从而避免占用大量传输带宽.
- 隐私保护. 对于城市大脑或者某些第三方分析系统而言, 只能获得特征而不是图像或视频本身, 因此能支持更好的隐私保护, 一定程度可防止监控图像或视频被滥用.
- 可软件定义. 由于在摄像头端固化神经网络处理器而非分析识别算法, 因此特征学习模型和摄像机参数可以实时地从云中心更新, 从而能有效支持算法升级与更新.

需要指出, 当数字视网膜未大规模普及时, 而全国数千万监控摄像头不可能短时替换, 则可采用基于智能边缘节点的解决方案, 即利用部署在派出所、区县公安局等本地机房的设备形成智能边缘节点, 实现高清视频转码存储、特征提取、简单分析识别, 提取的紧凑特征流再通过视频专网上传至云端的视频图像分析处理中心. 因此, 只要通过设置使得现有监控摄像头的输出压缩视频质量较高 (即 $QP < 38$), 可根据实时监控应用需求或对本地已存储监控视频的离线处理应用需求, 在本地服务器增加特征实时提取与编码硬件或软件, 云端增加特征汇聚、实时特征解码器硬件或软件.

3 可能的演进路线

为利用数字视网膜来构筑智慧城市的 “慧眼”, 应积极布局与推进相关标准制定、芯片与硬件实现、支撑软件开发与软硬件开源社区, 并适时开展大规模测试与应用.

(1) 标准化. 为了大规模生产和部署视网膜相机, 标准化是近期需要推进的首要任务. 目前, 在国内外科学家和技术专家的推动下, 已经或正在制定几个可用于数字视网膜的技术标准, 包括 IEEE 1857.4 和 AVS2 场景视频编码标准, MPEG 视觉特征紧凑描述子标准 CDVS 及其扩展 CDVA. 目前, 也正在讨论如何制定一些标准来解决 CNN 模型在 TensorFlow、Caffe2 等不同算法平台上的统一表示、相互转换和模型压缩问题, 从而使其可在 FPGA 和 ARM 等资源受限的硬件平台上有效实现和保存. 此外, 特征可互操作性、端云协作计算协议、城市数字视网膜的建设规范等也是潜在需要标准化的项目. 例如, 特征可操作性标准应定义在从不同模型提取的或由不同厂商提供的深度特征间如何进行自动识别、解析和交互.

(2) 软硬件开源. 开源软件的发展对人工智能产业与应用的发展起到了巨大的推动作用. 因此, 为了促进数字视网膜的广泛应用和产品化, 需要提供一些开源工具来支持相关技术和产品的开发, 包括场景视频转码工具、特征提取和压缩工具、大规模特征流聚合器, 以及基于特征的可视化图像视频大数据分析和挖掘工具. 此外, 还可以提供一些开源硬件工具来开发使用基于 FPGA 的数字视网膜.

(3) 大规模测试床. 需要建立一个大规模测试平台来评估和展示数字视网膜架构的技术优势. 这个平台至少应该包括上万路的监控摄像头, 地理上覆盖一个中等以上城市, 从而可以在真实场景中评估与数字视网膜相关的算法和技术. 目前, 国家自然科学基金和科技部 “云计算与大数据” 重点专项已经进行了部分项目布局, 预计在不久的将来可以取得一些显著进展.

4 结语

我国已明确提出 “到 2020 年, 基本实现全域覆盖、全网共享、全时可用、全程可控的公共安全视频监控建设联网应用”, 但是如果没有重大技术突破, 数千万摄像头根本无法实现 “全网共享” 的实时数据汇聚, 更不可能实现 “全时可用” 的联网分析识别, “数据大” 变不成 “大数据”, 巨大潜在价值无法发掘. 数字视网膜是应对上述挑战的一种可行的颠覆性技术发展方向. 因此, 应加大在数字视网膜相关基础理论、方法与关键技术的研究, 并布局与推进相关标准制定、芯片与硬件实现、支撑软件开发与软硬件开源社区, 最终目标是构筑服务于智慧城市的 “慧眼”, 并与城市大脑一起来支撑城市的精细化运营和动态管理.

参考文献

1 Gao W, Tian Y H, Huang T J, et al. The IEEE 1857 standard: empowering smart video surveillance systems. IEEE Intell Syst, 2014, 29: 30–39

2 Gao W, Ma S W. An overview of AVS2 standard. In: Advanced Video Coding Systems. Berlin: Springer, 2015. 35–49

3 Silveira R A D, Roska B. Cell types, circuits, computation. Curr Opin Neurobiol, 2011, 21: 664–671

4 Zhang X G, Huang T J, Tian Y H, et al. Background-modeling based adaptive prediction for surveillance video coding. IEEE Trans Image Process, 2014, 23: 769–784

5 Duan L Y, Chandrasekhar V, Chen J, et al. Overview of the MPEG-CDVS standard. IEEE Trans Image Process, 2016, 25: 179–194

6 Ding L, Tian Y H, Fan H F, et al. Rate-performance-loss optimization for inter-frame deep feature coding from videos. IEEE Trans Image Process, 2017, 26: 5743–5757

7 Zhang X, Ma S W, Wang S S, et al. A joint compression scheme of video feature descriptors and visual content. IEEE Trans Image Process, 2017, 26: 633–647

Digital retina: revolutionizing camera systems for the smart city

Wen GAO[1], Yonghong TIAN[1*] & Jian WANG[2]

1. *School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871*, China*;
2. *Alibaba Group, Hangzhou* 311121*, China*
* Corresponding author. E-mail: yhtian@pku.edu.cn

Abstract The primary viewpoints presented in this article are as follows: (1) The method to real-time gather and aggregate all kinds of urban big data, especially image and video data from video surveillance networks, and subsequently analyze and mine the value of these big data in the city brain to effectively support the urban operation and management is a key problem in the development of smart cities. (2) Recently, some city brains are established to mine the large visual data source to obtain valuable insights about the activities in the city (e.g., the urban traffic status). However, it is recognized that compression will inevitably affect visual feature extraction, and consequently degrading the subsequent analysis and retrieval performance. More importantly, it is impractical to aggregate all video streams from hundreds of thousands of cameras distributed across the city into a city brain for big data analysis and retrieval. These issues and challenges are rooted in the camera framework currently in use. (3) To address these challenges, a new camera framework should be developed from the fact that retina can encode both pixels and features. Such a retina-like camera, or directly referred to as digital retina, is typically equipped with a globally unified timer and an accurate positioner, and can output two streams simultaneously, including a compressed video stream for online/offline viewing and data storage, and a compact feature stream extracted from the original image/video signals for visual analysis and search. By real-time feeding only the feature streams into the city brain, these digital cameras form a compound-eye camera system for the smart city. (4) To promote the wide application of digital retinas in the smart city, the relevant works should be addressed in the near future, including standardization, hardware implementation, open-source software development, and the deployment of large-scale testbeds.

Keywords smart city, city brain, digital retina

Wen GAO was born in 1956. He received his Ph.D. degree in electronics engineering from the University of Tokyo, in 1991. Currently, he is a Boya chair professor at the Peking University, and also serves as the president of CCF from February 2016. Professor Gao works in the areas of multimedia and computer vision, including video coding, video analysis, multimedia retrieval, face recognition, multimodal interfaces, and virtual reality. His most cited contributions are model-based video coding and face recognition.

Yonghong TIAN was born in 1975. He received his Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, in 2005. Currently, he is a full professor with the School of Electronics Engineering and Computer Science, Peking University. His research interests include machine learning, computer vision, and multimedia big data.