软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(10):3635–3655 [doi: 10.13328/j.cnki.jos.006631]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度学习的新型视频分析系统综述*

孟令睿, 丁光耀, 徐 辰, 钱卫宁, 周傲英

(华东师范大学 数据科学与工程学院, 上海 200062)

通信作者: 徐辰, E-mail: cxu@dase.ecnu.edu.cn

摘 要: 摄像设备在生活中的普及, 使得视频数据快速增长, 这些数据中蕴含丰富的信息. 早期, 研究人员基于传统的计算机视觉技术开发视频分析系统, 用于提取并分析视频数据. 近年来, 深度学习技术在人脸识别等领域取得了突破性进展, 基于深度学习的新型视频分析系统不断涌现. 从应用、技术、系统等角度, 综述了新型视频分析系统的研究进展. 首先, 回顾了视频分析系统的发展历史, 指出了新型视频分析系统与传统视频分析系统的区别; 其次, 分析了新型视频分析系统在计算和存储两方面所面临的挑战, 从视频数据的组织分布和视频分析的应用需求两方面探讨了新型视频分析系统的影响因素; 再次, 将新型视频分析系统划分为针对计算优化的系统和针对存储优化的系统两大类, 选取其中典型的代表并介绍其核心设计理念; 最后, 从多个维度对比和分析了新型视频分析系统, 指出了这些系统当前存在的问题, 并据此展望了新型视频分析系统未来的研究和发展方向.

关键词: 视频分析系统; 深度学习; 计算优化; 存储优化

中图法分类号: TP391

中文引用格式: 孟令睿, 丁光耀, 徐辰, 钱卫宁, 周傲英. 基于深度学习的新型视频分析系统综述. 软件学报, 2022, 33(10): 3635–3655. http://www.jos.org.cn/1000-9825/6631.htm

英文引用格式: Meng LR, Ding GY, Xu C, Qian WN, Zhou AY. Survey of Novel Video Analysis Systems Based on Deep Learning. Ruan Jian Xue Bao/Journal of Software, 2022, 33(10): 3635–3655 (in Chinese). http://www.jos.org.cn/1000-9825/6631.htm

Survey of Novel Video Analysis Systems Based on Deep Learning

MENG Ling-Rui, DING Guang-Yao, XU Chen, QIAN Wei-Ning, ZHOU Ao-Ying

(School of Data Science and Engineering, East China Normal University, Shanghai 200062, China)

Abstract: The popularity of camera devices in daily life has led to a rapid growth in video data, which contains rich information. Earlier, researchers developed video analytics systems based on traditional computer vision techniques to extract and then to analyze video data. In recent years, deep learning has made breakthroughs in areas such as face recognition, and novel video analysis systems based on deep learning have appeared. This paper presents an overview of the research progress of novel video analytics systems from the perspectives of applications, technologies, and systems. Firstly, the development history of video analytics systems is reviewed and the differences are pointe out between novel video analytics systems and traditional video analytics systems. Secondly, the challenges of the novel video analysis system are analyzed in terms of both computation and storage, and the influencing factors of the novel video analysis system are discussed in terms of the organization and distribution of video data and the application requirements of video analysis. Then, the novel video analytics systems are classified into two categories: Optimized for computation and optimized for storage, typical representatives of these systems are selects and their main ideas are introduced. Finally, the novel video analytics systems are compared and analyzed from multiple dimensions, the current problems of these systems are pointed out, and the future research and development direction of novel video analytics systems are looked at accordingly.

Key words: video analysis system; deep learning; computational optimization; storage optimization

* 基金项目: 国家自然科学基金(61902128); 上海市扬帆计划(19YF1414200)
本文由“智慧信息系统新技术”专题特约编辑邢春晓研究员、王鑫教授、张勇副研究员、于戈教授推荐.
收稿时间: 2021-07-20; 修改时间: 2021-08-30; 采用时间: 2021-12-24; jos 在线出版时间: 2022-02-22

随着大量摄像机部署在公共场所甚至个人家庭中, 这些设备产生的视频数据迅速增多. 调查显示: 2016 年产生的视频数据占所有互联网流量的 73%以上[1]; 一个中等规模的城市, 仅一天就能产生 PB 级数据量的视频[2]. 这些视频中蕴含的丰富信息可以帮助解决现实生活中的一些难题, 例如: 道路和交叉路口的视频有助于及时检测拥堵、违规和事故[3], 并为交通规划决策提供信息[4,5]; 室内的视频有利于检测异常情况并及时预警[6]等. 但是, 如果依靠人工观看视频的方式提取这些信息, 不仅耗时耗力且结果误差较大[7].

面对视频处理需求, 如何自动、高效地从视频数据中提取相应信息, 是视频分析系统的关键. 早在 20 世纪 90 年代, 国内外的很多公司, 包括 IBM、Virage 等, 都开发了视频分析系统进行图像检索与对象查询. 本文将这些系统称为传统视频分析系统. 但是, 随着视频数据的快速增长和视频分析应用需求的增加, 这些系统的不足逐渐显现, 主要表现在这些系统基于传统的计算机视觉的方法进行对象查询, 而这些对象查询所使用的特征需要人为地选择和提取. 这是一种半自动的实现方式, 这种实现方式导致系统的准确度低、查询对象有限、识别能力有限. 在之后的一段时间中, 视频分析系统的发展停滞不前. 直到 2012 年, 在 ImageNet 图像分类比赛上, 深度神经网络 AlexNet[8]取得了当时最好的结果, 这为视频分析系统提供了新的方向. 通过使用深度神经网络, 系统可以自动提取并学习对象丰富的特征, 并推理得到准确的结果. 此后, 深度神经网络(deep neural network, DNN)的种类逐渐增多, 功能逐渐多样化. 更进一步地, 因深度神经网络的兴起, 以深度神经网络为本质的深度学习方法也逐步在人工智能的多个领域中得到广泛应用. 时至今日, 国内外研究人员仍致力于使用深度学习方法来解决现实生活中的一些难题. 随着深度学习日益受到广泛关注, 近年来出现了很多基于深度学习的视频分析系统, 本文将这类系统称为新型视频分析系统.

然而, 在视频数据不断增长和深度神经网络层数逐渐增多的趋势下, 新型视频分析系统在计算和数据存储过程中都面临着挑战, 例如, 系统如何快速甚至实时地分析不断到达的数据、系统如何压缩数据以节省存储空间等. 面对这些挑战, 研究者们通过对视频数据组织分布的观察以及对视频分析应用需求的分析, 逐步探索了不同的优化方向, 并由此开发了相关的原型系统. 本文对这些系统进行了总结, 并将其归纳为两类系统: 针对计算优化的系统和针对存储优化的系统. 对于每一类系统, 本文介绍了其设计思想, 并从多个方面对其进行了总结和分析.

本文第 1 节回顾视频分析系统的发展历程. 第 2 节阐述新型视频分析系统面对的挑战及其影响因素. 第 3 节分类介绍典型的新型视频分析系统. 第 4 节对比总结这些系统. 第 5 节探讨未来值得关注的研究方向. 第 6 节总结全文.

1 视频分析系统的发展

在新型视频分析系统出现之前, 人们采用传统视频分析系统处理视频数据. 传统视频分析系统主要采用经典的计算机视觉的方法, 通过获取视觉内容和语义内容的方法, 实现对视频数据的处理与分析. 图 1 展示了传统视频分析系统的查询处理流程.

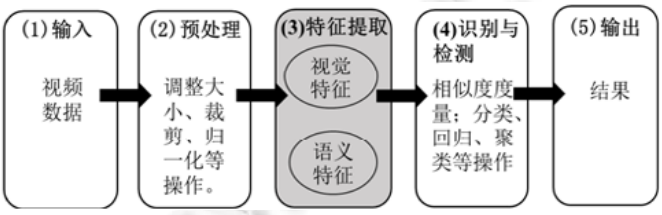

图 1 传统视频分析系统处理流程

这一流程可以分为数据输入、预处理、特征提取、识别与检测、结果输出这 5 个部分, 其中, 特征提取是传统视频分析系统的核心部分. 这些要提取的特征可以分为视觉特征和语义特征两大类: 视觉特征即图像通过线性变换得到的特征. 例如, 颜色、纹理、形状等; 语义特征则是通过分析图像中事物与事物之间的关系, 从中提取出的对象、事件等高级概念. 例如, 足球比赛的观众席上有一位球迷看起来很兴奋. 显然, 这两类特

征的含义不同, 因此, 对这两类特征的提取要求是不同的. 根据提取特征的不同, 可将传统视频分析系统分为基于内容查询(content-based image retrieval, CBIR)的系统和基于语义查询的系统这两类. 其中, 基于内容查询的系统具有代表性的有 QBIC[9]、Virage[10]、PhotoBook[11]、FourEyes[12]、VisualSEEK[13]等.

QBIC (query by image content)系统是 IBM 公司开发的第 1 个基于内容的商业化的图像检索系统, 图 2 展示了 QBIC 系统的架构. 可以看出, 该系统的处理流程符合传统视频分析系统的处理流程. 它对输入的图像或者视频进行预处理后, 再提取颜色、纹理、形状等视觉特征, 然后将这些特征存储在数据库中. 当用户进行查询时, 该系统根据用户提供的要查询图像的特征与数据库中保存的图像的特征进行相似度度量, 最终返回匹配度高的图像作为结果.

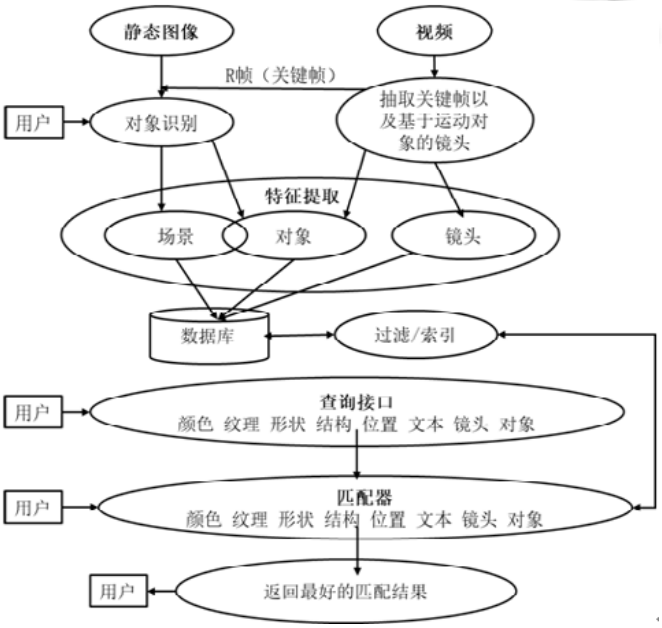

图 2 QBIC 系统[9]

例如, 当用户想要在一段交通视频中寻找消防车时, 用户需要为系统提供要检索的消防车的特征(如红色、长方体). 系统根据用户提供的特征进行查询, 并返回视频中存在消防车的图像. 系统处理该查询的具体过程为: 当视频采集完成后, 系统首先按帧将其分割为图像, 对于每一帧图像, 系统分析图像中对象的颜色和形状, 并将分析结果存入数据库中; 之后, 系统的匹配器基于相似度度量方法从数据库中寻找红色且形状为长方体的对象, 符合检索特征的对象将作为最终结果返回给用户. 由于系统能够提取的特征有限, 除了颜色和形状外, 系统无法提取该对象具体部件的特征. 例如, 消防车上必备的水枪、梯架、工作斗等, 这导致了检索的结果并不准确. 对于上述例子, 系统虽然获得了所有具备红色和长方体特征的对象, 但是得到的不一定是消防车, 还可能是红色的小汽车. 值得指出的是, 如果用户误以为消防车是蓝色的, 要求系统将蓝色且形状为长方体的对象作为消防车的查询结果, 这显然也会导致结果不准确. 此外, 仅通过颜色和形状, 系统可表达的查询对象也往往是有限的. 其他几个基于内容查询的系统的处理流程与 QBIC 类似. 表 1 总结了这些系统主要提取的视觉特征内容以及各个系统的设计特点.

但在实际应用中, 用户不仅想要获取图像的视觉特征信息, 也想获取图像表达的含义. 这些图像含义就是图像的语义特征, 它包含了人们对图像内容的理解. 这种理解是无法从图像的视觉特征获得的, 而是依赖于人们的认知. 例如, 在足球比赛中, 用户想查询的是进球瞬间对应的视频[14], 而不是整个视频里人或者其他物体的视觉特征, 即用户想要查询的内容与图像含义相关. 此时, 如果有一个球形的对象在一个矩形的对

象内停留, 同时伴有响亮的欢呼声和长哨声, 即可判断进球了. 显然, 这些欢呼声和哨声是无法通过视觉特征提取到的.

表 1 基于内容查询的系统总结

系统	主要提取的视觉特征	特点
QBIC	颜色、形状、纹理	无
Virage	色彩、布局、纹理、结构	将视觉特征分为通用特征和领域相关特征这两类
PhotoBook	形状、纹理、面部	设计了 3 个子系统分别提取对应的特征
FourEyes	形状、纹理、面部	PhotoBook 的改进版, 结合人的因素
VisualSEEK	颜色、纹理	面向 Web 的图像搜索引擎

基于这种查询需求, 研究者开发了 Stratification[15]、LHVDM[16]、Video modelling[14]等基于语义查询的系统. 这些系统的查询流程与基于内容查询的视频分析系统类似. 表 2 总结了这些系统提取语义特征的方式以及各个系统的特点. 一般而言, 这些系统都要求研究者拥有领域知识且明确用户需求, 研究者对相关领域知识的不了解和对用户查询内容的不清楚, 会导致最终结果的不准确.

表 2 基于语义查询的系统总结

系统	语义特征提取的方式	特点
Stratification	通过为视频的每一秒添加注释增加语义	繁琐且耗时
LHVDM	通过超链接的方式将关系的事件链接起来增加语义	繁琐且耗时
Video modelling	通过用户自定义的规则提取视频数据的语义	需要针对不同的场景定义不同的语义规则

通过以上对基于内容和基于语义的传统视频分析系统的介绍, 本文对传统视频分析系统存在的问题总结如下: (1) 该类系统提取的视觉和语义特征有限, 导致查询的对象有限; (2) 该类系统查询结果的准确度取决于用户查询时选取特征的好坏, 且研究者需要具备丰富的领域知识以确保结果的准确性; (3) 该类系统对特征的识别与检测技术多为相似度度量(如距离函数)或者仅含单层非线性变化的浅层学习结构(例如, 支持向量机 SVM、传统隐马尔可夫模型等). 这类技术的查询结果易受拍摄环境的影响, 比如光照变化、尺度变化、视角变化、遮挡以及背景的杂乱等都会对其造成较大的影响, 从而导致系统的识别能力有限[17].

深度神经网络的出现和深度学习的发展, 为视频分析系统提供了新的方向. 深度学习的概念最开始由 Hinton[18]提出, 它的提出来源于人们对人工神经网络的研究, 其动机在于模拟人脑的机制来自动学习数据样本的特征. 早期虽然也有人尝试利用一些通用的神经网络进行视频分析, 但是由于其深度太浅, 只能用于识别检测包含单一目标的图像[19]. 而且由于缺乏大规模的训练数据, 在利用这些浅层神经网络实现的应用中会出现过度拟合等现象[20], 导致查询效果不如传统的视频分析系统. 直到 2012 年, AlexNet[8]在 ImageNet 比赛上得到了最好的结果, 当时, AlexNet 识别效果超过了所有浅层的方法, 深度神经网络开始发展起来. 图 3 展示了一个深度神经网络的结构, 该网络由输入层、卷积层、池化层、全连接层和输出层组成[21]. 其中, 卷积层和池化层是该深度神经网络中用于特征提取的核心模块. 网络的每一层都可以用来自动提取图像不同的特征, 较低层会检测基本特征, 较高层则可检测更复杂的特征. 此后, 随着可供神经网络训练的数据规模越来越大, 神经网络训练过程中过拟合等问题得到缓解, 复杂的网络结构应运而生[22,23]. 例如, GoogLeNet[24]、Resnet[25]、Yolo[26]等. 在深度神经网络研究的基础上, 深度学习开始受到广泛的关注.

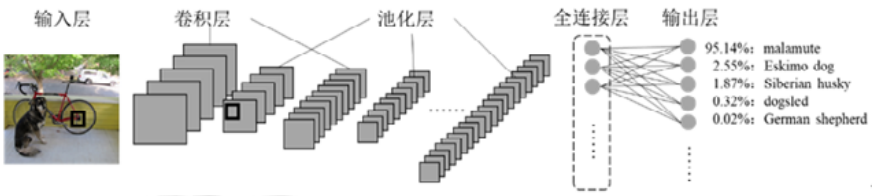

图 3 图像分类 CNN 的结构

虽然深度学习提供了较好的特征提取能力, 但由于深度神经网络的大规模参数特性, 例如利用深度神经

网络执行一次操作的计算总量可达数十亿次, 也给处理设备的计算能力带来巨大的挑战. 随着处理设备相关软件和硬件的发展, 该挑战也逐步被克服, 主要表现在图形处理器(graphics processing unit, GPU)的发展和GPU 计算库可用性的增强[27]. 与传统中央处理器(central processing unit, CPU)相比, GPU 的运算速度要高出一个数量级甚至以上, 使其对深度学习算法尤其有用. 同时, 随着开源软件包在深度学习领域的普及与应用, 很多深度学习库, 包括 TensorFlow[28]、Caffe[29]等, 在诸如卷积操作等高效 GPU 实现中得到广泛应用. 这些深度学习库提供了一系列深度学习组件和函数接口, 以方便用户直接调用. 因此, 深度学习开始在人脸识别[30,31]、语音识别[32,33]、机器翻译[34]、推荐系统[35,36]等人工智能领域中得到广泛应用.

在深度学习发展的背景下, 通过运用深度神经网络, 新型视频分析系统可以从大量的数据中自主学习不同抽象层次的特征且得到准确的结果. 图 4 展示了新型视频分析系统对查询需求的处理过程. 与传统视频分析系统的处理流程相比, 新型视频分析系统利用神经网络推理的方法取代了传统视频分析系统中特征提取的步骤[37]. 这类系统表达能力强、结果准确, 而且无需领域知识和与用户交互即可处理多对象的大规模数据. 在此基础上, 基于深度学习的新型视频分析系统不断涌现. 目前, 这类系统可实现三大类应用: 对象分类、对象检测、路径跟踪. 其中, 对象分类是获取对象所属的类别. 常见的对象分类应用又可细分为两类: 二分类查询应用和多分类查询应用. 二分类查询即判断视频当前帧是否包含要查询的对象, 而多分类查询则是要判断视频每一帧中包含的所有对象的类别. 对象检测是获取对象在图像中的位置, 路径跟踪是获取同一对象在连续时间内的位置信息. 图 5 展现了近年来在数据库相关领域和操作系统设计与实现相关领域的会议中出现的新型视频分析系统. 由图 5 可以看出, 新型视频分析系统已经成为了当前研究的热点之一.

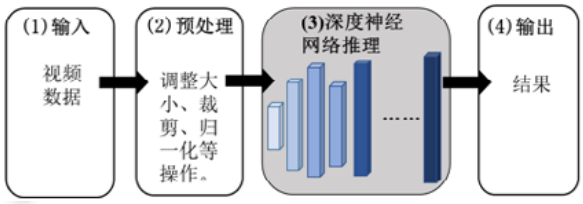

图 4 新型视频分析系统处理流程

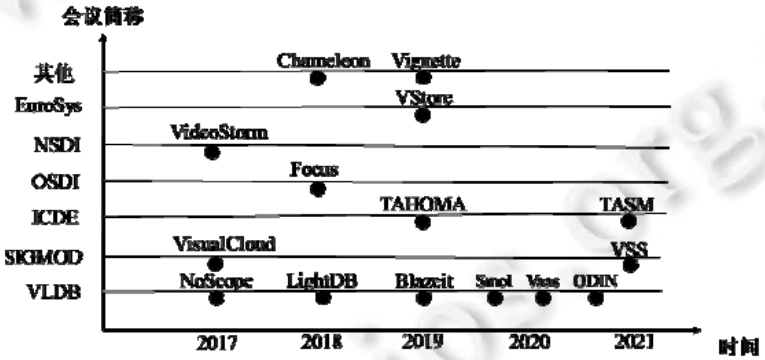

图 5 近年来出现的新型视频分析系统

2 新型视频分析系统的挑战与影响因素

本节首先从一个典型的视频分类应用出发, 将新型视频分析系统的处理流程分为计算和存储两部分, 然后分析与总结新型视频分析系统中来自这两方面的挑战. 最后, 本节从视频数据的组织分布以及视频分析系统的应用需求出发, 探讨影响新型视频分析系统性能的因素, 并指出系统可能的优化方向.

2.1 新型视频分析系统的挑战

新型视频分析系统面临的挑战来自于计算和存储两个方面. 图 6 展示了一个基于深度学习的视频分类应

用，该应用的处理流程与第 1 节提到的新型视频分析系统的处理流程相一致. 当用户要查询出现在该视频中的所有公交车时，系统读入待查询的视频数据并对其进行预处理，之后将预处理后的数据输入到 YoLo 深度神经网络中进行推理. 最后，系统将结果输出并反馈给用户. 在这个应用处理过程中，预处理和深度神经网络推理过程涉及到对数据的计算，输入和输出过程涉及到对数据的存储. 因此，本文将新型视频分析系统的处理过程分为计算和存储两部分. 同时，本文对新型视频分析系统面临的挑战，也将从这两个方面展开分析.

(1) 预处理、推理的计算开销限制了系统性能的提升. 如前所述，新型视频分析系统的计算过程分为预处理和推理两个部分. 影响两者计算开销的因素有多种.

- 首先，不同模型的推理性能差别较大，系统利用结构复杂的模型(如通用模型)进行推理较为耗时，而利用结构简单的模型(如专用模型)相对较快;
- 其次，数据的组织结构往往会影响系统性能，高分辨率的视频数据通常会产生较高的预处理与推理延迟. 更多地，视频数据中往往存在内容相似的数据帧，这些相似数据容易造成冗余计算，从而增加推理延迟;
- 最后，资源在预处理任务与推理任务之间的不合理分配，容易导致预处理或推理成为系统性能的瓶颈.

因此，基于上述分析，系统如何优化预处理、推理的过程以及合理分配资源，进而降低两者的计算开销，是新型视频分析系统所面临的挑战之一;

(2) 视频数据给系统带来了昂贵的存储和解压成本. 系统在获取视频数据后，首先将其保存至存储空间，随后在分析过程中，系统需要将这些数据解码成目标组织结构(如目标分辨率等). 然而，日益增长的视频数据要求系统提供更多的存储空间，使得存储成本不断增加. 与此同时，繁琐的解码操作通常较为耗时，反复执行这些操作会产生额外的开销，降低分析过程的响应延迟. 幸运的是，视频数据独有的特点为解决这些问题提供了机遇. 例如，在同一地方的多个摄像头拍摄的数据中，可能存在相似的部分，系统可以通过避免对相似部分的冗余存储来降低存储开销. 此外，多个应用可能会多次使用同一份数据，系统可以通过避免对同样的数据反复解码来降低解码开销. 因此，基于上述分析，系统如何利用视频数据的特点来减少存储与解码的数据量，从而降低存储与解码开销，是新型视频分析系统所面临的另一个挑战.

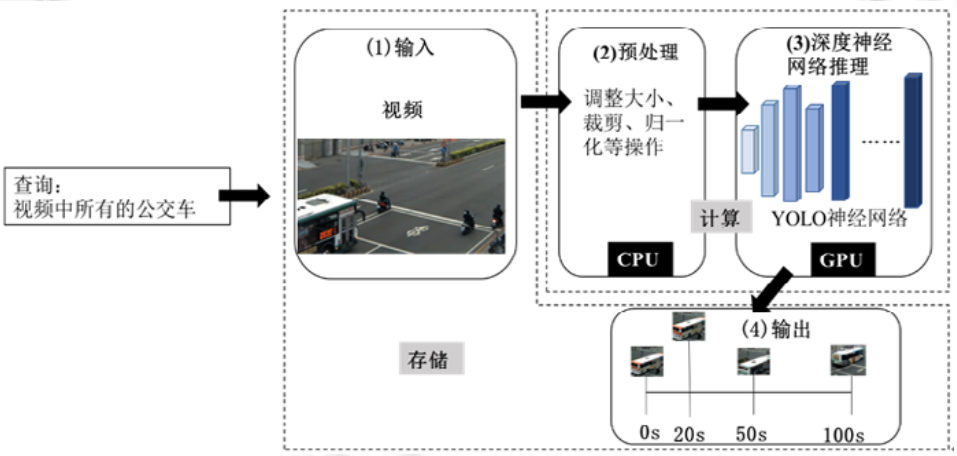

图 6 一个新型视频分类的应用

2.2 新型视频分析系统的影响因素

面对上节中提出的挑战，本节通过探究视频数据以及视频分析应用的特点，从视频数据的组织分布以及视频分析的应用需求两个方面，总结新型视频分析系统性能的影响因素，并指出利用这些因素提升系统性能

的新方向.

2.2.1 视频数据的组织分布

视频数据的组织分布可以从视频数据的组织结构和分布特性两方面加以分析.

- 一方面, 组织结构包括视频数据的物理属性和数据存储格式. 其中, 物理属性即数据对象性质和数据之间的关系. 例如, 视频的时长为 1 h、帧宽度为 1 920、帧高度为 1 028、数据速率为 1 474 kb/s、总比特率为 1 600 kb/s、帧率为 25.00 帧/s 等, 这些都属于视频数据的属性. 此外, 在数据存储格式上, 由于视频数据属于非结构化数据且没有规范的存储模型, 因此视频数据的存储形式丰富多样(如 H264[38]、HEVC[39]等);
- 另一方面, 分布特性包含了视频的空间和时间特性, 其展现出的是视频数据在空间和时间上的变化规律. 例如, 当前视频是上海市某一交叉路口的数据, 视频画面中有红绿灯、行人、车辆、花坛等多个对象, 随着时间的推移, 行人和车辆的数量在发生变化, 而花坛却是静止的.

对于视频数据的组织结构, 视频数据具有的帧率、分辨率、通道数等多种物理属性都会对视频数据的存储和计算性能产生影响. 例如, 未压缩的视频需要更大的存储空间[40]; 低帧率、低分辨率视频数据往往拥有较低的预处理和推理延迟. 因此, 在实际应用处理过程中, 系统可以通过调整视频的属性参数, 优化存储和计算过程, 进而提升系统性能.

对于视频数据的分布特性, 在空间上, 当从单个的位置固定的摄像头获取数据时, 所得到的视频数据角度单一. 系统利用这些数据进行推理时, 神经网络模型中部分层的中间结果可能不会影响最终结果, 这些层引起了多余的无效计算. 因此, 系统可以考虑精简模型, 去除这些无效的推理层, 从而加快推理的速度. 与此同时, 在实际情况中, 同一位置往往会布置多个摄像头, 方便用户从不同角度观察同一地点的情况. 这些摄像头捕获的数据存在交叉的部分, 容易造成冗余存储与计算. 因此, 系统可以考虑在存储时去除重复的部分来节省存储空间; 在计算时仅处理一次重复数据来避免冗余计算, 进而加速计算. 在时间上, 视频数据随时间的变化而呈现不同的特性. 当视频帧之间的时间间隔较小时, 视频帧内容相似度较高, 即存在相似性. 例如, 在一个视频中, 相邻几帧的图像几乎相同. 针对这类相似性, 视频分析系统高效的做法是: 通过判断这些帧的相似性, 系统仅处理相似帧中的 1 帧或者仅处理前后帧中差异的内容, 从而加速计算. 值得注意的是: 当视频帧之间的时间间隔较大时, 视频帧内容相似度较低, 即存在不平衡的现象. 例如, 在交通道路捕获的视频中, 白天车辆较多, 晚上车辆较少. 不平衡现象往往会影响系统性能, 例如, 在面对车辆检测负载时, 系统推理车辆较多的视频帧较为耗时, 而推理车辆较少的视频帧则相对较快. 因此, 系统需要能够自动判断不平衡现象, 并制定相应的应对方案.

2.2.2 视频分析的应用需求

视频分析的不同应用对实时性和准确度有着不同的要求. 有些应用对视频的分析仅仅是为了回答“事后查询”, 这些应用通过统计过去一段时间内的数据用于分析决策等, 对实时性的要求不高. 例如, 通过统计过去一个月内店铺的人流量, 店长可以合理安排服务人员. 然而还存在一些应用, 需要对源源不断到来的视频进行实时分析, 即要求较高的实时性. 例如, 用于 AMBER 警报的车牌识别器需要及时获取车牌数据[41], 以发现异常, 及时报警. 在实际处理过程中, 为了提高应用的实时性, 系统可以考虑在计算时为应用提供更多的资源、换用推理耗时较小的模型等. 除此之外, 有些应用对视频的处理结果要求具有较高的准确度. 例如, 在高速路口收费站的车牌识别器[42]需要准确地识别出不同车辆的车牌号码, 否则会造成计费错误. 有些应用则可以容忍一定的结果错误, 例如在道路中通过统计汽车数量调整交通信号灯的应用, 可以接受计数结果在一定范围内的误差, 而且这些误差并不会对最终的结果产生较大的影响. 因此, 基于上述分析, 系统可以针对不同应用的实时性、准确性要求, 合理地调整神经网络模型和资源分配策略等系统配置.

3 新型视频分析系统介绍

如第 2 节所述, 新型视频分析系统的挑战主要来自于计算和存储两个方面. 因此, 本节将从对计算进行优

化和对存储进行优化两方面展开，并分类介绍现有的新型视频分析系统.

3.1 针对计算进行优化的系统

在新型视频分析系统中，模型推理描述了计算的主要过程，以视频数据作为该过程的输入，这一过程依赖于计算资源. 因此，新型视频分析系统一般从模型、数据以及资源这 3 个方面对计算进行优化.

- 针对模型进行计算优化的策略：新型视频分析系统通过精简神经网络模型来优化计算. 常用的两种优化技术分别为：压缩通用的神经网络模型和训练专用化的神经网络模型. 这两种优化技术都达到了精简神经网络模型的目的，从而加快了系统的推理速度. NoScope[43]、Blazeit[7,44]、TAHOMA[45]、Focus[46]、Deluceva[47]、ODIN[48]等系统都采用了这种策略来优化计算；
- 针对数据进行计算优化的策略：新型视频分析系统通过改变数据的组织结构以及减少神经网络模型所需处理的数据量来优化计算. 通过改变数据的组织结构，预处理和推理速度得到了显著提升；减少神经网络模型处理的数据量，进一步加速了推理过程. NoScope、Foucs、Smol[49]、MIRIS[50]等系统都采用了这种策略来优化计算；
- 针对资源进行计算优化的策略：新型视频分析系统通过动态的任务调度和资源分配来优化计算. 通过对任务的合理调度，系统能够高效地利用 GPU 资源，从而降低了应用的计算延迟. 更进一步地，通过在多个应用间合理分配资源以满足不同应用的需求，平衡多个应用的计算延迟. Smol、Chameleon[51]、VideoStorm[52]等系统都采用了这种策略来优化计算.

3.1.1 针对模型进行计算优化

神经网络模型的推理延迟影响了计算的时间. 目前来看，现有的神经网络模型层次多，使得推理非常耗时. 为了降低推理延迟，现有的新型视频分析系统采用压缩通用神经网络模型或训练专用化的神经网络模型来解决这一问题.

- 压缩是一种以牺牲精确度为代价来加快推理速度的技术，该技术包括去除部分卷积层、矩阵修剪等. 例如，Focus 系统通过压缩后的神经网络模型过滤没有对象的数据. 然而，相比原始的神经网络模型，压缩后的神经网络模型将损失精确度，因此，系统需要结合其他技术来提高计算结果的精确度；
- 除了压缩通用的神经网络模型这项技术外，一种更为流行的技术是训练专用化的神经网络模型. 这项技术利用了前文提及的视频数据的一些特性. 例如，在摄像头角度固定的情况下，视频中的对象仅会从固定角度出现，而且在特定场景下，视频中的对象种类相对较少. 基于这些特性，系统通过训练专用化的神经网络模型来执行特定的推理任务. 与压缩后的神经网络相比，专用化的神经网络在执行推理时能够得到更准确的结果，但却牺牲了神经网络模型的通用性. 我们下文要介绍的 NoScope、Blazeit、TAHOMA、Deluceva、ODIN、MIRIS 等系统都采用了这项技术进行推理.

在上述系统中，NoScope、Blazeit、TAHOMA 利用专用化的神经网络模型实现了对象分类. 当使用专用化的神经网络模型进行分类推理时，与训练数据类似的测试数据会得到准确的结果. 然而，如果测试数据与训练数据相差较大，那么专用化的神经网络模型无法实现准确的判断. 为了准确判断测试数据中对象的类别，这3个系统都会设置置信度阈值. 在该设置的基础上，系统会计算出测试数据的置信度，当置信度大于设定的阈值时，则输出测试数据的分类结果；反之，系统或许需要作进一步处理，处理的方式取决于具体的系统. NoScope 系统是由斯坦福实验室开发的较早出现的新型视频分析系统，该系统用于二分类应用中. 图 7 展示了 NoScope 系统的架构.

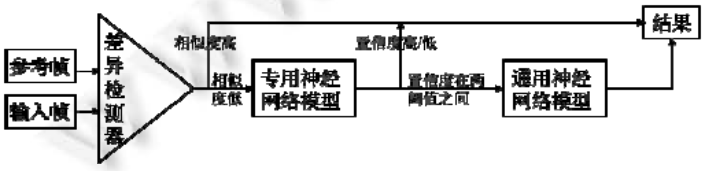

图 7 NoScope 架构[43]

如图 7 所示, 系统除了使用前文提及的高置信度阈值外, 还设置了一个低置信度阈值来过滤无关的结果. 当专用神经网络模型检测到测试数据的置信度大于高阈值时, 表明当前数据一定包含所查询对象; 反之, 当置信度小于低阈值时, 表明当前数据一定不包含所查询对象; 而当置信度在两个阈值之间时, 表明结果不太确定, 系统需要利用通用的神经网络模型作进一步的判断. 然而, NoScope 并不支持两类重要的查询, 即聚合(aggregation)和有限制条件(cardinality-limit)的查询. 因而, 开发 NoScope 系统的研究者们在 NoScope 系统的基础上开发了新的系统 Blazeit. 该系统定义了一种新的查询语言 FrameQL (简称 FQL), 以支持这两类查询. 值得指出的是, Blazeit 处理这两类查询的过程与 NoScope 的处理过程类似.

NoScope 和 Blazeit 这两个系统仅能支持二分类查询, 而 TAHOMA 系统利用级联的思想实现了多分类查询. 如图 8 所示, 该系统将多个专用的神经网络模型连接起来, 分别用于检测不同类别的对象. 在进行分类推理的过程中, 每经过一个专用神经网络模型, 就输出测试数据的置信度. 当置信度高于设定的阈值时, 表明当前分类结果准确, 输出对应的类别标签; 反之, 再通过下一个专用神经网络模型进行推理, 直至得到置信度高的结果.

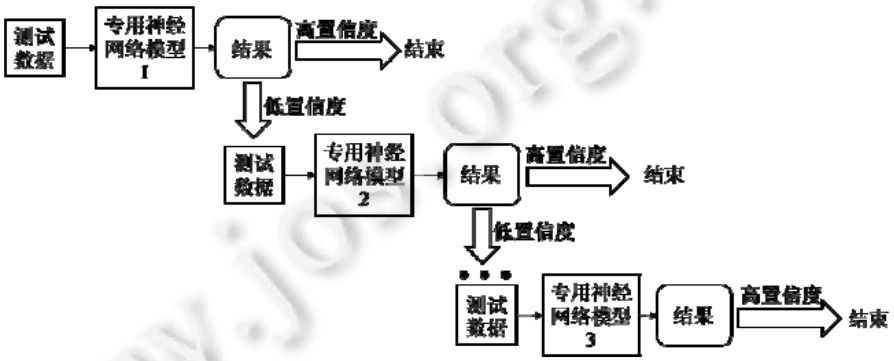

图 8 级联的方法[45]

此外, ODIN 和 Deluceva 系统通过专用化的神经网络模型实现了对象检测. 这类系统在实现对象检测时结合了视频数据存在的时间和空间特性. 具体来说, 当多个视频帧之间的时间间隔较小时, 这些视频帧的相似度较高. 因此, 系统无需处理每一帧的全部内容, 仅需处理相邻帧之间差异的部分以加快推理速度. 基于这一特性, 系统可以训练一种特殊的专用化模型——基于增量的模型. 此外, 当视频出现漂移(drift)现象时, 即视频的场景发生变化时, 视频数据内容与训练数据相差较大, 从而影响了之前的通用化神经网络模型的准确性. 为了尽快得到准确的结果, 系统需要训练专用化的模型来减少因重新训练通用化神经网络模型带来的开销.

Deluceva 系统使用了基于增量的模型来进行对象检测. 在图 9 展示的 Deluceva 系统架构中, 该系统使用视频的第 1 帧来初始化模型的参数. 然后, 对于之后的视频帧, 系统仅将该帧与前一帧数据之间的增量输入到模型中. 但是, 并非所有的增量都会输入到模型中, 系统会根据上一帧的检测结果过滤掉部分增量. 如果检测结果与预期结果相差较小, 系统会过滤掉后续的部分增量, 以加快检测. 然而, 过滤的增量过多时会降低结果的准确度, 因此在对象检测过程中, 系统会动态调整过滤的比率, 以在用户需要的精确度和处理速度之间达到平衡.

ODIN 系统利用专用化的神经网络模型解决了视频中漂移现象所带来的问题, 该系统使用对抗性自动编码器[53]和一种无监督的算法来检测视频中的漂移现象. 当检测到漂移现象时, ODIN 系统调用漂移恢复算法来训练针对新数据的专用化神经网络模型.

除了上述用于对象分类和对象检测的系统外, 这里还存在着利用专用化的神经网络模型实现路径跟踪的系统 MIRIS. 一般来说, 实现路径跟踪的前提是获取相邻帧中所有对象的位置. 这一过程影响了系统的计算时间. 为了加速这一过程, MIRIS 训练了专用化的神经网络模型来对视频进行处理.

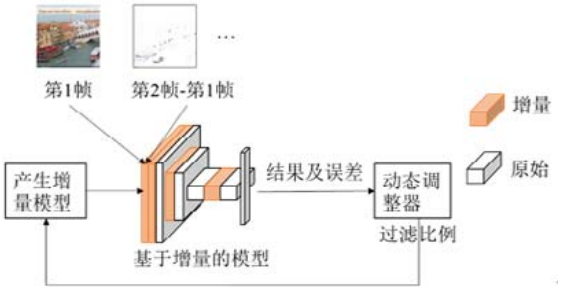

图 9 Deluceva 系统架构[47]

3.1.2 针对数据进行计算优化

针对模型的优化技术通过降低模型推理延迟来加速计算. 然而, 不仅是模型推理延迟, 数据预处理延迟也影响了计算的时间. 一般来说, 高分辨率的视频数据通常会产生较高的预处理与推理延迟. 因此, 现有的新型视频分析系统通常通过降低视频数据的分辨率来降低预处理与推理的延迟. 与此同时, 为了从数据角度进一步降低推理延迟, 现有的新型视频分析系统通过减少模型推理的数据量来解决这一问题.

低分辨率数据一方面避免了预处理过程进行复杂的大小调整、裁剪等工作, 另一方面也降低了推理的负载, 从而加速了计算. 然而, 使用分辨率低的数据会影响计算结果的精确度, 因此系统需要结合一些策略来增强计算结果的精确度. 例如, Smol 系统不仅通过低分辨率数据或者部分解码的数据来加快预处理和推理, 进一步地, 其还使用了通用化的神经网络模型来保证计算结果的精确度, 以弥补低分辨率数据的缺陷. 最终的实验结果表明: 与使用高分辨率数据和专用化的模型相比, 此种方法可以获得更低的计算延迟.

为了进一步降低模型推理的延迟, 现有的新型视频分析系统通常采取“相似过滤”策略来减少模型所需处理的数据量. “相似过滤”策略的目的是, 在模型推理前过滤视频数据中的相似帧. 为了寻找这些相似的视频帧, 现有系统利用了视频数据内容的相似性. 例如, 当不同的视频帧中存在同一类物体时, 这些视频帧的内容拥有较高的相似度(即聚类[54,55]思想); 当视频帧之间的时间间隔较小时, 其内容也拥有较高的相似度(即时间局部性).

Focus 系统用于解决多分类查询, 其利用了聚类思想来寻找某一类别的相似帧. 如图 10 所示, Focus 系统在视频获取阶段过滤掉没有对象的帧[56]. 在过滤完成后, 系统提取剩余视频帧中对象的特征, 并将特征相同的视频帧放入一个集合中. 为了方便且快速地检索这些集合, 系统还创建了对象类别标签到对象集合的索引. 其次, 当用户要查询 X 类的数据时, 系统通过索引找出 X 类所对应的集合, 并得到每一个集合的聚类中心帧. 然后, 系统仅需对该中心帧进行推理, 从而判断这个集合是否为 X 类对应的集合: 若是, 则系统将整个集合包含的视频帧返回给用户.

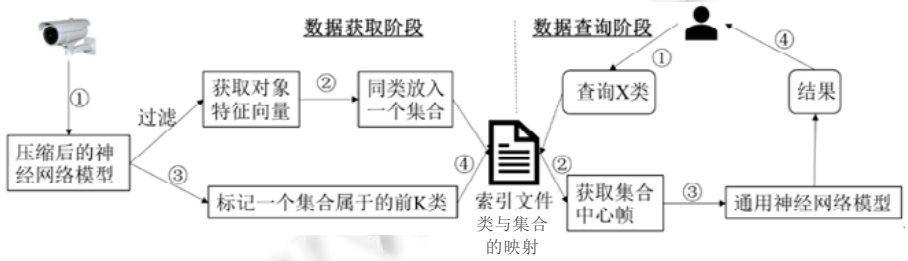

图 10 Focus 系统架构[46]

NoScope 系统针对二分类查询负载, 它利用视频数据的时间局部性来寻找相似帧. 具体来说, 该系统设计了图 7 所示的差异检测器来判断帧与帧之间的差异. 如果当前帧与参考帧相似, 系统就直接输出与参考帧相同的结果; 反之, 系统则将当前帧输入到第 3.1.1 节中提到的专用神经网络中进行推理.

MIRIS 系统用于处理路径跟踪负载. MIRIS 同样利用了视频数据的时间局部性, 但与 NoScope 不同的是, 由于跟踪负载期望获取检测对象的位置变化情况, 因而 MIRIS 无需判断视频帧之间的差异, 它通过采样方式跳过部分相似帧. 如图 11 所示, MIRIS 系统通过降低视频数据采样频率的方法来跳过部分帧, 从而减少神经网络所需处理的数据量. 然后, 系统对采样到的视频帧进行推理, 并获取到视频帧中所有对象的位置. 根据这些位置信息, 系统推理出同一对象可能的路径. 然而, 采样频率的降低影响了推理结果的准确性, 从而导致推理出的路径并不一定是准确的. 为了提高推理的准确性, MIRIS 系统使用两个自定义的过滤器来过滤错误的路径. 在这之后, 系统需要进一步判断过滤后的路径是否是准确的路径. 此时, 系统会提高数据采样的频率, 从而增加神经网络所要处理的数据量, 完成准确路径的确认过程. 在 MIRIS 系统之后, 该系统的研究者在 Vaas[57]系统上演示了 MIRIS 的实现. Vaas 系统除了通过改变采样的频率来加速路径跟踪外, 还进一步让系统支持对象分类和对象检测等查询.

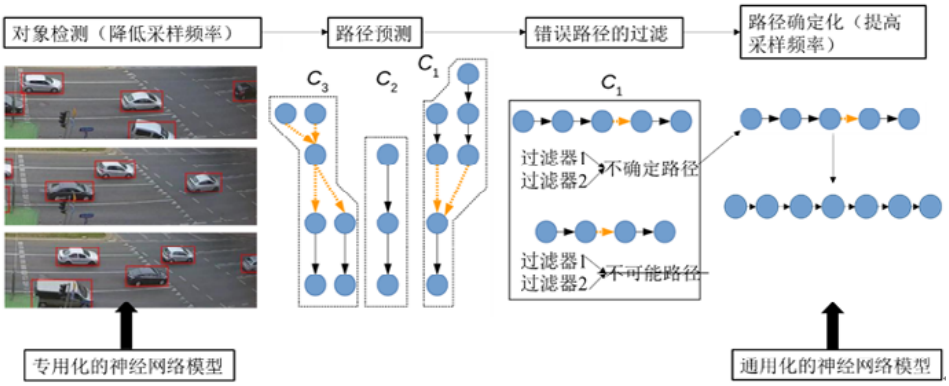

图 11 MIRIS 系统架构[50]

3.1.3 针对资源进行计算优化

不同类型的硬件资源影响着预处理和推理任务的处理延迟. 一般来说, 无论是预处理或是推理任务, 其在 GPU 上的处理延迟均远低于 CPU. 在计算过程中, 预处理与推理任务常以流水线形式进行, 也就是说, 预处理与推理任务在计算过程中会同时执行. 更进一步地, GPU 资源较为稀缺, 因而新型视频分析系统需要合理调度预处理与推理任务以便高效地利用 GPU 资源, 从而降低计算的延迟. 此外, 新型视频分析系统还会面临同时处理多个计算应用(例如, 对象分类、路径跟踪等)的场景. 在这种场景下, 为了平衡多个应用的计算延迟, 避免出现某些应用因无可用资源而一直等待的现象, 新型视频分析通常需要在多个应用间合理分配资源来解决这一问题.

在预处理与推理任务调度方面, 现有的新型视频分析系统常采用静态的任务调度策略. Smol 系统在计算前通过估计预处理和推理的延迟来调度预处理和推理任务的运行位置. 若预处理延迟较高, 则将预处理任务调度至 GPU 上运行; 反之, 则将推理任务调度至 GPU.

在应用间的资源分配方面, 现有新型视频分析系统通常采用动态资源分配策略. 如图 12 所示: Chameleon 系统在运行时, 动态搜索预处理与推理任务中帧率、分辨率和神经网络模型等参数的最佳配置, 为这两个任务分配相应的资源, 并会根据负载的变化改变配置, 从而降低计算延迟.

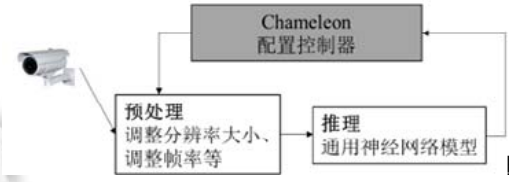

图 12 Chameleon 系统架构[51]

与 Chameleon 系统仅考虑单个计算应用内的资源分配不同, VideoStorm 系统是一个考虑在多个应用间合理分配资源的新型视频分析系统. 图 13 展示了 VideoStorm 系统的架构, VideoStorm 系统在运行时, 动态收集各个应用的运行信息, 并根据这些运行信息来调整各个应用的可用资源. 通过资源动态分配, VideoStorm 系统能够有效平衡各个应用的计算延迟. 例如, 假设此时有两个视频分析应用 *A* 和 *B*, 其中, 应用 *A* 已获取机器 2 的所有资源, 应用 *B* 还未提交, 但其资源需求跨越了机器 1 和机器 2. 当应用 *B* 提交并执行到机器 2 上时, VideoStorm 系统会实时调整应用 *A*、*B* 在机器 2 上的可用资源, 从而避免发生应用 *B* 一直等待的情况.

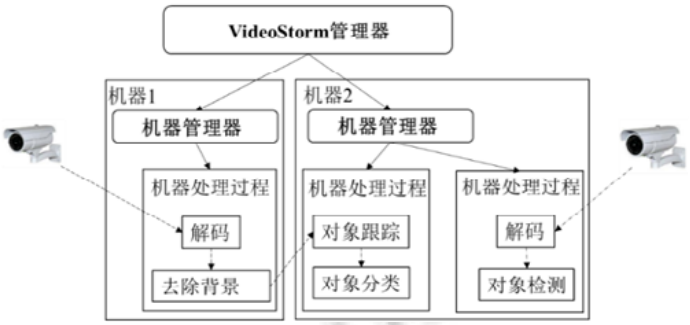

图 13 VideoStorm 系统架构[52]

3.2 针对存储进行优化的系统

上一节从模型、数据以及资源角度分析了针对计算过程进行优化的策略. 然而, 对于新型视频分析系统而言, 除了计算过程外, 如第 2.1 节所述, 系统还需要考虑根据视频应用的特点对存储进行优化. 通常, 存储过程包括两个步骤: 将视频数据采样并保存至存储空间; 将数据从存储空间解码至分析系统. 因此, 现有的视频分析系统一般都从存储空间和数据加载速度两个方面针对存储进行优化.

- 降低占用空间进行存储优化的策略: 新型视频分析系统通过压缩的方法来降低存储空间. 具体地, 系统通过合并或过滤相似的视频帧来减少所需存储的视频数据量, 从而降低数据存储成本. VSS[58]、LightDB[59]、Vignette[60]等系统都采用这种策略来优化存储;
- 提高加载速度进行存储优化的策略: 新型视频分析系统通过缓存、细粒度分割的方法来提高数据解码速度. 通过缓存数据的多种常用格式, 系统能够避免部分繁琐的解码过程. 此外, 为了减少对目标无关数据的解码开销, 系统将视频图像分割为包含目标对象的图块和不包含目标对象的图块, 系统通过仅解码包含目标对象的图块来避免多余的解码开销, 从而提高解码速度. VStore[61]、VSS、TASM[62]、LightDB、VisualCloud[63]、Vignette 等系统都采用了这种策略来优化存储.

3.2.1 降低占用空间进行存储优化

视频数据的日益增长, 要求系统提供更多的存储空间, 使得存储成本不断增加. 为了降低存储成本, 现有的视频分析系统往往采用数据压缩策略, 将视频中相似的数据进行合并或过滤.

VSS 是采用压缩策略的典型系统, 该系统观察到同一地点的多个摄像头捕获到的多个视频画面间会存在相似的部分, 而这些相似的数据会导致冗余存储. 为了缓解这一问题, 如图 14 所示, 该系统利用特征检测算法[55]来识别视频数据中的相似帧, 然后通过平移、旋转以及投影等操作, 将相似帧加以合并. 通过对相似帧进行合并, VSS 系统减少了所需存储的数据量, 从而降低了存储成本.

值得指出的是: 除了摄像头能够产生视频数据外, 现实生活中还存在通过虚拟技术产生的虚拟现实(VR)数据. 随着 VR 技术在城市规划、教育等领域的广泛应用, 系统也需要对 VR 数据进行对象检测、对象分类等操作, 从而支持虚拟现实增强、深度图生成等工作. 例如, 一名用户携带一个 VR 视频感知设备, 该设备在获取实时 VR 视频后, 会通过神经网络模型检测用户当前观看视野中的对象, 并通过矩形框标识这些对象, 从而达到突出显示的目的[59]. 通常, VR 视频数据由三维立体画面组成. 因而, 与传统的二维视频数据相比, VR 视

频数据拥有更多的信息量. 这意味着 VR 视频数据会占用更多的存储空间.

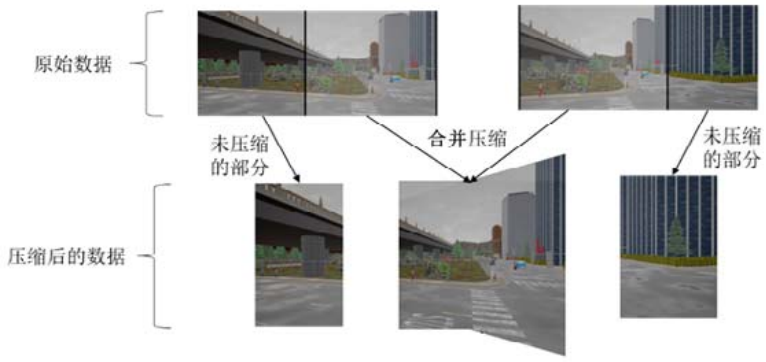

图 14 合并压缩[58]

为了减少 VR 视频数据所需存储空间, LightDB 系统使用投影方法将 VR 数据转化为传统的视频数据, 之后使用二维编码器对这些数据编码. 此外, 如图 15 所示, 该系统还基于相似度检测技术来识别并过滤视频数据中的相似帧. 在 LightDB 的基础上, Vignette 系统对 LightDB 进行了扩展. 该系统采用一种基于感知的压缩方式来对数据进行压缩, 即在获取数据特征的基础上针对数据进行部分采样, 从而过滤掉部分数据, 以达到节约存储空间的目的.

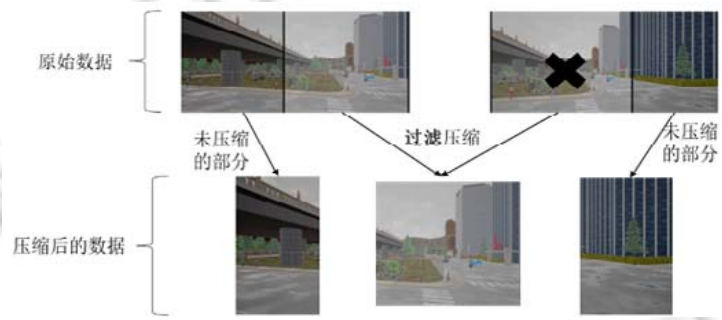

图 15 过滤压缩

3.2.2 提高解码速度进行存储优化

数据的解码速度影响了应用的响应延迟. 为了获得较低的响应延迟, 现有系统通常采用缓存和细粒度分割等方法来提高数据的解码速度. 通常, 存储系统往往以视频数据的默认格式对数据进行存储. 然而, 当用户在系统上运行多类应用时, 不同的应用对于数据的组织结构要求有所不同. 因而, 系统需要反复对数据执行繁琐的解码操作. 为了提高数据的解码速度, 系统可缓存多种视频数据的组织结构形式, 从而减少不必要的解码操作. 除了缓存这一技术外, 系统还可将视频帧分割为不重叠的图块, 然后在解码时仅针对部分包含目标对象的图块进行解码, 从而提高解码速度.

VStore 和 VSS 系统均通过缓存来提高数据的解码速度. 具体来说, 当视频流到达时, VStore 系统会为该视频流生成多种组织结构的数据并进行缓存. 值得注意的是: 缓存会占用一定的存储空间, 无限缓存数据将快速消耗存储空间. 因此, 系统不能无限缓存数据, 而是需要通过一些策略来限制缓存. 因此, VStore 会为这些生成的不同组织结构的数据设定存活时间. 此外, VStore 设定了一个缓存阈值, 当缓存数据量大于该阈值时, VStore 会逐步删除剩余存活时间小的缓存数据. 与 VStore 系统不同的是, VSS 并非在视频流到达时对视频数据进行缓存, 而是在完成视频数据的解码后将解码的数据缓存起来. 与 VStore 系统类似的是, VSS 系统自定义了一种类似 LRU 的策略用于清除缓存. 图 16 展示了一个例子用于说明缓存如何加快解码的速度. 如图 16

所示, m_0 表示一段存储在系统上的 100 mins 视频数据, 该数据按块进行存储. 此外, m_1 和 m_2 分别为 m_0 中分块 2 和分块 4 的缓存. 值得注意的是, m_1、m_2 与 m_0 的分辨率保持一致, 但却采用了不同的格式进行存储, 即 H264. 当系统需要使用 m_0 数据并要求数据格式为 H264 时, 系统可直接载入分块 2 和分块 4 的缓存, 即 m_1 和 m_2, 而无需对完整的 m_0 数据进行解码.

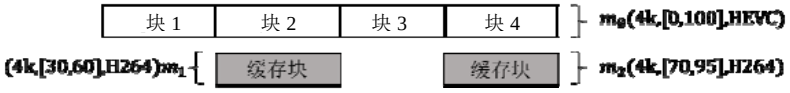

图 16 数据缓存策略

TASM 系统使用了细粒度分割的方法来提高数据的解码速度. 该系统观察到视频数据存在内容局部性, 即当前视频帧上所有的内容并非都与应用相关. 基于这一特性, TASM 在视频流到达时将每个视频帧分割成不重叠的图块, 各个图块中包含了不同的对象. 当需要获取视频数据中的部分对象时, TASM 会过滤掉不包含目标对象的图块, 而只对包含目标对象的图块进行解码, 从而提高数据的解码速度. 图 17 提供了一个单向车道车辆检测的例子来说明细粒度分割是如何提高数据的解码速度的. 如图 17 所示, 该视频帧可以被分割为图块 0 和图块 1, 其中, 图块 1 包含了该侧车道的所有车辆信息. 显然, 在对视频帧进行解码时, 系统可仅对图块 1 进行解码而无需对图块 0 进行解码.

图 17 基于细粒度分割的数据

此外, LightDB 和 VisualCloud 系统也利用了细粒度分割的方法来加快数据的解码速度. 然而, 与上述系统不同的是, LightDB 和 VisualCloud 系统针对 VR 视频数据进行了优化. 这两个系统观察到尽管 VR 视频提供了 360°的观看视角, 但用户在同一时刻仅从某一固定视角来观看 VR 视频. 进一步地, 用户往往倾向于从与前一时刻相近的视角来观看 VR 视频. 例如, 用户在第 5 秒时从 90°的视角观看 VR 视频, 而在第 6 秒时可能会从 80°–100°的视角来观看 VR 视频. 因此, 基于这一特点, 系统可利用方向预测的方法来获取用户接下来一段时间内观看 VR 视频的方向. 然后, 系统可仅针对预测方向的 VR 视频数据进行解码, 从而加快了数据的解码速度.

4 新型视频分析系统总结与分析

上一节从计算优化和存储优化两方面介绍了近年来的基于深度学习的新型视频分析系统. 基于上述内容, 本节首先通过图 18 展示了新型视频分析系统与性能影响因素以及解决方案之间的关系; 然后, 本节从多个维度出发, 总结并分析了针对计算进行优化的系统和针对存储进行优化的系统; 最后, 本节总结了当前系统研究存在的不足, 为之后的研究人员提供一定的参考.

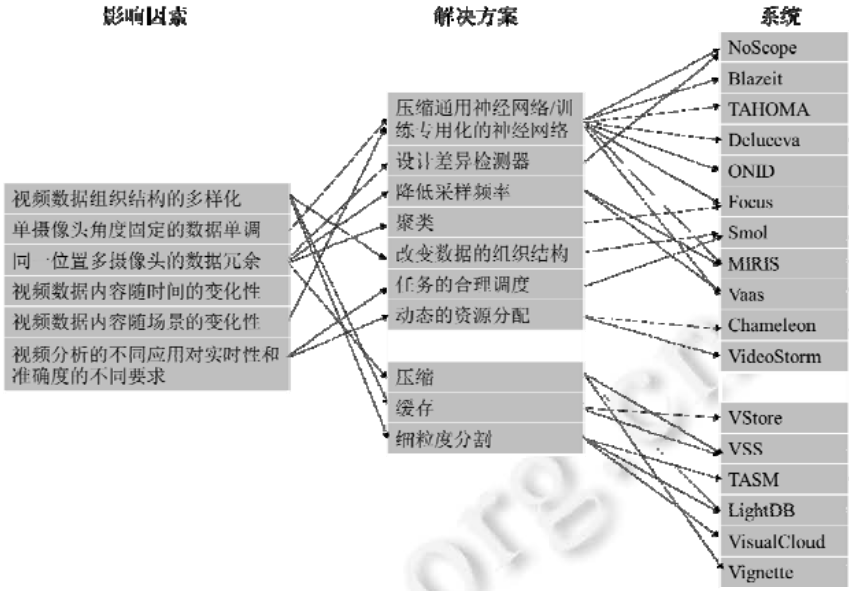

图 18 影响因素-解决方案-系统的关联关系

4.1 针对计算进行优化的系统的总结

表 3 从应用场景、查询类别、运行环境、时效性、开源情况等方面分析了针对计算进行优化的系统, 展示了这些系统之间的共性和差异.

表 3 对面向计算过程优化的新型视频分析系统的总结

系统	应用场景	查询类别		神经网络类别	运行环境	实时性	开源情况
		分类查询	聚合查询				
NoScope	分类	二分类查询	不支持	专用+通用	单机	※	是(https://github.com/stanford-futuredata/noscope)
Blazeit	分类	二分类查询	支持	专用+通用	单机	※	是(https://github.com/stanford-futuredata/blazeit)
TAHOMA	分类	多分类查询	不支持	专用	单机	※	否
Deluceva	检测	–	不支持	专用	单机	※※※	否
ONID	检测	–	不支持	专用	单机	※※※	否
Focus	分类	多分类查询	不支持	专用+通用	单机	※※	否
Smol	分类	二分类查询	支持	通用	单机	※	是(https://github.com/stanford-futuredata/smol)
MIRIS	检测、跟踪	–	不支持	专用	单机	※	是(https://github.com/favyen/miris)
Vaas	分类、检测、跟踪	二分类查询	不支持	专用+通用	单机	※	否
Chameleon	检测、跟踪	–	不支持	通用	单机	※	否
VideoStorm	分类、检测、跟踪	二分类查询	支持	通用	分布式	※※※	否

- 应用场景. 上述针对计算进行优化的系统涉及到的应用包括 3 类, 分别为对象检测、对象分类和对象跟踪: 对象分类应用查询对象所属的类别; 对象检测应用查询对象的边界框位置; 对象跟踪应用往往与对象检测应用相关联, 用于查询对象边界框随时间变化的序列. 大部分系统只处理其中一类应用, 小部分系统可以处理这 3 类应用. 例如, NoScope、Blazeit、Focus、TAHOMA、Smol 系统只能用于对象分类; MIRIS、Chameleon 系统只能用于对象检测和路径跟踪; 仅 MultiScope、VideoStorm、Vaas 系统适用于 3 类应用场景;
- 查询类别. 上述针对计算进行优化的系统主要实现了两类查询, 分别为分类和聚合查询. 分类查询又可以分为二分类查询和多分类查询. 部分实现对象分类的系统, 包括 NoScope、Blazeit、Smol、

VideoStorm 等系统, 一次配置只能识别出单个类别对应的对象, 即回答当前帧是否是某一类对象; 另一部分系统则可以一次识别多个类别对象的数据. 例如, Focus、TAHOMA 系统. 除了分类查询外, Blazeit、Smol、VideoStorm 系统还可以回答聚合查询;

- 运行环境. 上述针对计算进行优化的系统有两种不同的运行环境, 分别为单机和分布式. 大部分系统, 例如 NoScope、Blazeit、Focus、TAHOMA、Smol、Deluceva、MIRIS、ONID 等, 都是在单机环境下运行, 仅 VideoStorm 系统支持在分布式下运行;
- 实时性. 实时性是视频分析系统的一个重要特性, 它代表了系统处理查询任务的速度. 基于不同的实现, 不同系统的应用时延也有所不同. 具体来说, 一类系统用来回答“事后查询”, 诸如 NoScope、Blazeit、Smol、Deluceva、MIRIS、ONID 等系统, 通过统计过去一段时间内的数据进行未来的决策, 实时性不高; 另一类系统则可实时处理视频数据, 例如 VideoStorm、Deluceva、Focus 等系统. 这类系统多用于实时交通规划、实时计费等场景中, 但实时性也存在差别;
- 开源情况. 开源使全球信息技术领域发生了全局性、持续性的重大变化, 在社会基础设施建设方面也发挥着越来越重要的作用. 在上述针对计算进行优化的系统中, NoScope、Blazeit、Smol、MIRIS 系统实现了开源.

4.2 针对存储进行优化的系统的总结

表 4 从数据存储的形式、数据存储的位置、存储空间的优化、数据解码的优化、数据读取的性能等方面分析了针对存储进行优化的系统, 展示了这些系统之间的共性和差异.

表 4 对面向存储管理优化的新型视频分析系统的总结

系统	数据存储的形式	数据存储的位置	存储空间的优化	数据加载的优化	数据读取的性能
VStore	原始帧	本地	未优化	缓存	※
VSS	物理块(block)	本地	压缩优化	缓存	※※
TASM	图块(tail)	本地	未优化	细粒度分割	※
LightDB	图块(tail)	本地	压缩优化	细粒度分割	※※
VisualCloud	图块(tail)	本地	未优化	细粒度分割	※
Vignette	图块(tail)	云端	压缩优化	细粒度分割	※

- 数据存储的形式. 上述针对存储进行优化的系统存储视频数据的形式有 3 种, 分别为帧存储、物理块(block)存储和图块(tail)存储. 具体来说, VStore 系统按帧存储数据, VSS 系统以物理块的形式存储数据, TASM、LightDB、VisualCloud、Vignette 系统以图块的形式存储数据. 其中, 物理块并不会破坏完整的视频帧, 它仍然是一段连续时间的帧序列, 而图块相当于将视频的每一帧分割为不重叠的部分;
- 数据存储的位置. 上述针对存储进行优化的系统存储视频的位置有两个, 分别为本地存储和云端存储. VStore、VSS、TASM、LightDB、VisualCloud 的数据存放在本地, Vignette 的数据存放在云端;
- 存储空间的优化. 上述针对存储进行优化的系统通过压缩的方法缩小数据存储所需的空间. 具体来说, VSS 系统支持对相似的数据进行合并, LightDB 系统支持对相似的数据进行过滤, Vignette 系统支持对信息量小的数据进行基于感知的过滤. 这 3 个系统的不同在于: VStore 和 TASM 系统是对采样后的数据进行处理, 而 Vignette 是在采样的过程中就完成了对数据的处理. VStore、TASM、VisualCloud 系统则未考虑对相似数据的存储优化问题;
- 数据解码的优化. 上述针对存储进行优化的系统通过缓存和细粒度分割的方法降低数据解码的代价. 具体来说, VStore、VSS 系统通过缓存的方法存放组织结构不同的数据来避免反复的解码操作. TASM、LightDB、VisualCloud、Vignette 系统通过细粒度划分的方法, 只解码与用户查询相关的数据内容, 从而减少了解码的数据量;
- 数据读取的性能. 上述针对存储进行优化的系统对两类数据的读取进行了优化. 具体来说, 当用户从系统中读取现实生活中摄像头捕获的数据时, VStore 系统不支持读取 HEVC、H264 等格式. VSS 和

TASM 系统则支持任意格式数据的读取. 更进一步地, VSS 提供了方便用户对数据操作的 API, VStore 系统提供了多个常用算子的接口. 但是, 由于 VStore 数据配置的形式, VStore 系统读取数据的质量不如 VSS 和 TASM 系统. 因此, 总体上, VSS 系统读取数据的性能要优于 TASM 和 VStore 系统. 对于 VR 数据, LightDB 提供了多个数据操作的接口和声明式的 VRQL 查询语言, VisualCloud、Vignette 则未执行相关优化. 因此, 总体上, LightDB 系统读取数据的性能要优于 VisualCloud 和 Vignette 系统.

4.3 当前系统存在的不足

近年来, 基于深度学习的新型视频分析系统不断涌现, 但相关研究方兴未艾. 这些研究的不足之处主要表现在以下 3 个方面.

(1) 计算与存储的优化技术分离. 通过前述分析可以看出, 现有的新型视频分析系统针对计算进行优化或针对存储进行优化. 因此, 当前新型视频分析系统的优化技术呈现计算与存储的优化分离, 同时, 对计算和存储进行优化的系统亟待研究;

(2) 横向扩展能力有限. 现有的新型视频分析系统大部分仅支持单机运行, 缺乏横向可扩展能力. 单机部署意味着所有应用业务均在一台服务器上, 因而对服务器的配置要求极高且价格昂贵. 当业务规模扩大到一定程度时, 单台服务器可能难以承受, 影响了系统的正常运行. 此外, 一旦这台服务器宕机, 将产生严重的后果;

(3) 支持的查询类别单一. 从表 3 可以看出, 当前的新型视频分析系统通常仅支持分类查询, 查询类别相对简单. 为了进一步统计或分析查询结果, 往往需要用户进行编程实现. 此外, 单一的查询类别也限制了系统对复杂场景的支持.

5 新型视频分析系统未来的研究方向

新型视频分析系统具有广泛的应用前景, 本节简要讨论新型视频分析系统未来的研究方向.

(1) 数据处理与新型视频分析技术的深度融合. 大规模数据处理平台具有横向高可扩展能力, 针对计算和存储均进行了优化, 支持丰富的查询接口, 这些特性为构建新型视频分析系统提供了有力的技术支撑. 如何结合新型视频分析应用的负载特征, 量身定制新型视频分析系统, 是未来值得关注的研究方向. 此外, 像评价数据处理系统性能那样制定公认的评测基准及数据集, 有助于推进新型视频分析系统的性能测评;

(2) 新硬件赋能的新型视频分析加速技术. 现有的新型视频分析系统普遍采用了 GPU 硬件进行推理, 显存容量可能会限制并行处理的视频帧数量, 阻碍了性能的提升. 因此, 如何在显存受限情况下提升系统性能是研究的难点之一. 例如, 可以研究如何使用 NVLink 等新型高速互联通道加速数据传输, 充分利用新硬件的优势, 是未来值得关注的研究方向. 此外, 如何利用 TPU 等新型硬件来加速新型视频分析, 也是未来的发展趋势;

(3) 云端协同场景下的新型视频分析系统构建. 当前, 大规模数据处理应用往往都部署在云平台上, 视频分析系统可充分利用分布式计算平台的并行计算能力[64]及云平台的计算资源弹性分配机制[65]. 如何优化数据分区操作以及多台机器之间的通信管理等, 将是影响系统性能的关键因素. 除了云平台以外, 终端设备的计算能力正在逐步增强, 边缘计算正得到广泛关注. 在视频分析的应用场景中, 如果分析处理能够部分或全部在终端完成, 那么数据向云平台传输的开销将大大降低. 因此, 如何利用云端协同来构建新型视频分析系统, 是未来值得关注的研究方向.

6 结束语

本文首先回顾了视频分析系统的发展历程, 着重指出了传统视频分析系统的缺点, 突出了基于深度学习的新型视频分析系统优势. 然后, 本文分析了目前新型视频分析系统面临的挑战以及对系统性能的影响因素, 并且详细介绍了现有新型视频分析系统如何根据影响因素来解决前述挑战. 最后, 本文从不同维度对比总结

了新型视频分析系统，指出了现有系统的不足，并展望了新型视频分析系统未来的研究方向.

References:

[1] Cisco. The all bytes era: Trends and analysis. 2017 (in Chinese). https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html

[2] Cao SD, Hua Y, Feng D, Sun YY, Zuo PF. High-performance distributed storage system for large-scale high-definition video data. Ruan Jian Xue Bao/Journal of Software, 2017, 28(8): 1999−2009 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5203.htm [doi: 10.13328/j.cnki.jos.005203]

[3] Buch N, Velastin SA, Orwell J. A review of computer vision techniques for the analysis of urban traffic. IEEE Trans. on Intelligent Transportation Systems, 2011, 12(3): 920−939. [doi: 10.1109/TITS.2011.2119372]

[4] Tang Z, Naphade M, Liu MY, Yang X, Birchfield S, Wang S, Kumar R, Anastasiu D, Hwang JN. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: Proc of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. Los Angeles: IEEE, 2019. 8789−8798. [doi: 10.1109/CVPR.2019.00900]

[5] Shirazi MS, Morris BT. Vision-based turning movement monitoring: Count, speed & waiting time estimation. IEEE Intelligent Transportation Systems Magazine, 2016, 8(1): 23−34. [doi: 10.1109/MITS.2015.2477474]

[6] Jia SJ, Hu SP, Yang MZ, Liu ST. Indoor target anomaly detection based on Dense_YOLO. Journal of Dalian Jiaotong University, 2019, 40(3): 102−107 (in Chinese with English abstract). [doi: 10.13291/j.cnki.djdxac.2019.03.020]

[7] Kang D, Bailis P, Zaharia M. Challenges and opportunities in DNN-based video analytics: A demonstration of the Blazeit video query engine. In: Proc. of the 9th Biennial Conf. on Innovative Data Systems Research. 2019. http://cidrdb.org/cidr2019/papers/p141-kang-cidr19.pdf

[8] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84−90. [doi: 10.1145/3065386]

[9] Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, Gorkani M, Hafher J, Lee D, Petkovie D, Steele D, Yanker P. Query by image and video content: The QBIC system. Computer, 1995, 28(9): 23−32. [doi: 10.1109/2.410146]

[10] Hampapur A, Gupta A, Horowitz B, Shu CF, Fuller C, Bach JR, Gorkani M, Jain RC. Virage video engine. In: Proc. of the Storage and Retrieval for Image and Video Databases V: Int'l Society for Optics and Photonics. San Jose: SPIE, 1997. 188−198. [doi: 10.1117/12.263407]

[11] Pentland A, Picard RW, Sclaroff S. Photobook: Content-based manipulation of image databases. Int'l Journal of Computer Vision, 1996, 18(3): 233−254. [doi: 10.1117/12.171786]

[12] Minka T. An image database browser that learns from user interaction [Ph.D. Thesis]. Cambridge: Massachusetts Institute of Technology, 1996.

[13] Smith JR, Chang SF. VisualSEEk: A fully automated content-based image query system. In: Proc. of the 4th ACM Int'l Conf. on Multimedia. New York: ACM, 1997. 87−98. [doi: 10.1145/244130.244151]

[14] Petkovic M, Jonker W. A framework for video modelling. In: Proc. of the Int'l Conf. on Applied Informatics. Innsbruck: Springer, 2000.

[15] Aguierre STG, Davenport G. The stratification system: A design environment for random access video. In: Proc. of the Int'l Workshop on Network and Operating System Support for Digital Audio and Video. Cambridge: ACM, 1992. 250−261.

[16] Jiang H, Elmagarmid AK. Spatial and temporal content-based acces to hypervideo databases. VLDB Journal, 1998, 7(4): 226−238. [doi: 10.1007/s007780050066]

[17] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science, 2015, 349(6245): 255−260. [doi: 10.1016/j.compbiomed.2019.02.017]

[18] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436−444. [doi: 10.1038/nature14539]

[19] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks, 2015, 61: 85−117. [doi: 10.1016/j.neunet.2014.09.003]

[20] Lu F, Liu CH, Huang CY, Yang Y, Xie Y, Liu CX. Overview on deep learning-based object detection. Computer Systems & Applications, 2021, 30(3): 1−13 (in Chinese with English abstract). http://www.c-s-a.org.cn/1003-3254/7839.html [doi: 10.15888/ j.cnki.csa.007839]

[21] Zhou FY, Jin LP, Dong J. Review of convolutional neural network. Chinese Journal of Computers, 2017, 40(6): 1229−1251 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2017.01229]

[22] Dubey SR. A decade survey of content based image retrieval using deep learning. IEEE Trans. on Circuits and Systems for Video Technology, 2021, 8215(c): 1−17. [doi: 10.1109/TCSVT.2021.3080920]

[23] Wang Z, Yan M, Liu S, Chen JJ, Zhang DD, Wu Z, Chen X. Survey on testing of deep neural networks. Ruan Jian Xue Bao/Journal of Software, 2020, 31(5): 1255−1275 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5951.htm [doi: 10.13328/j.cnki.jos.005951]

[24] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1−9. [doi: 10.1109/CVPR.2015.7298594]

[25] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770−778. [doi: 10.1109/CVPR.2016.90]

[26] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779−788. [doi: 10.1109/CVPR.2016.91]

[27] Song J, Xiao L, Lian ZC, Cai ZY, Jiang GP. Overview and prospect of deep learning for image segmentation in digital pathology. Ruan Jian Xue Bao/Journal of Software, 2021, 32(5): 1427−1460 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6205.htm [doi: 10.13328/j.cnki.jos.006205]

[28] Abadi M, Agarwal A, Barham P, Brevdo E, Chen ZF, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia YQ, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng XQ. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv: 1603.04467, 2016.

[29] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. Orlando: ACM, 2014. 675−678. [doi: 10.1145/2647868.2654889]

[30] Taigman Y, Yang M, Ranzato M, Wolf L. Deepface: Closing the gap to human-level performance in face verification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1701−1708. [doi: 10.1109/CVPR.2014.220]

[31] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 815−823. [doi: 10.1109/CVPR.2015.7298682]

[32] Kim S, Dalmia S, Metze F. Gated embeddings in end-to-end speech recognition for conversational-context fusion. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 1131−1141. [doi: 10.18653/v1/P19-1107]

[33] Hosseini-Kivanani N, Vasquez-Correa JC, Stede M, Nöth E. Automated cross-language intelligibility analysis of Parkinson's disease patients using speech recognition technologies. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence: ACL, 2019. 74−80. [doi: 10.18653/v1/P19-2010]

[34] Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 1412−1421. [doi: 10.18653/v1/D15-1166]

[35] Wang Z, Tan Y, Zhang M. Graph-based recommendation on social networks. In: Proc. of the 12th Asia-Pacific Web Conf. (APWeb 2010). Busan: IEEE, 2010. 116−122. [doi: 10.1109/APWeb.2010.60]

[36] Huang LW, Jiang BT, Lv SY, Liu YB, Li DY. Survey on deep learing based recommender systems. Chinese Journal of Computers, 2018, 41(7): 1619−1647 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2018.01619]

[37] Sun ZJ, Xue L, Xu YM, Wang Z. Overview of deep learning. Application Research of Computers, 2012, 29(8): 2806−2810 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-3695.2012.08.002]

[38] Wiegand T, Sullivan GJ, Bjontegaard G, Luthra A. Overview of the H.264/AVC video coding standard. IEEE Trans. on Circuits and Systems for Video Technology, 2003, 13(7): 560−576. [doi: 10.1109/TCSVT.2003.815165]

[39] Sullivan GJ, Ohm JR, Han WJ, Wiegand T. Overview of the high efficiency video coding (HEVC) standard. IEEE Trans. on Circuits and Systems for Video Technology, 2012, 22(12): 1649−1668. [doi: 10.1109/TCSVT.2012.2221191]

[40] Jia CM, Ma HC, Yang WH, Ren WQ, Pan JS, Liu D, Liu JY, Ma SW. Video processing and compression technologies. Journal of Image and Graphics, 2021, 26(6): 1179−1200 (in Chinese with English abstract). http://kns.cnki.net/kcms/detail/detail.aspx?FileName=ZGTB202106001&DbName=CJFQ2021

[41] AMBER Alert, U.S. Department of justice. http://www.amberalert.gov/faqs.htm

[42] SR 520 Bridge Tolling, WA. https://www.wsdot.wa.gov/Tolling/520/default.htm

[43] Kang D, Emmons J, Abuzaid F, Bailis P, Zaharia M. NoScope: Optimizing neural network queries over video at scale. arXiv: 1703.02529v3, 2017.

[44] Kang D, Bailis P, Zaharia M. Blazelt: Optimizing declarative aggregation and limit queries for neural network based video analytics. Proc. of the VLDB Endowment, 2019, 13(4): 533−546. [doi: 10.14778/3372716.3372725]

[45] Anderson MR, Cafarella M, Ros G, Wenisch TF. Physical representation-based predicate optimization for a visual analytics database. In: Proc. of the 35th IEEE Int'l Conf. on Data Engineering. Macau SAR: IEEE, 2019. 1466−1477. [doi: 10.1109/ICDE.2019.00132]

[46] Hsieh K, Ananthanarayanan G, Bodik P, Venkataraman S, Bahl P, Philipose M, Gibbons PB, Mutlu O. Focus: Querying large video datasets with low latency and low cost. In: Proc. of the 13th Symp. on Operating Systems Design and Implementation. Carlsbad: USENIX, 2018. 269−286. https://www.usenix.org/system/files/osdi18-hsieh.pdf

[47] Wang J, Balazinska M. Deluceva: Delta-based neural network inference for fast video analytics. In: Proc. of the 32nd Int'l Conf. on Scientific and Statistical Database Management. Vienna: IEEE, 2020. 1−12. [doi: 10.1145/3400903.3400930]

[48] Suprem A, Arulraj J, Pu C, Ferreira J. ODIN: Automated drift detection and recovery in video analytics. Proc. of the VLDB Endowment, 2020, 13(11): 2453−2465. [doi: 10.14778/3407790.3407837]

[49] Kang D, Mathur A, Veeramacheneni T, Bailis P, Zaharia M. Jointly optimizing preprocessing and inference for DNN-based visual analytics. Proc. of the VLDB Endowment, 2020, 14(2): 87−100. [doi: 10.14778/3425879.3425881]

[50] Bastani F, He S, Balasingam A, Gopalakrishnan K, Alizadeh M, Balakrishnan H, Cafarella M, Kraska T, Madden S. MIRIS: Fast object track queries in video. In: Proc. of the 2020 ACM SIGMOD Int'l Conf. on Management of Data. Portland: ACM, 2020. 1907−1921. [doi: 10.1145/3318464.3389692]

[51] Jiang J, Ananthanarayanan G, Bodik P, Sen S, Stoica I. Chameleon: Scalable adaptation of video analytics. In: Proc. of the 2018 Conf. of the ACM Special Interest Group on Data Communication. Budapest: ACM, 2018. 253−266. [doi: 10.1145/3230543.3230574]

[52] Zhang H, Ananthanarayanan G, Bodik P, Philipose M, Bahl P, Freedman MJ. Live video analytics at scale with approximation and delay-tolerance. In: Proc. of the 14th Symp. on Networked Systems Design and Implementation. Boston: USENIX, 2017. 377−392. https://www.usenix.org/system/files/conference/nsdi17/nsdi17-zhang.pdf

[53] Makhzani A, Frey B. Pixelgan autoencoders. In: Proc. of the Advances in Neural Information Processing Systems 30: Annual Conf. Neural Information Processing Systems 2017. Long Beach: MIT, 2017. 1975−1985. https://arxiv.org/pdf/1706.00531.pdf

[54] Cao F, Estert M, Qian W, Zhou AY. Density-based clustering over an evolving data stream with noise. In: Proc. of the 2006 SIAM Int'l Conf. on Data Mining. Bethesda: Society for Industrial and Applied Mathematics, 2006. 328−339. [doi: 10.1137/1.9781611972764.29]

[55] O'Callaghan L, Mishra N, Meyerson A, Guha S, Motwani R. Streaming-data algorithms for high-quality clustering. In: Proc. of the 18th Int'l Conf. on Data Engineering. San Jose: IEEE, 2002. 685−694. [doi: 10.1109/ICDE.2002.994785]

[56] Cropley J. Top video surveillance trends for 2016. Technical Report, 2016. https://www.scati.com/en/news/top-video-surveillance-trends-for-2016_326.html

[57] Bastani F, Moll O, Madden S. VAAS: Video analytics at scale. Proc. of the VLDB Endowment, 2020, 13(12): 2877−2880. [doi: 10.14778/3415478.3415498]

[58] Haynes B, Daum M, He D, Mazumdar A, Balazinska M, Cheung A, Ceze L. VSS: A storage system for video analytics. In: Proc. of the 2021 ACM SIGMOD Int'l Conf. on Management of Data. Xi'an: ACM, 2021. 685−696. [doi: 10.1145/3448016.3459242]

[59] Haynes B, Mazumdar A, Balazinska M, Ceze L, Cheung A. LightDB: A DBMS for virtual reality video. Proc. of the VLDB Endowment, 2018, 11(10). [doi: 10.14778/3231751.3231768]

[60] Mazumdar A, Haynes B, Balazinska M, Ceze L, Cheung A, Oskin M. Perceptual compression for video storage and processing systems. In: Proc. of the ACM Symp. on Cloud Computing. Santa Cruz: ACM, 2019. 179−192. [doi: 10.1145/3357223.3362725]

[61] Xu T, Botelho LM, Lin FX. VStore: A data store for analytics on large videos. In: Proc. of the 14th EuroSys Conf. 2019. Dresden: ACM, 2019. 1−17. [doi: 10.1145/3302424.3303971]

[62] Daum M, Haynes B, He D, Mazumdar A, Balazinska M. TASM: A tile-based storage manager for video analytics. In: Proc. of the 37th IEEE Int'l Conf. on Data Engineering. Chania: IEEE, 2021. 1775−1786. [doi: 10.1109/icde51399.2021.00156]

[63] Haynes B, Minyaylov A, Balazinska M, Ceze L, Cheung A. Visualcloud demonstration: A DBMS for virtual reality. In: Proc. of the 2017 ACM Int'l Conf. on Management of Data. Chicago: ACM, 2017. 1615−1618. [doi: 10.1145/3035918.3058734]

[64] Rashid ZN, Zebari SRM, Sharif KH, Jacksi K. Distributed cloud computing and distributed parallel computing: A review. In: Proc. of the 2018 Int'l Conf. on Advanced Science and Engineering. Berlin: Springer, 2018. 167−172. [doi: 10.1109/ICOASE.2018.8548937]

[65] Sahni J, Vidyarthi DP. Heterogeneity-aware elastic scaling of streaming applications on cloud platforms. The Journal of Supercomputing, 2021, 1−28. [doi: 10.1007/s11227-021-03692-w]

附中文参考文献:

[1] 思科. 皆字节时代: 趋势与分析. 2017. https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html

[2] 操顺德, 华宇, 冯丹, 孙园园, 左鹏飞. 面向海量高清视频数据的高性能分布式存储系统. 软件学报, 2017, 28(8): 1999−2009. http://www.jos.org.cn/1000-9825/5203.htm [doi: 10.13328/j.cnki.jos.005203]

[6] 贾世杰, 胡斯平, 杨明珠, 刘舒婷. 基于 Dense_YOLO 的室内目标异常检测. 大连交通大学学报, 2019, 40(3): 102−107. [doi: 10.13291/j.cnki.djdxac.2019.03.020]

[20] 陆峰, 刘华海, 黄长缨, 杨艳, 谢禹, 刘财喜. 基于深度学习的目标检测技术综述. 计算机系统应用, 2021, 30(3): 1−13. http://www.c-s-a.org.cn/1003-3254/7839.html [doi: 10.15888/j.cnki.csa.007839]

[21] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 1229−1251. [doi: 10.11897/SP.J.1016.2017.01229]

[23] 王赞, 闫明, 刘爽, 陈俊洁, 张栋迪, 吴卓, 陈翔. 深度神经网络测试研究综述. 软件学报, 2020, 31(5): 1255−1275. http://www.jos.org.cn/1000-9825/5951.htm [doi: 10.13328/j.cnki.jos.005951]

[27] 宋杰, 肖亮, 练智超, 蔡子赟, 蒋国平. 基于深度学习的数字病理图像分割综述与展望. 软件学报, 2021, 32(5): 1427−1460. http://www.jos.org.cn/1000-9825/6205.htm [doi: 10.13328/j.cnki.jos.006205]

[36] 黄立威, 江碧涛, 吕守业, 刘艳博, 李德毅. 基于深度学习的推荐系统研究综述. 计算机学报, 2018, 41(7): 1619−1647. [doi: 10.11897/SP.J.1016.2018.01619]

[37] 孙志军, 薛磊, 许阳明, 王正. 深度学习研究综述. 计算机应用研究, 2012, 29(8): 2806−2810. [doi: 10.3969/j.issn.1001-3695.2012.08.002]

[40] 贾川民, 马海川, 杨文瀚, 任文琦, 潘金山, 刘东, 刘家瑛, 马思伟. 视频处理与压缩技术. 中国图像图形学报, 2021, 26(6): 1179−1200. http://kns.cnki.net/kcms/detail/detail.aspx?FileName=ZGTB202106001&DbName=CJFQ2021

孟令睿(1998－), 女, 硕士生, CCF 学生会员, 主要研究领域为大规模数据处理系统.

丁光耀(1996－), 男, 博士生, 主要研究领域为大规模数据处理系统.

徐辰(1988－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为大规模数据处理系统, 分布式机器学习系统, 面向新硬件的数据管理技术.

钱卫宁(1976－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为可扩展事务处理, 大数据管理系统基准评测, 海量数据分析处理及其应用, 数据驱动的计算教育学.

周傲英(1965－), 男, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为数据库, 数据管理, 数据驱动的计算教育学, 教育科技、物流科技等基于数据的应用科技.