智能视频监控技术综述

黄凯奇 陈晓棠 康运锋 谭铁牛

(中国科学院自动化研究所模式识别国家重点实验室&智能感知与计算研究中心 北京 100190)

摘 要 随着摄像头安装数量的日益增多,以及智慧城市和公共安全需求的日益增长,采用人工的视频监控方式已经远远不能满足需要,因此智能视频监控技术应运而生并迅速成为一个研究热点.智能视频监控技术是一个跨领域的研究方向,它的研究内容丰富,应用领域广泛多样.文中对智能视频监控技术的发展历史、研究现状以及典型算法的现状给了比较全面的综述.首先从底层、中层、高层对智能视频监控技术进行分类,分别对目标检测、目标跟踪、分类识别以及行为分析算法进行归纳总结;然后对典型算法的优缺点进行分析,给出了典型算法在现有研究数据库上的性能对比,并对待解决问题和难点进行了总结;最后对智能视频监控技术在物联网背景下存在的挑战以及未来发展趋势进行了探讨.

关键词 智能视频监控;智慧城市;公共安全;物联网

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2015.01093

Intelligent Visual Surveillance: A Review

HUANG Kai-Qi CHEN Xiao-Tang KANG Yun-Feng TAN Tie-Niu

(*National Laboratory of Pattern Recognition & Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

Abstract Due to the rapid increase of the number of cameras used in the video surveillance and the huge needs of the smart city and public security, video surveillance by human beings is no longer suitable. Hence, intelligent video surveillance emerges and becomes one of the hottest research points. Intelligent video surveillance is an interdisciplinary research area that has abundant research interests and diverse applications. This paper summarizes the history, the state-of-the-art, and various applications of the intelligent videosurveillance. Firstly, this paper classifies the algorithms by low level, middle level and high level, then analyses their advantages and disadvantages, and compares the performances on different datasets, as well as presents the outstanding issues; finally, we discuss the future research trends of intelligent video surveillance in the context of Internet of Things.

Keywords intelligent videosurveillance; smart city; publicsecurity; Internet of Things

1 引 言

当今社会,人口众多,存在众多安全隐患.随着人们对安全性要求的提高以及经济条件的改善,监控摄像头的个数增长速度越来越快,覆盖的范围也越来越广.传统的视频监控仅提供视频的捕获、存储和回放等简单的功能,用来记录发生的事情,很难起

收稿日期:2013-09-08;最终修改稿收到日期:2015-03-16.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316302)、国家自然科学基金(61322209)、国家科技支撑计划(2012BAH07B01)资助.黄凯奇,男,1977年生,博士,研究员,国家自然科学基金优秀青年基金获得者,中国计算机学会(CCF)高级会员,曾任IEEE北京分会副秘书长,主要研究领域为计算机视觉、模式识别、智能视觉监控.E-mail:kqhuang@nlpr.ia.ac.cn.陈晓棠,女,1987年生,博士,助理研究员,主要研究方向为计算机视觉、模式识别.康运锋,男,1981年生,工程师,主要研究方向为智能视频监控应用.谭铁牛,男,1964年生,博士,研究员,中国科学院院士,主要研究领域为生物特征识别、智能视频监控、网络数据理解与安全.

到预警和报警的作用.若要保证实时监控异常行为并及时采取有效措施,就需要监控人员一刻不停的监看视频,这种情况下,监控人员容易疲惫,尤其面对多路监控视频时,往往目不暇接,很难及时对异常做出反应.因此这就迫切需要智能视频监控,来辅助监控人员的工作.

众多的摄像头,庞大的监控网络,瞬间就会产生海量视频数据,如何从这些海量数据中高效地提取出有用的信息,就成为智能视频监控技术要解决的问题.具体地讲,智能视频监控技术就是为了让计算机像人的大脑,让摄像头像人的眼睛,由计算机智能地分析从摄像头中获取的图像序列,对被监控场景中的内容进行理解,实现对异常行为的自动预警和报警.

20 世纪末以来,随着计算机视觉的发展,智能视频监控技术得到广泛的关注和研究,并随着安全的日益重视,也成为当前的研究热点.智能视频监控包括在底层上对动态场景中的感兴趣目标进行检测、分类、跟踪和识别,在高层上对感兴趣目标的行为进行识别、分析和理解.智能视频监控技术可以广泛应用于公共安全监控、工业现场监控、居民小区监控、交通状态监控等各种监控场景中,实现犯罪预防、交通管制、意外防范和检测、老幼病残监护等功能,能够显著提高监控效率,降低监控成本,具有广泛的研究意义和应用前景.

目前已有较多工作对智能视频监控技术的各方面进行总结和阐述,Bouwmans 等人[1-2]从背景建模及行人检测方面对目标检测技术进行了介绍,Yilmaz 等人[3]、Wang[4]和 Wu 等人[5]从单摄像机和多摄像机跟踪方面对目标跟踪算法进行较为详细的介绍,Huang 等人[6]、Andreopoulos 等人[7]、Zhang 等人[8]对图像中目标分类识别算法进行了介绍,行为识别得到了较多的关注,Hu 等人[9]、Morris 等人[10]和 Aggarwal 等人[11]多次对行为识别算法及相关数据库工作进行综述性的介绍.随着研究的进展,相继有很多智能视频监控系统被开发[12-16],如早期卡内基梅隆大学开发的 VSAM(Visual Surveillance and Monitoring)系统[12]、英国雷丁大学和 INRIA 等多个研发机构合作研究的 ADVISOR(Annotated Digital Video for Intelligent Surveillance and Optimized Retrieval)系统[13]、IBM 开发的 SSS(Smart Surveillance System)系统[14]、中佛罗里达大学研发的 Knight 系统[15]、中国科学院自动化研究所研发的 Vstar 系统等[16].不少学者也对智能视频监控系统进行了较为全面的介绍,如 Huang 等人[16]、Hu 等人[17]和 Valera 等人[18],在文章中他们对系统中涉及到的多个部分的算法进行了介绍.和算法及系统相对的是,在技术应用方面,目前还是处于尝试阶段,Valera 等人[18]介绍了相关系统在交通、地铁、港口等方面的应用,Huang 等人[16]介绍了相关系统在地铁、大型活动中的应用,有关综述性的文章尚不多.

智能视频监控技术涉及的内容比较多,已有的文献中相当一部分的综述性工作都是对某一类算法进行介绍,如目标检测算法综述[1-2]、目标跟踪算法综述[3-5]等等.也有一些学者从系统角度对多个模块算法进行了介绍,如文献[17,19],这些工作对于智能视频监控技术的发展发挥了重要的作用.然而随着时代的发展,智能视频监控技术也在突飞猛进,近几年来各类优秀算法层出不穷,需要较好的梳理总结;另一方面,之前的综述性文献更多的侧重于从算法原理进行介绍,较少对算法的性能进行比较.本文认为,智能视频监控技术作为计算机视觉和模式识别技术的重要组成部分,是面向安全应用而产生的,性能的优劣是评价算法的重要指标.智能视频监控技术从 2000 年左右发展至今,取得了许多很好的工作进展,也得到了一些有效的应用.作者及其团队在这一领域的研究和应用方面进行了较为长期的工作,本文试图对智能视频分析技术从底层、中层、高层 3 个方面对现有技术进行归纳整理,不仅仅对这一技术在不同层面的典型算法研究现状、存在的瓶颈进行归纳整理,而且对他们的性能评价等试图进行较为全面地总结和深入地探讨.

具体地,本文分别从底层、中层和高层 3 个方面对智能视频监控技术进行概述.算法部分涵盖了目标检测、目标分类、目标跟踪和行为分析 4 个方面,分别介绍了相关研究意义及应用领域、研究现状,并给出了各类典型算法在不同数据集上的性能评测,同时对待解决的问题与难点也进行了讨论;并对物联网时代的智能视频监控技术存在的挑战以及未来发展趋势进行了探讨.本文第 2 节对智能视频监控技术的兴起进行介绍;第 3 节对智能视频监控系统的算法进行介绍;第 4 节对其在大数据环境下的挑战及方向进行讨论.

2 智能视频监控技术的兴起

视频监控是安全防范的重要组成部分,监控的第一要务是用最短时间从被监控的地方获取尽可能多的信息反馈,从信息获取和处理对象的角度而言,早期完全依靠人来获取和处理信息,如我国明朝的

东厂，通过密布的耳目源源不断获得各种信息并形成决策；之后也有利用其他生物本身特有的感知器官来进行监控，如庭院的守门之犬，通过灵敏的听觉与嗅觉及时为主人提供异常信息；之后生物器官也发展成为一些相关的设备，例如，中国历史上最早利用外部设备进行探测的监控系统当属乔家大院的“万人球”，是在清朝末年由水银玻璃制成的镜子，通过它可以看到房间内任何角落的一举一动，且不变形，有如现在的全景摄像机. 直到 19 世纪 70 年代真正发展出的视频监控，开始利用摄像头来获取信息，与这些大部分的信息处理是依靠人来进行处理决策不同的是，智能视频监控开始尝试利用机器智能来辅助人类进行信息处理. 以下将对视频监控技术的发展进行简单介绍.

随着信息技术的进步和市场的需求逐步发展，视频监控技术的发展可以初略的分为 3 个阶段①，如表 1 所示.

表 1　视频监控技术的发展

	产生时间	核心技术	核心设备	特点	不足
第一代（模拟化）	20 世纪 70 年代	光学成像技术和电子技术	摄像头电视墙	技术成熟、价格低廉，安装简单（看得到）	图像质量差、有线传输，难以适应大规模监控
第二代（数字化）	20 世纪 90 年代	数字压缩编码技术和芯片技术	DVR DVS	图像质量好、模块化管理（看得好）	视频数据量大，不易存储和使用
第三代（智能化）	2000 年左右	计算机视觉和模式识别	尚无	智能内容分析（用得好）	分析算法对环境的要求较高

(1) 第一代：模拟视频监控系统

随着光学成像技术和电子技术的发展，监控摄像机的制造和使用成为可能，为了满足利用电子设备代替人或者其他生物进行监控的需求，大约在 20 世纪 70 年代，世界迎来了电子监控系统，这个时期以闭路电视监控系统（Closed Circuit Television，CCTV）为主，也就是第一代模拟视频监控系统[20]. 一般利用同轴电缆传输前端模拟摄像机的视频信号，由模拟监视器进行显示，而存储由磁带录像机完成. 这一代技术价格较为低廉，安装比较简单，适合于小规模的安全防范系统.

(2) 第二代：数字视频监控系统

由于磁带录像机存储容量太小，线缆式传输限制了监控范围等缺点，随着数字编码技术和芯片技术的进步，20 世纪 90 年代中期，数字视频监控系统随之而生. 初期采用模拟摄像机和嵌入式硬盘录像机（Digital Video Recorder，DVR），这个阶段被称为半数字时代，后期发展成为利用网络摄像机和视频服务器（Digital Video Server，DVS），成为真正的全数字化视频监控系统.

DVR 的大量应用使得监控系统可以容纳更多的摄像机，存储更多的视频数据，从而使得摄像机的数量得到海量的提升. 嵌入式和网络通信技术的发达使得图像编码处理单元由后台走向了前端，视频图像在摄像机端编码后经网络传到后台，数字化的视频监控系统应用范围广，扩展性能好，使用和维护简单，适用于超过 100 路、1000 路，甚至城市级规模的安全防范系统，但监控规模扩大的同时带来了对视频内容理解的需求，可以说，数字化技术的发展是智能化技术发展的前提和基础.

(3) 第三代：智能视频监控系统

随着第二代数字视频监控技术的进步，大规模布控成为可能. 同时，随着全球范围安全形势的日益严峻，全世界范围内对视频监控系统的需求空前高涨，各国部署的摄像头越来越密集. 2006 年英国有 450 万个由闭路电视控制的摄像头，每个英国人平均每天会被拍到 300 次②；2008 年美国安装的摄像机已经超过了 2000 万台；2010 年中国超过 1000 万个监控摄像头用于城市监控与报警系统[21].

摄像头的增加带来了大规模防范的可能，即可以获取海量的视频数据用于实时报警和事后查询. 但是对以人为主的使用对象而言，大规模视频数据也带来巨大的挑战. 美国圣地亚国家实验室专门做了一项研究，结果表明，人在盯着视频画面仅仅 22 min 之后，人眼将对视频画面里 95% 以上的活动信息视而不见②.

基于以上需求，智能视频监控系统应运而生，其中最核心的部分是基于计算机视觉的视频内容理解技术，通过对原始视频图像经过背景建模、目标检测与识别、目标跟踪等一系列算法分析，进而分析其中的目标行为以及事件，从而回答人们感兴趣的“是谁、在哪、干什么”的问题[22]，然后按照预先设定的安全规则，及时发出报警信号. 智能视频监控系统有别于传统视频监控系统最大的优势是能自动地全天候进行实时分析报警，彻底改变了以往完全由安保人员对监控画面进行监视和分析的模式；同时，智能

① 对视频监控技术的划分，实际上没有严格的界限和定义，也有“模数混合”、“高清时代”、“IP 监控”等，本文统一称之为数字化视频监控系统.

② CNN. Smart cameras Spot Shady Behavior. http://edition.cnn.com/2007/TECH/science/03/26/fs.behaviorcameras/

技术将一般监控系统的事后分析变成了事中分析和预警，不仅能识别可疑活动，还能在安全威胁发生之前提示安保人员关注相关监控画面并提前做好准备，从而提高反应速度，减轻人的负担，达到用电脑来辅助人脑的目的.

这一技术也得到学界和产业界的认可，美国电气和电子工程师协会(IEEE)在其成立 125 周年大会上，突出展示了 7 项被认为很可能改变世界的技术，其中就包括智能视频监控技术的核心-图像和视频的内容分析技术①. 国际知名视频监控市场网站 IPVM 在 2012 年针对高级会员做了一项投票，选出监控行业未来最具影响的技术，得票最高的便是智能化背景下的视频分析技术，其次是由海量高清监控摄像机带来的大规模视频数据存储②.

接下来本文将对智能视频监控技术的核心算法的发展进行介绍.

3 智能视频监控算法

3.1 智能视频监控算法框架

智能视频监控研究的主要内容就是如何从原始的视频数据中提取出符合人类认知的语义理解，即希望计算机能和人一样自动分析理解视频数据. 比如，判断场景中有哪些感兴趣目标，历史运动轨迹，从事什么行为，以及目标之间的关系等. 一般而言，智能视频监控研究中对视频图像的处理可以分为 3 个层次，如图 1 所示.

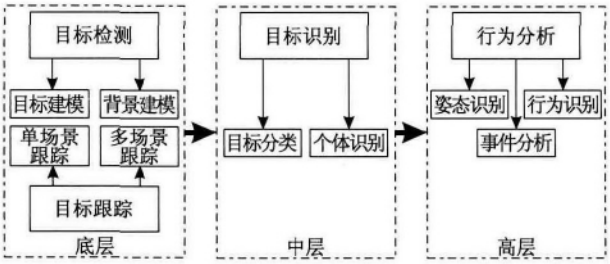


图 1 智能视频监控算法流程

(1) 底层. 主要是从视频图像采集终端获取图像序列，对感兴趣目标进行检测和跟踪，以便对目标进行后续处理分析，主要解决目标在哪里的问题. 本文中目标检测部分可分为目标建模和背景建模. 目标跟踪是为了获得运动目标的活动时间、位置、运动方向、运动速度、大小、表观(颜色、形状、纹理)等，可分为单场景目标跟踪和跨场景目标跟踪.

(2) 中层. 主要是在底层的基础上提取运动目标的各种信息并进行相关判断. 这部分内容包括目标识别. 目标识别是为了对目标进行分类进而识别目标的身份，可分为目标分类和个体识别. 中层的分析为底层处理到高层行为理解搭建了一座桥梁，填补了底层与高层之间的语义间隔. 主要是为了解决目标是什么的问题.

(3) 高层. 高层处理完成对目标的行为进行分析和理解. 高层的语义蕴含着特定的语义场景，往往和具体的应用紧密相关. 行为分析可以分为姿态识别、行为识别和事件分析，主要为了解决目标在干什么的问题.

总而言之，智能视频监控研究的主要目的就是要让计算机回答感兴趣目标在哪里，是什么，在干什么，甚至预测感兴趣目标下一步的行为. 以下本文将分别介绍相关层次.

3.2 目标检测

目标检测是从视频或者图像中提取出运动前景或感兴趣目标，也就是确定当前时刻目标在当前帧的位置，所占大小. 因此目标检测在智能视频监控算法中处于基础地位，目标检测性能的好坏直接影响了后续目标跟踪等算法、目标分类与识别的性能.

根据处理的数据对象的不同，目标检测可以分为基于背景建模的运动目标检测方法和基于目标建模的检测方法. 基于背景建模的方法要求感兴趣目标是保持运动的，并且背景是保持不变的. 当背景发生变化的时候，基于背景建模的方法会将变化背景误检为运动前景，而在运动目标静止一段时间后，也会被归为背景. 因此该方法难以用于背景变化的场景，如手持摄像机或车载摄像机拍摄时. 该方法一般可以达到实时性的要求，因此在采用固定摄像机的应用中广泛使用.

基于目标建模的前景提取方法不受应用场景的限制，不但可以对固定摄像机拍摄的视频进行感兴趣目标的检测，也可以处理单帧静态图像或移动摄像机拍摄的视频. 该方法由于扫描的窗口数目巨多，检测速度较慢，一般很难实时，因此在要求实时性的实际系统中难以应用，两者之间的比较如表 2 所示.

表 2 目标检测方法分类

特点	基于背景建模	基于目标建模
源数据	视频	图像/视频
目标	运动	静止/运动
背景	固定	固定/运动
算法速度	较快	较慢
受遮挡影响	影响较小	影响较大，容易漏检

① http://www.ieee125.org
② http://ipvm.com/report/the_next_big_thing_2012

3.2.1　基于背景建模的目标检测

基于背景建模的方法通过分析视频图像的底层特征，构建背景模型来分割出运动前景，并给出运动前景的位置、大小、形状等信息，并随时间不断更新背景模型.

如何构造鲁棒的背景模型是基于背景建模的运动目标检测算法的关键，目前已有大量的工作来解决这个问题，如帧间差分、均值滤波[23]、中值滤波[24]、最大值最小值滤波[25]、线性滤波[26]、非参数模型[27]、近似中值滤波[28-29]、基于高斯假设的迭代方法[30-31]、基于聚类的方法[32]、基于隐马尔科夫的方法[33]、基于自回归模型的方法[34]、基于在线学习的方法[35-36]以及基于时空背景随机更新的 VIBE 方法[37]. 其中，混合多高斯背景建模方法(Gaussian Mixture Model，GMM)[30]是目前普遍应用的一种前景提取方法. 图 2 显示了在复杂场景中使用 GMM 方法进行前景检测的结果，可以看到检测性能受动态背景、摄像机抖动等因素的干扰比较严重.

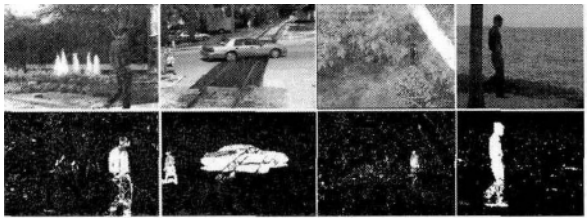


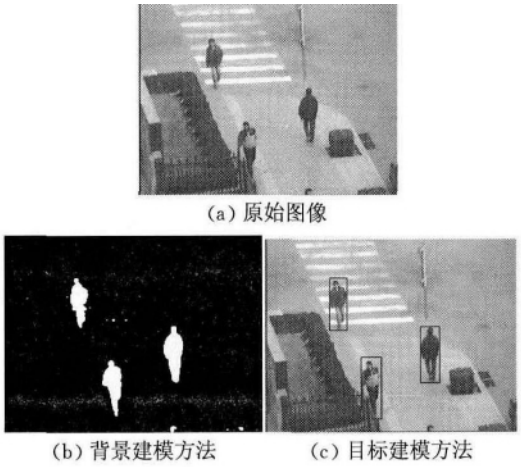
图 2　GMM 方法的检测结果[42]

为了改善前景分割的效果，一些算法[38-40]对前景物体也构造模型. 此外，针对监控场景的特殊性，对运动前景检测的结果有必要进行后处理，如减少缓慢运动或静止的目标突然加速后在原位置留下的“鬼影”[24]以及去除阴影[41]等.

3.2.2　基于目标建模的目标检测

基于目标建模的检测方法对大量训练目标进行学习，训练分类器，在图像多个尺度上做滑动窗口扫描，判定各窗口是目标还是背景，从而得到该图像中所有感兴趣目标的大小和位置. 与基于背景建模的方法不同，通过目标建模方法提取的目标是一个包围框，该方法不能得到目标的轮廓. 如图 3 所示. 基于目标建模的目标检测方法不受场景限制，可以应用于移动摄像头下的目标检测，检测结果不需要再进行个体分割.

基于目标建模的目标检测方法研究的内容有很多，如何从景物的原始灰度图像中提取图像的描绘特征关系到整个系统的可靠性与精度. 因此，如何建立鲁棒高效准确的目标表述模型及分类器是这类方法的核心问题. 基于目标建模的检测方法一般采用滑动窗口的策略. 根据建模的方法不同，基于滑动窗口的目标检测主要分为刚性全局模板检测模型、基于视觉词典的检测模型、基于部件的检测模型和深度学习模型，其他模型中有语法模型以及生物启发特征模型等. 它的一般框架流程图如图 4[43].



(a) 原始图像

(b) 背景建模方法　　(c) 目标建模方法

图 3　背景建模与目标建模的目标检测结果

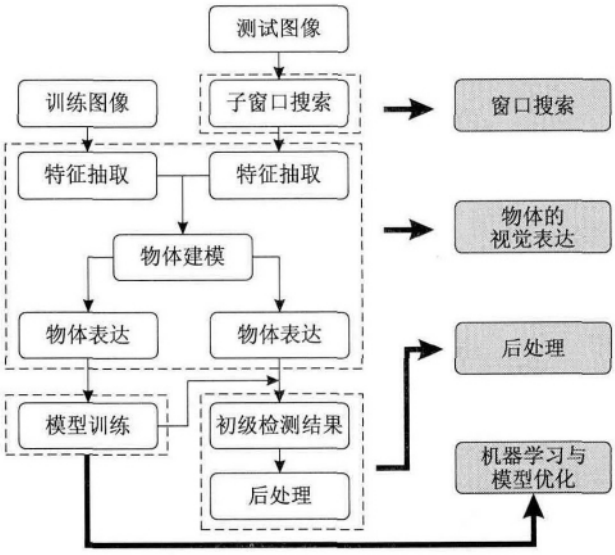


图 4　基于滑动窗口的目标检测系统框架

(1) 刚性全局模板检测模型[44-45]. 假定目标刚性不变，试图通过固定的窗口大小和特征方式完成对目标的表达. 最具代表性的工作是来自法国 INRIA 的 Dalal 等人[44]为解决静态图像的行人检测问题提出的梯度方向直方图(Histogram of Oriented Gradients，HOG)特征，成为近年以来最有影响力和最为成功的特征之一.

(2) 基于视觉词典(Bag-Of-visual Words，BOW)的目标检测模型. 又称为词袋模型，这类方法从训练库中目标抽取局部特征，如 SIFT[46]、SURF[47]

等，学习一个视觉词典（例如，使用 Kmeans 聚类，并把每个聚类中心作为一个视觉单词）；然后给定一幅图片，抽取其局部特征，在学习到的视觉词典上进行投票就得到该图像基于视觉词典的特征表述；最后采用滑窗搜索加 SVM 分类的方法，就可以判定各窗口是否包含目标. 该方法利用局部特征的尺度不变性、仿射不变性以及视角不变性，来解决目标检测中多视角或部分遮挡等比较棘手的问题.

（3）基于部件模型的检测模型. 视觉词典模型丢失了目标的空间布局信息，为弥补这方面的不足，提出了基于部件的目标检测方法. 基于部件模型的目标检测方法将一个目标建模分为整体模型和各个部分模型的综合，有利于解决遮挡目标和多姿态目标等情况[48-50]. 该方法最早可以追溯到 1973 年 Fischler 和 Elschlager[51] 提出的 Pictorial 结构，它认为一个目标是由部分和部分之间的结构组成. 2003 年，Fergus 等人[49] 提出星座模型，该模型既考虑了部件的表观信息，又考虑了每个部件之间的相对位置信息以及部件的尺度信息. 在此基础上，Felzenszwalb 等人[48] 提出可形变部件模型（Deformable Part Based Model，DPBM）. DPBM 包含 3 部分内容：全局模型、部件模型和形变描述模型. 其中，全局模型用来刻画目标的全局结构特性，部件模型用来刻画目标的局部结构，形变描述模型刻画各部件的形变. 形变部件模型奠定了当今物体检测算法研究的基础，也成为后续 PASCAL VOC 竞赛物体检测任务的基础框架.

（4）基于深度学习的目标检测模型. 近几年来，深度学习（deep learning）方法迅速成为研究热点，它主要通过多层神经网络来模仿人大脑的多层抽象机制来实现对数据的抽象表达，将特征学习和分类器整合到一个框架中. 目前深度学习方法在目标检测、分类识别等领域都取得了很好的性能. 一个典型的基于深度学习的目标检测方法包括从输入图像上提取区域块，用卷积神经网络计算每个区域块的特征，最后用线性 SVM 分类器对每个区域块进行分类等步骤[52]. 深度学习模型天然强大的数据表达能力，必然会将目标检测、分类的研究推向新的高度. 当然，目前深度学习模型还存在着解释性差、模型复杂度高、优化困难、计算强度高等诸多问题，这些都需要研究者们进一步的思考.

3.2.3 算法性能评测

本节将分别对背景建模方法和目标建模方法进行评测. 这类算法通常采用召回率和准确度来反映算法的有效性. 其中，召回率代表的是检测结果中正确检测的个数占全部答案包的比重. 准确度指的是检测结果中正确检测的个数占全部检测结果的比例.

（1）背景建模方法评测

为了评测已有的基于背景建模的前景提取方法在不同场景下的性能，Brutzer 等人[53] 人工合成了 SABS（Stuttgart Artificial Background Subtraction）数据集①，模拟多种对背景建模较为挑战的复杂场景，如噪声、动态背景、运动前景与背景表观相似、开关灯等. 并在此数据集上对基于背景建模的目标检测算法的性能进行评测，结果如表 3 所示. 表中性能指标为 F 分数[53-54]，F 分数是准确率与召回率的加权评价指标，F 分数越高反映算法的性能越好. 各种场景下最好的性能用粗体显示.

表 3 在不同复杂场景数据集上的结果[53]

方法	原始	动态背景	独立初始化过程	变暗	开关灯	有噪声的夜晚	前景与背景表观相似	前景与背景表观不相似	压缩编码 H.264
McFarlane[55]	0.614	0.482	0.541	0.496	0.211	0.203	0.738	0.785	0.639
Stauffer[31]	**0.800**	0.704	0.642	0.404	0.217	0.194	0.802	0.826	0.761
Oliver[56]	0.635	0.552	—	0.300	0.198	0.213	0.802	0.824	0.669
McKenna[57]	0.522	0.415	0.301	0.484	0.306	0.098	0.624	0.656	0.492
Cheng[36]	0.766	0.641	0.678	**0.704**	**0.316**	0.047	0.768	0.803	0.773
Kim[58]	0.582	0.341	0.318	0.342	—	—	0.776	0.801	0.551
Zivkovic[59]	0.768	0.704	0.632	0.620	0.300	**0.321**	**0.820**	**0.829**	0.748
Maddalena[60]	0.766	**0.715**	0.495	0.663	0.213	0.263	0.793	0.811	0.772
Barnich[61]	0.761	0.711	**0.685**	0.678	0.268	0.271	0.741	0.799	**0.774**

为了能够描述不同复杂场景下背景建模算法的有效性，我们对表 3 进一步统计，选取了在应用中常见的 6 类场景，对表中 9 种算法的平均性能进行统计，总结了如图 5 所示的规律. 可以看出随着场景复杂度的提升，基于背景建模的算法性能下降显著，其中视频编码等对运动检测的影响不大，但是光线、噪声等对这类算法的影响较大. 图中性能指标为 F 分数[54].

① http://www.vis.uni-stuttgart.de/index.php?id=sabs

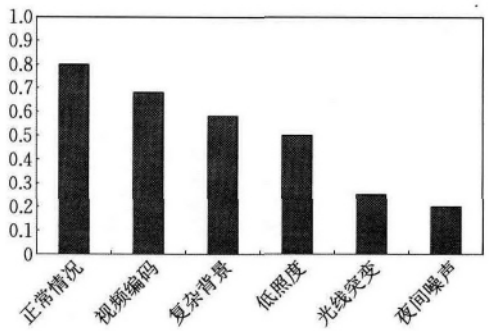


图 5　基于背景建模的前景提取算法随场景变化的性能变化

(2) 目标建模方法评测

PASCAL VOC 数据库是目标检测领域最为公认的测评数据库之一，大部分主流的目标检测方法都会在该数据集上进行测试[6].

将现有的多种方法在 PASCAL VOC 2007 数据库的结果进行对比，如表 4 所示. MKL[62] 是一种刚性全局模板目标检测模型，它采用了 4 种类型的多尺度的特征，再利用多核学习算法对这些特征进行融合. BOW[63] 利用局部特征的尺度不变性、仿射不变性等，来解决目标检测中多视角或部分遮挡问题，是早期主要采用的方法之一. Boosted[64] 方法是基于部件模型的目标检测方法. Boosted 提出了一种局部结构化描述子，并纳入到局部结构化物体建模框架中，取得了 2010 年 VOC 竞赛的冠军. R-CNN[52] 是一种基于深度学习的目标检测方法，相对于其余几种方法，深度学习的方法取得了最好的性能.

表 4　在 PASCAL VOC2007 数据集上的目标检测结果 (单位：%)

类别	MKL[62]	BOW[63]	Boosted[64]	R-CNN[52]
飞机	37.6	15.2	36.7	60.3
自行车	47.8	15.7	59.8	62.5
鸟	15.3	9.8	11.8	41.4
船	15.3	1.6	17.5	37.9
瓶子	21.9	0.1	26.3	29.0
公交车	50.7	18.6	49.8	52.6
小汽车	50.6	12.0	58.2	61.6
猫	30.0	24.0	24.0	56.3
椅子	17.3	0.7	22.9	24.9
牛	33.0	6.1	27.0	52.3
饭桌	22.5	9.8	24.3	41.9
狗	21.5	16.2	15.2	48.1
马	51.2	3.4	58.2	54.3
摩托车	45.5	20.8	49.2	57.0
人	23.3	11.7	44.6	45.0
盆栽	12.4	0.2	13.5	26.9
羊	23.9	4.6	21.4	51.8
沙发	28.5	14.7	34.9	38.1
火车	45.3	11.0	47.5	56.6
显示器	48.5	5.4	42.3	62.2
均值	32.1	10.1	34.3	48.0

基于目标建模的检测方法的性能测评一般采用在信息检索系统中广泛使用的平均准确度(Average Precision，AP)指标作为测评指标. AP 分数是从准确率/召回率曲线计算得到的，因此可以准确的反映出系统的实际检测性能. 具体为，在召回率曲线上进行均匀采样得到相应的准确度，将这些采样得到的准确度的平均值作为 AP 分数. 表 4 中的指标均为 AP.

3.2.4　待解决问题与难点

基于背景建模的检测方法只适用于固定摄像机拍摄的场景，但是在固定场景中干扰因素也很多，如光照环境变化、地板或玻璃反光、阴影、局部动态背景物体(摇晃的树枝、喷泉)等. 这些都是现实场景中常见的问题，极大的影响了算法的性能，给前景提取带来很大挑战.

基于目标建模的检测方法在应用中也有诸多挑战，如巨大的类内差、嘈杂的环境、各式的姿态、严重的遮挡、不同的光线条件、巨大的尺度差异、很小的类间差、严重的形变、低质量图片等等. 而且在应用中，该方法需要事先手工标定大量训练样本，并且在不同的应用场合有可能需要重新标定不同的样本，训练不同的分类器，带来大量的人力开销和费用支出. 另外，由于采用滑动窗口策略，该方法时间消耗比较大，一般难以实时.

上述这些问题依然没有很好的解决. 研究者们依然需要从特征、物体描述和分类器等诸多方面来思考如何提高算法的精度、效率与鲁棒性.

3.3　目标跟踪

目标跟踪用来确定我们感兴趣的目标在视频序列中连续的位置，也就是定位目标“在哪里”. 目标跟踪问题是计算机视觉领域的一个基本问题，是智能视频监控的一个重要环节，具有广泛的应用价值. 目标跟踪可以记录感兴趣目标的历史运动轨迹和运动参数，为更高层的目标的行为分析与理解打下基础[65].

根据应用场景的不同，可以将目标跟踪算法分为单场景目标跟踪和多场景目标跟踪两类. 单场景目标跟踪包括单目标和多目标跟踪，多场景目标跟踪可以分为重叠场景和非重叠场景目标跟踪. 表 5 总结了单场景目标跟踪算法、重叠场景目标跟踪算法，以及非重叠场景目标跟踪算法的各项特性. 例如在单场景中，同一目标在连续两帧的空间位置是很接近的；在重叠场景目标跟踪中，目标经过重叠场景从一个场景进入另一个场景，可以利用连续的空间关系确定进入新场景的目标身份；在非重叠场景目标跟踪中，由于场景之间盲区的存在，不同场景对相同目标的观测在时间空间上都会存在很大差异.

表 5 目标跟踪算法分类及特点

特点	单场景目标跟踪	重叠场景目标跟踪	非重叠场景目标跟踪
时间空间	均连续	空间连续	均有很大间断
时间同步	—	严格要求	要求
摄像机标定	不需要	一般需要	不需要
跟踪范围	较小	较大	很大
受遮挡影响	严重	较小	严重
应用范围	普遍应用	特殊场合	普遍应用

3.3.1 单场景目标跟踪

单场景下的目标跟踪致力于解决指定的单个目标的持续跟踪，也就是在单个摄像机拍摄的视频中只跟踪指定的一个目标. 它与目标检测的关系有两种，一种是在目标检测的基础上，对前景目标进行表观建模，然后按照一定的跟踪策略，找到目标的当前最佳位置（也称之为生成式跟踪）；另一种是目标跟踪与目标检测同时进行，也称为基于检测的跟踪，基本思路是将跟踪问题看作是前景和背景的二分类问题，通过学习分类器，在当前帧搜索得到与背景最具区分度的前景区域（也称之为判别式跟踪）.

对于第 1 种目标跟踪算法，可以是基于特征点[66-67]、基于轮廓[68-71]或基于核[72-73]. 代表性方法如基于光流特征的跟踪算法（Kanade-Lucas-Tomasi，KLT）[72]等. 第 2 种基于检测的目标跟踪算法[74-78]日趋成为目标跟踪算法的主流，代表性方法有 Grabner 等人[78]和 Santner 等人[79]提出的一种基于在线特征提升（Online AdaBoost，OAB）的跟踪方法，Babenko 等人[76]提出的基于多示例学习（Multiple Instance Learning，MIL）的跟踪方法等，Wen 等人[80]利用在线的空时上下文结构信息来辅助跟踪.

上述目标跟踪算法都涉及两个问题：目标表观建模和跟踪策略[42,81]. 目标表观模型是对目标的描述，根据目标的表观数据进行建模，它是跟踪算法的核心模块，表观模型的好坏对跟踪的准确性和鲁棒性起着决定性的影响. 其代表性方法有颜色直方图[71]、梯度方向直方图（Histogram of Oriented Gradients，HOG）[44]、基于核密度估计的表观模型[82-83]、混合高斯模型[84]、基于子空间学习的表观模型[85-86]、基于分块的表观模型[87]、稀疏表达模型[88]等. 跟踪策略用来在当前帧图像中找到最优的目标位置，代表性方法有均值漂移算法[71]、卡尔曼滤波[89]、隐马尔科夫模型[90]和粒子滤波[91]等.

和单目标跟踪不同，多目标跟踪需要处理的问题更多，多目标跟踪存在的挑战包括但不限于：目标的自动初始化，目标间的遮挡推理以及联合状态优化所带来的巨大的计算量等问题[92]. 有些研究将多目标跟踪问题看成一个候选区域和已有目标轨迹之间的数据关联（Data Association）问题，代表性方法如多假设跟踪器（Multiple Hypothesis Tracker，MHT）[93]和联合概率数据关联滤波（Joint Probabilistic Data Association Filter，JPDAF）[94]. 也有一些研究将跟踪问题看成是贝叶斯状态空间推断问题，如 Isard 等人[95]提出一个贝叶斯多目标跟踪器，通过背景建模获得前景区域，然后将多目标似然方程融入粒子滤波的框架，来实现多目标跟踪. 目前的多目标跟踪研究大多没有显示考虑目标的遮挡问题，而且计算效率也不是很高. 如何考虑目标间的交互，并且提高多目标跟踪的效率是一个值得研究的问题.

3.3.2 多场景目标跟踪

多场景目标跟踪是在多摄像机监控网络下为每个运动目标建立唯一的身份标识，从而保证对目标进行全局的持续跟踪.

一个多摄像机跟踪系统由至少 2 路摄像机组成. 每路摄像机都运行着一个单场景跟踪算法，各路单场景跟踪算法既相互独立又相互依赖. 独立性体现在每路摄像机要分别检测、跟踪目标，直至目标离开其视域；依赖性体现在当某路摄像机检测到一个新目标时，要与其他各路摄像机交换信息，以确定该目标的身份，即是新进入系统的，还是已经在其他场景下出现过的. 如果属于前者，则给该目标建立一个新的标号；如果属于后者，则继续沿用其原来的标号. 从而保证了每个目标在该系统中能够被唯一标识.

对于重叠场景的目标跟踪问题，由于采用了多个摄像机从不同视角观测相同区域，这个空间关系为跨场景目标持续跟踪提供了有利条件. 解决此问题主要有两类方法：一种是先确定各摄像机视野之间的视域分界线[96]，即确定两个场景之间重叠区域的位置，出现在视域分界线位置的目标对应于另一个场景下刚进入视野的目标，基于此来确定新进入的目标的身份；另一类方法是基于单应性矩阵来确立不同场景下观测目标的对应关系[97]，利用单应性矩阵，计算目标在对应场景下的位置，从而将不同场景下的目标进行关联.

非重叠场景目标跟踪问题不同于重叠场景的目标跟踪，更不同于传统的单场景目标跟踪. 各场景之

间的监控盲区导致不同摄像机观测到的同一目标的时间以及位置是不连续的，即非重叠场景的目标跟踪中存在严重的时空信息缺失，这给解决非重叠场景的目标跟踪问题增加了难度.

和单场景目标跟踪相比，非重叠场景的目标跟踪具有两个有特色的研究内容，即摄像机网络拓扑估计和跨摄像机目标再识别[98].

(1) 多摄像机网络拓扑估计

拓扑估计是通过学习得到描述多摄像机系统中各摄像机连接关系的拓扑结构. 如图 6 所示，非重叠场景下的多摄像机网络拓扑结构通常包含 3 个要素：第 1 个要素为拓扑结点，它表示各摄像机视野内进出口区域如图 6 中的①～⑧；第 2 个要素为结点之间的连接关系，它表示实际环境中两个进出口区域之间是否存在一条直接连通的物理路径，如图 6 中的实线段表示可见的直接连通的物理路径，虚线段表示不可见的直接连通的物理路径；第 3 个要素为对应于每个连接的转移时间概率分布，表示运动目标在对应的两个进出口区域之间转移所耗时间的概率分布，如图 6 虚线段上的概率分布图.

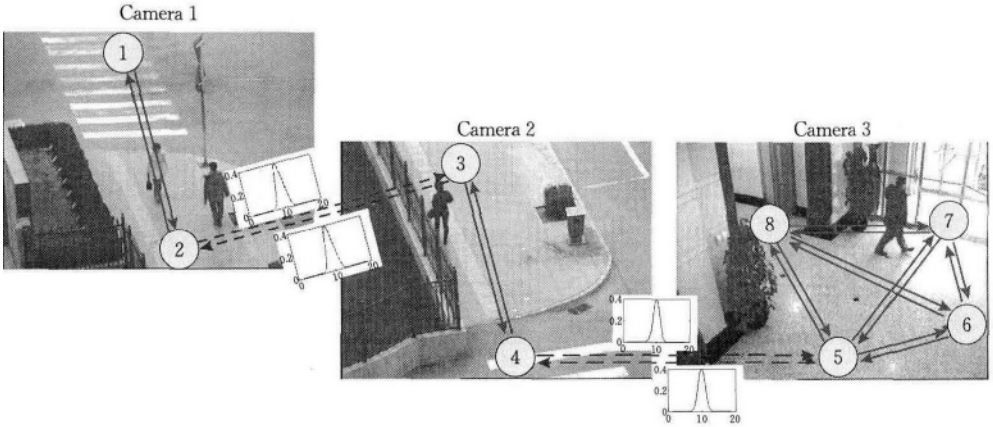


图 6　非重叠场景多摄像机网络的拓扑结构

近年来涌现了许多方法用于估计多摄像机网络拓扑结构，按照研究策略的不同，可以分为基于目标识别与跟踪的方法和基于互相关函数的方法两类. 基于目标识别与跟踪的方法要求跨摄像机间运动目标的对应关系已知(人为标定或者认为跨摄像机目标识别与跟踪已经解决)[99-100]或者要求先采用目标识别或目标跟踪方法学习不同摄像机观测的运动目标之间的对应关系[101]. 基于互相关函数的拓扑估计方法不要求跨摄像机的目标对应关系已知，也不要求学习这个对应关系. Makris 等人[102]通过计算一个摄像机结点观测的到达事件的时间序列和另一个摄像机结点观测的离开事件的时间序列的互相关函数，来估计对应的两个摄像机结点之间的连通性以及转移时间概率分布. 后续有许多工作通过融合目标的表观信息对原始的基于互相关函数的方法进行改进[103-104]，来增强对错误的目标对应引入噪声的鲁棒性.

(2) 跨摄像机目标再识别

为了消除不同场景的不同光照条件对目标表观造成的影响，研究者们提出学习摄像机之间的颜色转移函数[102,105-106]来消除不同光照条件对目标表观颜色的影响.

跨摄像机目标再识别方法按照表观建模方式和匹配策略的不同，可以分为基于鲁棒特征的方法、基于机器学习的方法和基于转移模型的方法 3 类. 基于鲁棒特征的方法[105]就是要对运动目标建立对视角、姿势、光照等变化鲁棒性好的表观模型，然后根据相似度度量方法来匹配不同摄像机下观测的目标表观. 基于机器学习的方法[107-109]在使用一种或多种简单特征对运动目标表观建模之后，利用机器学习的方法来学习两个观测目标的表观模型之间的相似度或距离，再利用其进行目标匹配. 基于转移模型的方法[110]通过建立特征的转移模型来模拟目标表观随摄像机的变化.

3.3.3　算法性能评测

本节将分别对单目标跟踪、多目标跟踪和多场景目标跟踪方法进行评测.

(1) 单目标跟踪性能评测

用于单目标跟踪评测的公开视频序列有很多，如 VIVID[111]和 CAVIAR① 等. 文献[5]将各种数据集完善标定后进行整理汇总，总结了用于单目标跟踪评测的 50 个视频序列.

① CAVIAR: Context Aware Vision using Image-based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

单目标跟踪性能的评测有两种，第1种是根据跟踪中心误差在整个序列上的平均值，这种指标在某些情况下不能公平地全面地反映跟踪算法的性能. 例如，当一种算法在绝大部分视频帧中都能很精确的覆盖跟踪目标，只有极少数几帧跟丢时，有可能导致该算法的评价指标非常低. 第2种评价指标计算算法估计的目标状态满足一定精度范围的视频帧数占整段视频帧数的比例. 通常认为跟踪结果方框和标定方框的重叠面积大于50%(称为重叠指标)或者跟踪结果的中心到标定方框的中心距离小于20个像素(称为精度指标)的视频帧是满足精度范围的[112].

文献[5]中将多种典型的单目标跟踪算法按照上述评价指标进行了性能评测，其中性能最好的几种算法结果如表6所示.

表6 典型算法在50个视频序列上的平均跟踪性能

评测指标	SCM[113]	Struck[114]	TLD[77]	VTD[115]
重叠指标	0.499	0.473	0.437	0.416
精度指标	0.648	0.656	0.608	0.576

(2) 单场景多目标跟踪性能评测

单场景多目标跟踪的公开数据集，如PETS2009 S2L1①、TRECVID08(i-LIDs)②、CAVIAR、ETH③、TUD Stadtmitte④等，视角多样，人群稀疏密度、遮挡程度等差异较大. 如PETS2009 S2L1数据集中人群密集程度相对较为稀疏，而TRECVID08数据集人群较为密集，遮挡非常严重.

单场景多目标跟踪算法的评测指标常用两个指标MOTP(Multiple Object Tracking Precision)和MOTA(Multiple Object Tracking Accuracy)[56]，MOTP和MOTA反应了跟踪算法的准确率，二者都是数值越大性能越好.

此外，南加利福尼亚大学的Li等[116]研究者也定义了一些有意义的评价指标，并且在相关论文中大量采用. 这些指标从另一个方面评价跟踪算法的性能，如一条轨迹中目标标号改变的次数等.

表7显示了几种代表性多目标跟踪算法在公开数据集上的性能. 从表7可以看出比较高的数据集主要集中于PETS2009 S2L1，CEM算法从MOTA和MOTP两个指标上看相比于其他几种方法性能最好. 但是由于PETS2009 S2L1数据集人群较为稀疏，因而MOTA和MOTP分数较高，而像TUD和ETH数据集，由于摄像头视角比较低，遮挡相对更严重，跟踪性能较差.

表7 在公开数据集上的多目标跟踪结果

数据集	方法	MOTA	MOTP
PETS2009 S2L1	K-shortest[117]	0.770	0.630
PETS2009 S2L1	FLP[118]	0.820	0.560
PETS2009 S2L1	CEM[119]	0.959	0.787
TUD	CEM[119]	0.618	0.632
PETS2009 S2L1	PF_TBD[120]	0.750	0.600
ETH	PF_TBD[120]	0.729	0.700

(3) 多场景目标跟踪性能评测

多场景目标跟踪领域常用的公开数据集一般都是重叠场景的，如PETS2009 S2L1、TRECVID08(i-LIDs)等.

非重叠场景的公开数据集非常少，目前公开的数据库有用于测试跨场景之间目标再识别算法性能的i-LIDs和VIPeR⑤. VIPeR提供了1264张图片，包含632个人，每人两张图片，该数据集采集时光照和视角变化都很大，比较挑战，目前现有的跨摄像机目标再识别算法一般都在此数据集上测试. 表8总结了几种代表性目标再识别算法的性能. 其中，评价指标为识别结果正确的目标个数占目标总数的百分比.

表8 目标再识别算法性能 (单位：%)

数据集	测试目标数	PCH[121]	PRDC[112]	LMNN-R[109]	ELF 200[107]
VIPeR	316	≈12	15.66	≈20	≈12
i-LIDS	30	—	44.05	—	—

i-LIDs中有部分场景是非重叠的，但是由于人群密集，遮挡情况非常严重，一般使用该数据集的标定数据来评测目标再识别的性能，而不能直接实现跨场景的目标跟踪. CASIA最近也公开了一个非重叠场景的数据集⑥，人群密度相对i-LIDs较低，可以用于测试非重叠场景的目标跟踪算法. 由于非重叠场景的多摄像机目标跟踪发展还处于起步阶段，算法性能不稳定，受具体场景的影响非常大，目前该领域的跟踪算法多采用非公开的较小规模的数据集来测试效果. 表9对现有的几种非重叠场景的多摄像机目标跟踪算法在不同数据库上的性能进行了总结. 其中跟踪正确率的定义为整个大场景中正确跟踪的目标数与目标总数之比.

① PETS 2009 Benchmark Data. http://www.cvg.rdg.ac.uk/PETS2009/a.html#s2l1
② TRECVid 2008 Evaluation for Event Detection. http://www.itl.nist.gov/iad/mig/tests/trecvid/2008/
③ Robust Multi-Person Tracking from Mobile Platforms. http://www.vision.ee.ethz.ch/~aess/dataset/
④ http://www.d2.mpi-inf.mpg.de/node/428/
⑤ http://vision.soe.ucsc.edu/?q=node/178
⑥ http://www.datatang.com/org/76804

表 9　非重叠场景目标跟踪算法性能统计

	数据集							
	场景					目标		
	数目	环境	视频时长	光照变化	视角变化	数目	转移次数	跟踪正确率
文献[100]	3	室外	15 min	不变	较小	27	45	0.90
文献[105]	2	室外	12 min	较小	较小	—	32	0.94
文献[30]	5	室内	1.5 h	较大	较大	—	50	0.92
文献[122]	5	室内/室外	20 min	较大	较大	14	44	0.93

3.3.4　待解决问题与难点

虽然关于跟踪算法的研究已经持续了很多年，研究者们提出了各种各样的跟踪方法，但是还没有形成一个适用于所有应用场合的统一理论框架或体系，而且目标跟踪在实际应用中遇到的很多难点问题依然没有得到很好的解决，例如光照突变、遮挡、姿态/视角变化、相似物体与杂乱背景干扰等等. 另外，算法的跟踪准确率与运行效率很难同时兼顾. 目前绝大多数跟踪算法，或者实时性好，但准确率低；或者准确率高，但效率低而无法实用. 这使得高层视觉技术的研究和应用受到了极大的约束.

对于多场景的目标跟踪问题，虽然可以利用重叠场景的丰富的空间信息解决单场景下比较棘手的遮挡问题，但是由于经济条件和计算复杂度的限制，一般在实际应用中，更普遍的是非重叠场景的摄像机网络. 因此，非重叠场景下的目标跟踪问题，除了面临单场景目标跟踪要面临的上述问题以外，还引入了新的挑战. 例如，不同摄像机安装的角度不同，所处的光照环境不同，甚至摄像机的参数不同等诸多因素都使得不同摄像机下观测到的同一个运动目标的表观有很大区别，因此跨摄像机目标匹配和识别问题很难解决；不同场景之间的监控盲区导致不同场景下的相同目标的不同观测在时间和空间上都不连续，这种时空信息的缺失给跨摄像机目标关联带来了极大挑战.

综上，速度快、精度高、复杂场景以及大规模多场景的目标跟踪仍然是一个有待解决的研究热点.

3.4　目标分类与识别

目标分类与识别任务要求回答一张图像中是否包含某种物体，也就是判别图像中所包含物体的类别，进而识别出目标的身份①. 作为高层计算机视觉应用的基础，它在很多视觉领域得到了广泛应用，例如行人跟踪以及大规模图像检索等. 因此，目标识别在人类的生活中扮演着越来越重要的角色，也在不断改变着人的生活. 尤其在近五六年中，在不同场景，不同难度和不同规模的数据库的催生下，大量的识别算法被提出，目标识别技术也取得了长足的进步.

在近十年中，有两个最主要方法被广泛的使用，分别是词袋模型(Bag-of-Words)和深度学习模型. 词袋模型从 2005 年开始被广泛认可，并在很多主流数据库上和例年的 PASCAL VOC 目标识别竞赛中都取得了最好的结果[123-124]. 从 2012 年开始，深度学习模型取得了突破性的进展，在大规模数据库 ImageNet-1000 上取得了比词袋模型高出 10%的分类精度，并且迅速成为研究热点引领了近两年的研究潮流.

3.4.1　词袋模型

词袋模型最初产生于自然语言处理领域，通过建模文档中单词出现的频率来对文档进行描述与表达. Csurka 等人[125]于 2004 年首次将词袋的概念引入计算机视觉领域，由此大量的研究工作集中开始于词袋模型的研究，并逐渐形成了由特征提取、特征聚类、特征编码、特征汇聚和分类器分类 4 部分组成的标准目标分类框架[6]. 词袋模型中大量的工作集中在特征编码和特征汇聚方面.

(1) 特征编码. 特征编码密集提取的底层特征中包含了大量的冗余与噪声，为提高特征表达的鲁棒性，需要使用一种特征变换算法对底层特征进行编码，从而获得更具区分性、更加鲁棒的特征表达，这一步对目标识别的性能具有至关重要的作用. 因而，大量的研究工作都集中在寻找更加强大的特征编码方法上，重要的特征编码算法包括向量量化编码[126]、核词典编码[127]、稀疏编码[128]、局部线性约束编码[129]、显著性编码[130]、Fisher 向量编码[131]、超向量编码[132]等.

(2) 特征汇聚. 空间特征汇聚是特征编码后进行的特征集整合操作，通过对编码后的特征，每一维都取其最大值或者平均值，得到一个紧致的特征向量作为图像的特征表达. 这一步得到的图像表达可以获得一定的特征不变性，同时也避免了使用特征集进行图像表达的高额代价. 最著名的如空间金字塔匹配(Spatial Pyramid Matching，SPM)[63]，其也成为当前基于词袋模型的图像分类框架中的标准步骤.

① 个体识别包括人脸、虹膜、步态等多方面，统称为生物特征识别，本文不包括此部分.

3.4.2 深度学习模型

深度学习模型是另一类目标识别算法，在近两年取得了巨大的突破. 其基本思想是通过有监督或者无监督的方式学习层次化的特征表达，来对目标进行从底层到高层的描述. 深度学习中的每一个节点代表一个神经元，这种层次化很好的符合了人脑的神经元处理结构，并通过引入反馈机制模拟人脑的认知过程.

以卷积神经网络为例，卷积神经网络主要包括卷积层和汇聚层. 卷积层通过使用固定大小的滤波器与整个图像进行卷积，来模拟 Hubel 和 Wiesel 提出的简单细胞[133]. 汇聚层则是一种降采样操作，通过取卷积得到的特征图中局部区块的最大值、平均值来达到降采样的目的，并在这个过程中获得一定的不变性. 汇聚层用来模拟 Hubel 和 Wiesel 理论中的复杂细胞. 在每层的响应之后通常还会有几个非线性变换，如 sigmoid、tanh、relu 等，使得整个网络的表达能力得到增强. 在网络的最后通常会增加若干全连通层和一个分类器，如 softmax 分类器、RBF 分类器等. 卷积神经网络中卷积层的滤波器是各个位置共享的，可大大降低参数的规模，对防止模型过拟合是非常有益的；另一方面，卷积操作保持了图像的空间信息，适合于对图像进行表达. 主流的深度学习模型包括自动编码器(Auto-encoder)[134]、受限波尔兹曼机(Restricted Boltzmann Machine，RBM)[135]、深度信念网络(Deep Belief Nets，DBN)[136]、卷积神经网络(Convolutional Neural Netowrks，CNN)[137]、生物启发式模型等[138].

3.4.3 算法性能评测

在目标分类领域，公开数据库根据不同规模，不同复杂度可以分为多个等级. 文献[6]中总结了目标分类领域常用的几种公开数据库，并分析了各数据库的难度. 几种代表性的公开数据库有 Caltech-101[139]、Caltech-256[140]、PASCAL VOC 2007[141]、ImageNet[142]等. 表 10 给出了几种代表性的基于视觉词袋模型的目标分类算法在 PASCAL VOC 2007 数据库上的分类结果，将分类精度作为评测指标. 表 11 对不同数据集的特点进行了分析，并给出了在不同数据库上的词袋模型和深度学习模型的效果，可以看出词袋模型在较小数据集上有一定优势，深度学习模型在较大数据库上取得更好的效果.

表 10 主流目标分类与识别数据库

数据库	图像数目	类别数目	每类图像数目	图像大小/pixel	难度
MNIST	60000	10	6000	28×28	容易
CIFAR-10	60000	10	6000	32×32	中等
MPEG-7	1400	70	20	256×256～650×600	中等
15 Scenes	4485	15	200～400	约 300×250	容易
Caltech-101	9146	101	40～800	约 300×200	中等
Caltech-256	30607	256	80+	约 300×200	较难
PASCAL VOC 2007	9963	20	96～2008	约 470×380	很难
SUN397	108754	397	100+	约 500×300	很难
SUN2012	16873	8	2000	约 500×300	很难
Tiny Images	7900 万	75062	—	32×32	很难
ImageNet-1000	120 万	1000	—	约 500×400	较难
ImageNet	1400 万	10 万	1000	约 500×400	很难

表 11 主流目标识别数据库及最优性能

数据库	光照	遮挡	尺度	视角	形变	正确率
MNIST	小	无	无	无	小	深度学习/79[143]
CIFAR-10	小	无	无	中	小	深度学习/90.68[143]
MPEG-7	无	无	小	无	大	词袋/96.6[144]
15 Scenes	小	小	小	小	小	词袋/98.75[145]
Caltech-101	中	小	小	小	小	深度学习/86.5[146]
Caltech-256	中	中	中	中	中	深度学习/70.6*
PASCAL VOC 2007	大	大	大	大	大	深度学习/79[147]
SUN397	大	大	大	大	大	深度学习/42[148]
SUN2012	大	大	大	大	大	深度学习/——
Tiny Images	大	大	大	大	大	深度学习/——
ImageNet-1000	中	中	大	大	大	深度学习/89[147]
ImageNet	大	大	大	大	大	深度学习/——

注：* PASCAL Workshop on ECCV 2012.

这里我们将最为流行的词袋模型与卷积神经网络模型进行对比，发现两者其实是极为相似的. 在词袋模型中，对底层特征进行特征编码的过程，实际上近似等价于卷积神经网络中的卷积层，而汇聚层所进行的操作也与词袋模型中的汇聚操作一样. 不同之处在于，词袋模型实际上相当于只包含了一个卷积层和一个汇聚层，且模型采用无监督方式进行特征表达学习，而卷积神经网络则包含了更多层的简单、复杂细胞，可以进行更为复杂的特征变换，并且其学习过程有监督过程，滤波器权重可以根据数据与任务不断进行调整，从而学习到更有意义的特征表达. 从这个角度来看，卷积神经网络具有更为强大的特征表达能力，因此它在图像目标识别任务中的出色性能就很容易解释了.

综合表 10、表 11，可以看出，目标分类算法性能受数据库影响很大，在不同规模，不同复杂度的数据库上性能差别较大. 通过对多种目标分类算法在不同数据库上的分类性能进行分析，总结了如图 7 所示的规律. 图中性能指标为在不同规模的数据库上主流算法的分类精度. 可以看出随着数据库分类类别的增多以及图片规模的增大，算法的性能逐渐下降.

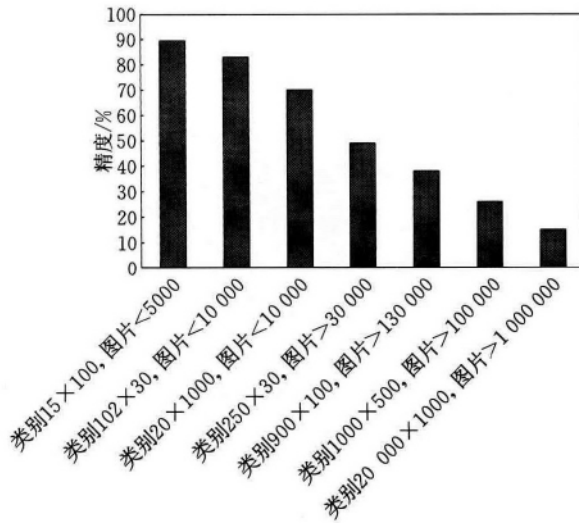


图 7　目标分类算法在不同数据库规模下的分类性能

3.4.4　待解决问题与难点

由于图像采集过程中光照条件、拍摄视角、距离的不同，物体自身的非刚体形变以及其他物体的部分遮挡、使得物体实例的表观特征产生很大的变化，而且目标以外的背景也会千差万别，使得提取局部特征或者中层特征时引入很多噪声和干扰，这些因素在实际应用中普遍存在，给目标分类与识别算法带来了极大的困难.

虽然现代高速计算机的计算能力已达到相当惊人的程度，其目标分类与检测的大多数工作还仅仅局限于一些小规模数据库上的简单识别问题，远远低于人类视觉系统可以轻松识别上万类物体的能力. 速度快、精度高、鲁棒性强的计算机目标分类与识别系统依然是努力的目标.

本文仅介绍了近十年来在目标识别方面两种代表性的方法，其他一些著名方法，比如：PCA、LDA、ICA、AdaBoost、SVM 等就不在此赘述了[7].

3.5　行为分析

行为分析是利用计算机视觉信息（图像或视频）来分析行为主体在干什么. 相对于物体检测和分类来说，人的行为分析是在其基础上实现更高层的目标，涉及到对人类视觉系统的更深层的理解，是计算机视觉领域中要解决的终极问题之一. 除了理论研究价值之外，行为分析也具有非常广泛的应用前景，如人机交互、智能视频监控、智能家居以及视频检索等.

受 Aggarwal 等人[11]工作的启发，本文按照信息提供的复杂程度不同，将行为分析方法分为静态姿态识别方法、运动行为识别方法和复杂事件分析方法，如图 8 所示. 静态姿态识别方法是以静态图像为研究目标，在图像检测和识别方法的基础上对人体的姿态进行分类识别. 后两类行为识别方法的研究目标是基于视频序列的，基于视频的行为识别方法是目前行为识别方法的主要研究方向. 本文将行为分析总结为图 8，并对每部分进行介绍.

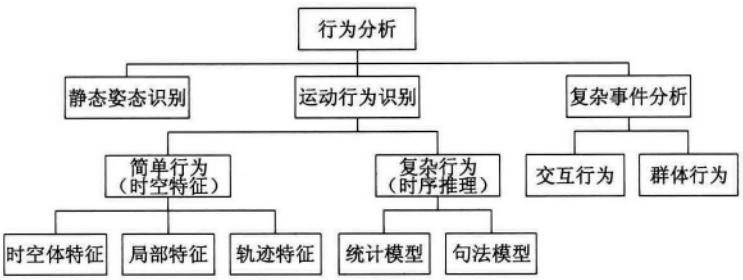


图 8　行为分析组成图

3.5.1 静态姿态识别

静态姿态识别就是通过对静态图像中目标的姿态估计来识别目标的行为.静态姿态识别的发展依赖于两个方面:图像目标分类和姿态模型表达.前者利用静态图像中目标分类的算法直接对静态图片中的人体行为进行分类,这类方法可以看成是对图像分类中物体表达的一种延伸[149],典型如 PASCAL VOC 行为分类竞赛[148,150].另一类更关注在人体姿态模型的表达.人体姿态模型的研究与计算机视觉的发展密切相关.

在 20 世纪 70 年代,由于计算能力和数据可视化手段的限制,对人体姿态估计的研究主要停留在人体模型描述及关节运动分析阶段.80 年代和 90 年代的研究提出基于(形状)模型的方法[151-152].2000 年以后的相关领域研究热点主要集中在通过各种设备捕捉装置获取更加接近真实的姿态估计上[153-154],2010 年微软推出 Kinect,利用 RGBD(RGB&Depth)检测跟踪人体骨架(随机森林算法)[155]来估计人的姿态,并成功运用到体感交互产品中,Yang 等人[156]采用部件组合的方式来对姿态进行估计也获得了较好的效果.

基于静态图像的人的行为识别方法虽然具有简单的优势,但由于缺少行为中的动态时序信息,对于一些描述动作转移的行为,如"坐下"和"站起",一张图像是无法达到分类和识别的目的的.因此,静态姿态识别目前通常是作为简单行为识别的一种方法,或者为运动行为提供基础信息,而以人体的动态时序信息为基础的基于视频的行为识别方法是行为识别领域的主流方法.

3.5.2 运动行为识别

人体运动是一个非常复杂的系统,具有很大的自由度和高度非线性的特点.利用运动信息来判断人的行为是行为分析的重要研究方向.早在 1973 年,Johansson[157]就通过在黑色背景中的人的关节点处贴亮点来获取关节点的运动轨迹数据.这些在单个图像中看似毫无意义的亮点在动态图像序列中通过相互运动能够明显的表达人的各种动作.按照对时序信息的使用程度其可以划分为时空特征方法和时序推理方法两类.

(1)时空特征方法

时空特征方法主要是在特征层面考虑运动的信息,具体而言是将一个包含行为的视频序列看作在时空维度上的三维立方体,然后从这个三维立方体中提取有效的行为特征.这类方法相对而言,主要面向简单行为.经典方法包括基于时空体模型[158-159]、局部特征[160-165]和时空轨迹[166-170]这 3 类.

① 时空体模型特征.基于时空体模型的方法是对整个三维立方体进行建模.如 Bobick 和 Davis[158]利用人体在三维立方体中沿时间进行投影,构造了运动能量图和运动历史图,然后利用模板匹配的方法对行为进行分类.运动历史图可以看作是人体在三维立方体中沿时间轴的加权投影,该投影不仅能反映出运动物体的姿态,还包含了不同姿势的时序性信息.

为了能在更复杂的场景下对人的行为进行识别,Ke 等人[159]利用层级的均值漂移算法对时空立方体进行分割并自动找到人的行为对应的时空区域,然后对该部分时空区域进行建模.此类将行为作为一个整体进行建模和分类的方法比较直观,对于识别一些简单场景的行为比较有效,但对于复杂场景的行为,由于光照、视角以及动态背景等因素的影响,此类方法的有效性将大大降低.Wang 等人[171]通过估计人的姿态的关节点,并采用数据挖掘的方式来获取行为时空结构的最优表达,在多个数据库上取得了很好的效果.

② 局部特征.因局部特征在图像识别领域的巨大成功,很多方法试图从时空立方体的局部出发,获取更多的时空局部特征.局部特征可以通过构建三维时空滤波器的方式快速的提取时空立方体中的兴趣点,这些底层的时空兴趣点具有旋转和尺度不变性,可以很好的提高行为识别方法的鲁棒性.

基于局部特征的行为识别方法首先通常是构建兴趣点检测子如 Harris3D 检测子[160]、SIFT 检测子[161]和 Hessian 检测子[162].然后构建局部特征描述子,利用在检测子检测到的兴趣点周围提取表观和运动信息形成局部特征向量.常用的有 Cuboid 描述子[162]、HOG3D 描述子[164],近年来也有学者从深度传感器出发构建描述子[163]和行为模型,取得了不错的效果[172].此类方法可以直接与词袋模型(Bag of words)结合得到局部特征视觉单词的直方图特征,将该直方图特征作为最终的行为特征送入分类器.除了简单的对不同的局部特征进行统计外,也有很多方法[165-166,173]利用局部特征在时空中的空间位置关系学习行为的中层表达.基于局部特征的方法对于处理真实场景中的行为识别问题具有很好的鲁棒性,因此,此类方法被用于互联网的视频分类中.然而,局部特征缺少行为的全局动态信息,这个缺陷决定基于局部特征的行为识别方法在识别效果

上是有上限的.

③ 时空轨迹特征. 时空轨迹由于能够在更大的时间范围内对行为的动态信息进行描述，因此可以有效的提高行为的表达能力.

时空轨迹是将人体在运动中的运动点沿时间轴连接在一起形成的轨迹曲线. 早期工作对轨迹的描述比较简单，Campbell 和 Bobick[166] 通过将一个行为的轨迹映射为一个相空间中的一条线，通过对相空间中曲线的划分来进行行为的识别. Lv 等人[167] 以及 Huang 等人[168] 提出了基于有限点如关节点轨迹的行为识别方法. 为了能够将轨迹方法的优势应用到复杂场景的行为识别中，Messing 等人[169] 以及 Wang 等人[170] 结合局部特征检测方法提出了基于局部兴趣点轨迹的行为识别方法. 时空的方法非常适合处理手势和单人行为，局部特征能够很好的提高行为识别方法的鲁棒性，基于轨迹的方法也能够通过获取更多的时序信息来极大的提高行为识别方法在真实场景行为数据库中的性能[174]. 这类方法的不足是缺少对行为的高层表达. 因此，这类方法不适合解决复杂的行为识别问题.

(2) 时序推理方法

基于时空的方法主要是从提取时序特征出发对运动行为进行识别，并没有充分考虑到时序之间的关联性，从而难以识别更复杂的行为. 复杂行为识别在简单行为识别结果基础上考虑到简单行为之间的时序关联，因此，复杂行为识别方法常出现层级现象. 经典的复杂行为识别方法可以分为统计模型方法[175-178] 和句法模型方法[179-181].

① 统计模型. 统计模型使用基于状态的统计模型来识别行为，子行为被看作概率状态，行为被看作沿这些子行为沿时间序列转移的一条路径. 底层的一些子行为可以通过上面提到的时序方法进行识别，这些子行为进一步的构成了一个高层行为序列. 在高层的模型中，每一个子行为在这个序列中作为一个观测值.

Nguyen 等人[175] 以及 Shi 等人[176] 利用隐马尔科夫模型(HMM)对子行为序列进行建模来进行复杂行为识别，Damen 等人[177] 则利用子行为构建动态贝叶斯网络(DBN)来实现复杂行为的识别问题. 利用 HMM 和 DBN 模型可以很好的对子行为序列进行建模，但对于描述一些具有空间关系的子行为，即子行为之间存在着时间的重叠，直接利用这两种模型则无法对复杂行为进行描述. 为了能够更好的描述复杂行为中子行为之间的相互关系，Tran 等人[178] 利用一定的先验知识构建了马尔科夫逻辑网络 MLNs(Markov Logic Networks)来对子行为之间的时空关系进行描述.

② 句法模型. 句法模型把子行为看作一系列离散的符号，行为被看作这些符号组成的符号串. 子行为可以通过上面提到的时空或时序方法进行识别，而复杂行为可以用一组生成这些子行为符号串的生成规则来表示，自然语言处理领域的语法分析技术可以被用来对这种生成规则进行建模，从而实现对复杂行为的识别. 这一类基于语法分析技术构建的模型被称为句法模型，常用的有上下文无关语法模型(Context-Free Grammars)和上下文无关的随机语法模型(Stochastic Context-Free Grammars)[179-180].

一般的句法模型也只能识别子行为序列构成的复杂行为，对于处理同时发生的子行为则无能为力. 为了克服这个局限，Zhang 等人[181] 在 CFG 的基础上加入了描述子行为之间复杂时空关系的逻辑连接，即 and、or 和 not，使得构建的句法模型可以解决子行为共同发生的问题.

3.5.3　事件分析

事件是指在特定条件或外界刺激下引发的行为，是更为复杂的行为分析，包括对目标、场景及行为前后关联的分析. 事件分析是行为分析的高级阶段，能够通过对目标较长时间的分析给出语义描述. 之前的行为识别可以是事件分析的基础，但事件分析也具有其特殊性，仅仅依赖于前述的行为识别并不能较好地解决事件分析. 按照事件的复杂程度，事件分析可以分为多人交互行为[150,182-184] 以及群体行为[185-187].

(1) 交互行为. 为了识别人与物的交互行为，首先要做的是识别物体和分析人的运动信息，然后联合这两种信息进行交互行为的识别. 早期的交互行为识别方法[182] 忽略物体识别和运动估计的相互影响，即先利用物体分类方法来识别物体，然后再识别这些物体参与的运动行为①. 没有利用物体识别和运动分析两者的相互关系，运动估计是严格依赖于物体检测的. 为了利用物体与动作之间的相互关系来提高物体检测和行为识别的性能，Moore 等人[183] 利用简单行为识别的结果来提升物体分类的性能. 一般情况下，行为识别还是依赖于物体分类

① 9 Ongoing Trends for Surveillance Analytics. http://www.securitymagazine.com/articles/83984-ongoing-trends-for-surveillance-analytics

的,但当物体分类出现错误时,行为信息通过构建的贝叶斯网络对物体分类进行补偿.更进一步的,Gupta 和 Davis[184]提出了一种概率模型来整合物体表观、人体对物体的动作以及动作对物体的反作用.这些信息通过贝叶斯网络被整合在一起来对物体和行为进行分类和识别.

(2)群体行为.群体行为是有一个或多个人群组成的行为.其研究对象是多人形成的群体.群体行为分析根据所要获取知识的不同,可以分为两类.第一类是每个个体在整个群体行为中发挥不同的作用[185-186].例如分析一个"做报告"的行为,需要分析报告者的行为和听众的行为.此类群体行为可以很自然的通过由多个个体的子行为构建的多层模型对群体行为进行表达.另外一类群行为是将个体的运动信息作为一个整体来进行群体行为分析.如"军队行军"和"游行"等都属于这类群体行为.在此类群体行为方法中[187],每个个体经常被当作一个点,然后利用这些点的轨迹对整体行为进行分析.

3.5.4 算法性能评测

本节对人的行为分析方法进行性能评测.从不同的分类可以看出,行为识别所要解决的问题是多样的,不同的行为对应着不同的解决方法.虽然现有的行为分析方法可以很准确的识别一些相对简单的行为,但对于处理复杂行为以及事件级行为来说,仍是一个非常具有挑战的研究方向.

为了能够对行为识别的研究现状有更为直观的了解,图 9 给出了当今流行算法在一些常用的行为数据库所能达到的最好结果.由于事件分析的数据库的评测标准不够统一,本文只列出了一般行为数据库(Weizmann[188]、KTH[189]、YouTube[190]、Hollywood2[191]、UCF Sports[192]、IXMAS[193]、UIUC[194]、Olympic Sports[173]、UCF50[195]、UCF101[196]、HMDB51[197]、MSR Action3D[198]、MSR Gesture3D[199]和 MSR Daily Activity3D[200])的最好结果.其中很多方法已经在 Weizmann 中达到 100%的识别率,Wang 等人[170]提出的基于轨迹的方法在 KTH、YouTube、UCF sports 和 IXMAS 中得到了最好的结果,文献[172]在 Hollywood2、HMDB51、Olympic Sports 和 UCF50 中得到了最好的结果.文献[174]在 UIUC 中取得了最好的结果.文献[173]在 MSR Action3D 中取得了最佳识别结果,而文献[171]中的方法在 MSR Gesture3D 和 MSR Daily Activity3D 中得到最佳结果.

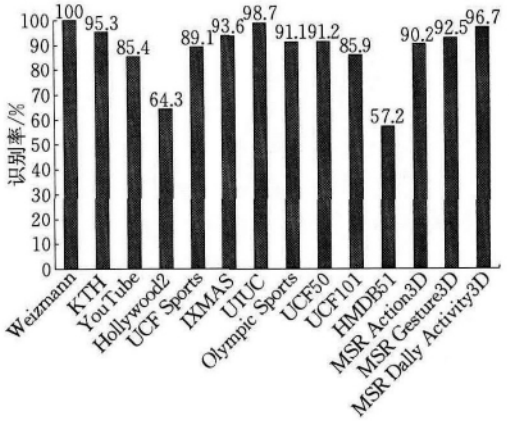


图 9 在常用数据库上现有较好行为识别算法的结果[170]

4 物联时代的智能视频监控挑战及方向

根据公开数据以及调研报告分析,2007 年全国摄像机市场规模大概 500 万台,之后每年大约 20%~30%的增长.被称为"高清元年"的 2010 年,高清监控摄像机开始规模化应用,并且保持高增长速度.根据调查,2012 年全国高清摄像机(包括 IPC、HD-SDI、720p、1080p 等全部在内)市场规模接近 400 亿人民币,总出货量已突破 360 万台,约占到摄像机市场总量的 16%,也就是说 2012 年全国摄像机实际架设数量接近 2000 万台.据不完全统计,目前全国架设摄像机数量接近 1.5 亿台,其中高清摄像机接近 1500 万台[21].

如果把摄像机看作人的眼睛,而智能视频系统或设备则可以看作人的大脑,视频监控就是物联网的感知环节少不了的"眼睛".海量的摄像头构成了视联网,产生了洪水般的视频数据:按照 3 Mbps 的平均码流,仅仅 1500 万高清摄像机部分的存储一小时就需要 20 EB,一个月需要 15 ZB,按照每块硬盘 2T 的容量,那么存储这些需要 750 亿块硬盘,耗资超过 100 万亿人民币,相当于接近 2012 年全国 GDP(519 322 亿元)的 2 倍.

海量的监控数据,标志着视频监控进入了大数据时代,可以用 4 个"V"概括,即 Volume、Variety、Value、Velocity[195].Volume,数据体量巨大,从 TB 级别,跃升到 ZB 级别;Variety,数据类型繁多,物理数据到网络数据(视频、图片、地理位置信息等等);Value,价值密度低,连续不间断监控过程中,可能有用的数据仅仅有一两秒;Velocity,增加速度快,1 秒定律,最后这一点也是和传统的数据挖掘技术有着本质的不同.视频数据也被认为是大数据中的最主

要组成[201].

4.1　物联时代面临的四大挑战

海量的视频监控数据给智能视频监控技术及系统应用带来了巨大的挑战，这些挑战可以归纳为跨场景、跨媒体、跨空间的几大挑战[202]：

(1) 监控节点泛在分布

目前全球的摄像头急剧增加，与监控相关摄像头的种类已经不仅仅限于常规固定摄像头，还包括移动摄像头，如移动电脑、手机等等，手机等手持设备数量已达数十亿级，而互联网主机数量也突破了十亿级. 这些监控节点包括监控摄像机在全球范围内几乎可以任意分布，快速组网，给管理与分析带来了极大挑战，如何使这些设备能够自组织成一个完整的体系是将来需要解决的一个难题[203].

(2) 监控数据海量混杂

如今，海量监控节点不断增加，监控数据的类型也从单一性扩展到多样性. 监控数据的类型已经不仅仅局限在监控摄像头，图像、语言、文本等信息都夹杂其中，这些数据作为监控的载体，都起到重要的作用. 如何在各种载体中获取有用信息，将大数据变成小数据是急需克服的问题[204].

(3) 监控对象种类繁多

监控范围的扩大使得监控对象的种类不断扩大，监控的对象从传统的像素级到目标级直至发展到事件级，从不同监控场景下分析出各种目标的行为、目标之间内在联系以及群体目标之间的事件级演变是面临的一大难题[205].

(4) 监控态势动态演变

作为指挥部门来说，发生突发事件之后，如何组织力量快速响应最重要. 这就需要一套完备的对于态势预判的技术，能够自动预测监控目标即将选择的行进路径，从而能够做出动态响应，目前该技术还不成熟，需要目标识别与跟踪等多项技术的突破来实现(见第 1107 页脚注①).

4.2　物联时代的智能视频监控发展方向

以物联网为基础的海量大数据对智能视频监控提出了巨大的挑战，但与此同时也能有效推动智能监控系统的发展. 如平安城市的联网监控，是将原本各个区域和楼宇分散的监控资源，如安防、交通、零售、家居中的监控资源，整合成一个地域范围的一体化监控资源，然后再将城域的资源整合成一个广域的资源. 在物联网的前提下，信息可以传送和集中的价值并不仅仅是跨越空间距离获取信息，更关键的是在这个基础上可以扩展各种各样的智能分析应用. 在物联时代，智能视频监控将发展成为新的视觉物联，视联网将具有如下特性，其也将是智能视频监控系统发展的新方向.

(1) 高效视觉网

监控网络随时组合、自动调整，自组织视联网使动态的监控任务和泛在的摄像头资源之间形成有序对接，提高分布式监控终端的利用效率和监控效果，其中分布式监控系统起到了重要作用. 分布式对象技术是伴随网络而发展起来的一种面向对象技术，它采用面向对象的多层客户端/服务器计算模型，将分布在网络上的全部资源按照对象的概念来组合，使得面向对象技术能够在异构的网络环境中得以全面和方便的实施，有效地控制系统的开发、管理和维护的复杂性. 分布式监控系统[206]基于分布式对象技术建立，一方面可以增强监控网络的可扩展性，可以随意增加不同功能的监控终端，另一方面可以提高系统的工作效率以及响应速度.

(2) 协同视觉网

实现物理监控区、监控单元与目标以及与网络空间感知数据的交流和互动，实现全方位的(更深入、更全面、更准确、更可靠)视频监控. 一方面在物理空间中，视频监控终端、移动监控终端等的泛在分布，构成一个庞大的视频监控网络，使感知数据海量混杂；另一方面在网络空间中，网址、论坛、博客、微博、话题倾向与传播情况、社会关系等构成了网络空间海量的感知数据. 单一的感知数据不能准确的把握公共安全事情的来龙去脉，只有对跨时间、跨地域、跨物理空间与网络空间的感知信息融合才能完成. 通过物理与网络空间信息的交互、协同与融合，建立二元空间协同感知、主动调制的立体化社会感知与预警模式，由二元空间的复杂关联特征与协同互动感知机制，实现二元空间协同的实体和社会群体发掘与关联分析、公共安全事件跨时空全局态势分析与预测. 建立二元空间需要融合多媒体数据[207-208]，发掘跨媒体数据的语义关联，有助于对异常态势的多维度预警.

(3) 主动视觉网

如果说高效视觉网获得更加海量的监控数据，协调视觉网获得多种类型的监控数据，两者还是在被动感知，主动视觉网则是在两者基础上实现真正的主动分析. 高效主动视觉网，充分利用“海量”与“不同媒体”数据之间的“结构特征”和“时序动态信息”进行统一表达和计算，显著提高视频监控内容语义理解水平. 从不同媒体数据(包括文本、图像、视频

等)中提取能表达公共安全事件的多方面特征,包括时间、地点、人物、时序、关系等,从而提高语义表达的完整性,对事件检测更加可靠性.跨媒体数据[209-212],提供了不同媒体数据之间的关系信息,可以通过鲁棒的结构学习海量跨媒体数据,得到关联信息可以表达同一事件的多方面特征,从而提高监控网络预警的可靠性.

智能视频监控技术作为最早应用于物联网的重要技术之一,其发展必将受到物联网的巨大影响.智能视觉监控技术涉及图像处理、图像分析、机器视觉、模式识别、人工智能等众多研究领域,是一个跨学科的综合问题,也是一个极具挑战性的前沿课题.

5 结 论

智能视频监控技术作为下一代视频监控发展的趋势,具有广阔的发展空间.但大规模应用中的核心关键技术尚处在积累阶段,需要深厚的技术功底以及不断地应用验证.物联网大数据时代给这一技术的发展带来了巨大的挑战,同时也迎来了更大的机会.美国政府在公布的"大数据研发计划"(Big Data Research and Development Initiative)中包含一个旨在为机器建立视觉智能的 Mind's Eye 项目.这一项目有别于传统的利用广泛选取对象来描述单一场景的机器视觉研究,旨在通过知觉认知获取行为和知识推理来增加场景的理解,这一计划可以被认为能够建立一个更完整的视觉智能效果.我们有理由相信,在不久的将来,依靠视频大数据的智能视频监控技术一定会具有人类一样的大智慧.

参 考 文 献

[1] Bouwmans T, El Baf F, Vachon B. Background modeling using mixture of Gaussians for foreground detection: A survey. Recent Patents on Computer Science, 2008, 1(3): 219-237

[2] Wojek C, Dollar P, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. IEEE Pattern Analysis and Machine Intelligence, 2012, 34(4): 743-761

[3] Yilmaz A, Javed O, Shah M. Object tracking: A survey. ACM Computing Surveys (CSUR), 2006, 38(4): 1-29

[4] Wang X. Intelligent multi-camera video surveillance: A review. Pattern Recognition Letters, 2012, 34(1): 3-19

[5] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2411-2418

[6] Huang Kai-Qi, Ren Wei-Qiang, Tan Tie-Niu. A review on image object classification and detection. Chinese Journal of Computers, 2014, 37(6): 1225-1240(in Chinese)
(黄凯奇,任伟强,谭铁牛.图像物体分类与检测算法综述.计算机学报, 2014, 37(6): 1225-1240)

[7] Andreopoulos A, Tsotsos J K. 50 years of object recognition: Directions forward. Computer Vision and Image Understanding, 2013, 117(8): 827-891

[8] Zhang X, Yang Y H, Han Z, et al. Object class detection: A survey. Association for Computing Machinery Computing Surveys (CSUR), 2013, 46(1):1311-1325

[9] Hu Qiong, Qing Lei, Huang Qing-Ming. A survey on visual human action recognition. Chinese Journal of Computers, 2013, 36(12): 2512-2524(in Chinese)
(胡琼,秦磊,黄庆明.基于视觉的人体动作识别综述.计算机学报, 2013, 36(12): 2512-2524)

[10] Morris B T, Trivedi M M. A survey of vision-based trajectory learning and analysis for surveillance. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(8): 1114-1127

[11] Aggarwal J K, Ryoo M S. Human activity analysis: A review. ACM Computing Surveys, 2011, 43(3): 16

[12] Collins R T, Lipton A J, Kanade T, et al. A system for video surveillance and monitoring: VSAM final report. Robotics Institute, Carnegie Mellon University: Technical Report CMU-RI-TR-00-12, 2000

[13] Siebel N T, Maybank S. The advisor visual surveillance system //Proceedings of the European Conference on Computer Vision Workshop on Applications of Computer Vision. 2004: 103-111

[14] Shu C F, Hampapur A, Lu M, Brown L, et al. IBM smart surveillance system (S3): An open and extensible framework for event based surveillance//Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance. Como, Italy, 2005: 318-323

[15] Shah M, Javed O, Shafique K. Automated visual surveillance in realistic scenarios. IEEE Transactions on Multimedia, 2007, 14(1): 30-39

[16] Huang Kaiqi, Tan Tieniu. Vs-star: A visual interpretation system for visual surveillance. Pattern Recognition Letters, 2010, 31(14): 2265-2285

[17] Hu W, Tan Tieniu, Wang Liang, et al. A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2004, 34(3): 334-352

[18] Valera M, Velastin S A. Intelligent distributed surveillance systems: A review. IEE Proceedings -Vision, Image and Signal Processing, 2005, 152(2): 192-204

[19] Wang Su-Yu, Shen Lan-Sun. Intelligent visual surveillance technology: A survey. Journal of Image and Graphics, 2007, 12(9): 1505-1514(in Chinese)
(王素玉,沈兰荪.智能视觉监控技术研究进展.中国图象图形学报, 2007, 12(9): 1505-1514)

[20] Lyon D. Surveillance Studies: An Overview. Cambridge: Polity Press, 2007

[21] 2013－2018 Chinese intelligent video surveillance market forecast report of supply and demand situation and development prospect(in Chinese)
(2013－2018 年中国智能视频监控市场供需现状及发展前景预测报告)

[22] Haritaoglu I, Harwood D, Davis L. W^4S: A real time system for detecting and tracking people in $2\frac{1}{2}D$ //Proceedings of the 3rd Internet Conference on Automatic Face and Gesture Recognition. Nara, Japan, 1998: 877-892

[23] David C, Gui V. Automatic background subtraction in a sparse representation framework//Proceedings of the Systems, Signals and Image Processing(IWSSIP). Bucharest, Romania, 2013: 63-66

[24] Cucchiara R, Piccardi M, Prati A. Detecting moving objects, ghosts, and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(10): 1337-1342

[25] Liu Z, Huang K, Tan T. Foreground object detection using top-down information based on EM framework. IEEE Transactions on Image Processing (TIP), 2012, 21(9): 4204-4217

[26] Li Dawei, Xu Lihong, Goodman E D. Illumination-robust foreground detection in a video surveillance system. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 23(10): 1637-1650

[27] Bayestehtashk A, Shafran I. Parsimonious multivariate copula model for density estimation//Proceedings of the Acoustics, Speech and Signal Processing. Vancouver, BC, 2013: 5750-5754

[28] Li Fuxin, Kim T, Humayun A, et al. Video segmentation by tracking many figure-ground segments//Proceedings of the IEEE International Conference on Computer Vision. Sydney, NSW, 2013: 2192-2199

[29] Prioletti A, Mogelmose A, et al. Part-based pedestrian detection and feature-based tracking for driver assistance: real-time, robust algorithms, and evaluation. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(3): 1346-1359

[30] Zhu Qingsong, Song Zhan, Xie Yaoqin. An efficient r-KDE model for the segmentation of dynamic scenes//Proceedings of the International Conference on Pattern Recognition. Tsukuba, Japan, 2012: 198-201

[31] Stauffer C, Grimson W. Learning pattern of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 747-757

[32] Butler D, Sridharan S, Bove V M Jr. Real-time adaptive background segmentation//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 2003, 3: 49-52

[33] Liu Z, Huang K, Tan T. Cast shadow removal using MRF based on hierarchical information. IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT), 2012, 22(1): 56-66

[34] Monnet A, Mittal A, Paragios N, Ramesh V. Background modeling and subtraction of dynamic scenes//Proceedings of the IEEE International Conference on Computer Vision. Nice, France, 2003: 1305-1312

[35] Grabner H, Bischof H. Online boosting and vision//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 260-267

[36] Cheng Li, Gong Minglun, Schuurmans D, Caelli T. Real-time discriminative background subtraction. IEEE Transactions on Image Processing, 2011, 20(5): 1401-1414

[37] Barnich O, Van Droogenbroeck M. ViBe: A universal background subtraction algorithm for video sequences. IEEE Transactions on Image Processing, 2011, 20(6): 1709-1724

[38] Wang Y, Loe K F, Wu J K. A dynamic conditional random field model for foreground and shadow segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(2): 279-289

[39] Sun J, Zhang W, Tang X, Shum H Y. Background cut//Proceedings of the European Conference on Computer Vision. Graz, Austria, 2006: 628-641

[40] Sheikh Y, Shah M. Bayesian modeling of dynamic scenes for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(11): 1778-1792

[41] Liu Zhou, Huang Kaiqi, Tan Tieniu. Cast shadow removal combining local and global features//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hyatt Regency Minneapolis, USA, 2007: 1-8

[42] Ding Jian-Wei. Object Detection and Tracking in Complex Scenes [Ph. D. dissertation]. University of Chinese Academy of Sciences, Beijing, 2012(in Chinese)
(丁建伟. 复杂场景中的目标检测与跟踪研究[博士学位论文]. 中国科学院大学, 北京, 2012)

[43] Zhang Jun-Ge. Object Detection Based on Visual Structure Representation and Modeling [Ph. D. dissertation]. University of Chinese Academy of Sciences, Beijing, 2013(in Chinese)
(张俊格. 基于视觉结构表达与建模的物体检测研究[博士学位论文]. 中国科学院大学, 北京, 2013)

[44] Dalal N, Triggs B. Histograms of oriented gradients for human detection//Proceedings of the IEEE Computer Vision and Pattern Recognition, San Diego, USA, 2005: 886-893

[45] Viola P, Jones M J. Robust real-time face detection. International Journal of Computer Vision, 2004, 57(2): 137-154

[46] Lowe D G. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91-110

[47] Bay H, Ess A, Tuytelaars T, et al. SURF: Speeded up robust features. Computer Vision and Image Understanding, 2008, 110(3): 346-359

[48] Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645

[49] Fergus R, Perona P, Zisserman A. Object class recognition by unsupervised scale-invariant learning//Proceedings of the IEEE Computer Vision and Pattern Recognition. Madison, USA, 2003: 264-271

[50] Girshick R B, Felzenszwalb P F, McAllester D. Object detection with grammar models//Proceedings of the Conference on Neural Information Processing Systems (NIPS). Granada, Spain, 2011: 442-450

[51] Fischler M, Elschlager R. The representation and matching of pictorial structures. IEEE Transactions on Computers, 1973, C-22(1): 67-92

[52] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA, 2014: 580-587

[53] Brutzer S, Höferlin B, Heidemann G. Evaluation of background subtraction techniques for video surveillance//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado, USA, 2011: 1937-1944

[54] Yao Jian, Odobez J. Multi-layer background subtraction based on color and texture//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, 2007: 1-8

[55] McFarlane N, Schofield C. Segmentation and tracking of piglets in images. Machine Vision and Applications, 1995, 8(3): 187-193

[56] Oliver N M, Rosario B, Pentland A P. A Bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 831-843

[57] McKenna S J, Jabri S, et al. Tracking groups of people. Computer Vision and Image Understanding, 2000, 80(1): 42-56

[58] Zivkovic Z, van der Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters, 2006, 27(7): 773-780

[59] Li Liyuan, Huang Weimin, Gu Irene Y H, et al. Foreground object detection from videos containing complex background //Proceedings of the 11th ACM International Conference on Multimedia. New York, USA, 2003: 2-10

[60] Maddalena L, Petrosino A. A self-organizing approach to background subtraction for visual surveillance applications. IEEE Transactions on Image Processing, 2008, 17(7): 1168-1177

[61] Barnich O, Van Droogenbroeck M. ViBE: A powerful random technique to estimate the background in video sequences// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei, China, 2009: 945-948

[62] Vedaldi A, Gulshan V, Varma M, Zisserman A. Multiple kernels for object detection//Proceedings of the IEEE 12th International Conference on Computer Vision. Kyoto, Japan, 2009: 606-613

[63] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York, USA, 2006: 2169-2178

[64] Zhang Junge, Yu Yinan, Huang Kaiqi, et al. Boosted local structured HOG-LBP for object localization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado, USA, 2011: 1393-1400

[65] Smeulders A W, Chu D M, Cucchiara R, et al. Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1442-1468

[66] He W, Yamashita T, Lu H, Lao S. SURF tracking//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 1586-1592

[67] Porikli F, Tuzel O, Meer P. Covariance tracking using model update based on Lie algebra//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 728-735

[68] Isard M, Blake A. Contour tracking by stochastic propagation of conditional density//Proceedings of the European Conference on Computer Vision. Cambridge, UK, 1996: 343-356

[69] Yilmaz A, Li X, Shah M. Object contour tracking using level sets//Proceedings of the IEEE Asian Conference on Computer Vision. Jeju, Korea, 2004

[70] Hager G, Dewan M, Stewart C. Multiple kernel tracking with SSD//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA, 2004: 790-797

[71] Comaniciu D, Ramesh V, Meer P. Real-time tracking of non-rigid objects using mean shift//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head Island, SC, 2000: 142-149

[72] Lucas B, Kanade T. An iterative image registration technique with an application to stereo vision//Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco, USA, 1981: 674-679

[73] Tomasi C, Kanade T. Detection and tracking of point features. Carnegie Mellon University: Technical Report CMU-CS-91-132, 1991

[74] Avidan S. Support vector tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001, 26(8): 184-191

[75] Avidan S. Ensemble tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, 29(2):494-501

[76] Babenko B, Yang M, Belongie S. Visual tracking with online multiple instance learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 983-990

[77] Kalal Z, Matas J, Mikolajczyk K. P-N learning: Bootstrapping binary classifiers by structural constraints//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 49-56

[78] Grabner H, Bischof H. Online boosting and vision//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 260-267

[79] Santner J, Leistner C, Saffari A, et al. PROST: Parallel robust online simple tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 723-730

[80] Wen L Y, Cai Z W, Lei Z, et al. Online spatio-temporal structural context learning for visual tracking//Proceedings of the European Conference on Computer Vision. Florence, Italy, 2012: 716-729

[81] Li Min. Object Tracking for Visual Surveillance [Ph. D. dissertation]. Institute of Automation, Chinese Academy of Sciences, Beijing, 2010(in Chinese)
(**李敏. 视觉监控中的目标跟踪研究[博士学位论文]. 中国科学院自动化研究所, 北京**, 2010)

[82] Yang C, Duraiswami R, Davis L. Efficicnet mean-shift tracking via a new similarity measure//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 176-183

[83] Han B, Davis L. Online density-based appearance modeling for object tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Beijing, China, 2005: 1492-1499

[84] Wang H, Suter D, Shen C. Adaptive object tracking based on an effective appearance filter. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(9): 1661-1667

[85] Ross D A, Lim J, Lin R, Yang M. Incremental learning for robust visual tracking. International Journal of Computer Vision, 2008, 77(1-3): 125-141

[86] Li X, Hu W, Zhang Z, et al. Robust visual tracking based on incremental tensor subspace learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Rio de Janeiro, 2007: 1-8

[87] Adam A, Rivlin E, Shimshoni I. Robust fragments-based tracking using the integral histogram//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 798-805

[88] Mei X, Ling H. Robust visual tracking using L1 minimization //Proceedings of the IEEE International Conference on Computer Vision. Miami, USA, 2009: 1436-1443

[89] Bar-Shalon Y, Fortmann T E. Tracking and Data Association. New York: Academic Press, 1998

[90] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition//Proceedings of the IEEE, 1989, 77(2): 257-286

[91] Isard M, Blake A. Condensation-conditional density propagation for visual tracking. International Journal of Computer Vision, 1998, 29(1): 5-28

[92] Ess A, Leibe B, Schindler K, et al, Robust multi person tracking from a mobile platform. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1831-1846

[93] Cox I J, Leonard J J. Modeling a dynamic environment using a Bayesian multiple hypothesis approach. Articial Intelligence, 1994, 66(2): 311-344

[94] Rasmussen C, Hager G D. Probabilistic data association methods for tracking complex visual objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 560-576

[95] Isard M, MacCormick J. Bramble: A Bayesian multipleblob tracker//Proceedings of the IEEE International Conference on Computer Vision. Vancouver, Canada, 2001: 34-41

[96] Khan S, Shah M. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. IEEE Transactions on Patterns Analysis and Machine Intelligence, 2003, 25(10): 1355-1360

[97] Black J, Ellis T. Multi-camera image tracking. Image and Vision Computing, 2006, 24(11): 1256-1267

[98] Chen Xiao-Tang. Object Tracking Across Multiple Cameras with Non-Overlapping Views [Ph. D. dissertation]. University of Chinese Academy of Sciences, Beijing, 2013(in Chinese)
(**陈晓棠. 非重叠场景下的跨摄像机目标跟踪研究[博士学位论文]. 中国科学院大学, 北京**, 2013)

[99] Javed O, Rasheed Z, Shafique K, Shah M. Tracking across multiple cameras with disjoint views//Proceedings of the IEEE International Conference on Computer Vision. Nice, 2003: 952-957

[100] Javed O, Shafique K, Shah M. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Computer Vision and Image Understanding, 2008, 109(2): 146-162

[101] Tieu K, Dalley G, Grimson W E L. Inference of non-overlapping camera network topology by measuring statistical dependence//Proceedings of the IEEE International Conference on Computer Vision. Beijing, China, 2005: 1842-1849

[102] Makris D, Ellis T, Black J. Bridging the gaps between cameras//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA, 2004: 205-210

[103] Zou Xiaotao, Bhanu B, Roy-Chowdhury A. Continuous learning of a multilayered network topology in a video camera network. EURASIP Journal on Image and Video Processing, 2009

[104] Niu Chaowei, Grimson E. Recovering non-overlapping network topology using far-field vehicle tracking data//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hong Kong, China, 2006: 944-949

[105] Javed O, Shafique K, Shah M. Appearance modeling for tracking in multiple non-overlapping cameras//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 26-33

[106] Jeong K, Jaynes C. Object matching in disjoint cameras using a color transfer approach. Machine Vision and Applications, 2008, 19(5-6): 443-455

[107] Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features//Proceedings of the European Conference on Computer Vision. Marseille, 2008: 262-275

[108] Zheng Wei-Shi, Gong Shaogang, Xiang Tao. Person re-identification by probabilistic relative distance comparison//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 649-656

[109] Dikmen M, Akbas E, Huang T S, Ahuja N. Pedestrian recognition with a learned metric//Proceedings of the Asian Conference on Computer Vision. Xi'an, China, 2010

[110] Montcalm T, Boufama B. Object inter-camera tracking with non-overlapping views: A new dynamic approach//Proceedings of the Canadian Conference Computer and Robot Vision. Ottawa, Canada, 2010: 354-361

[111] Collins R, Zhou X, Teh S K. An open source tracking testbed and evaluation web site//Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005). 2005

[112] Babenko B, Yang M-H, Belongie S. Robust object tracking with online multiple instance learning. IEEE Transactions on Patterns Analysis and Machine Intelligence, 2011, 33(7): 1619-1632

[113] Zhong W, Lu H, Yang M-H. Robust object tracking via sparsity-based collaborative model//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 1838-1845

[114] Hare S, Saffari A, Torr P H S. Sruck: Structured output tracking with kernels//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 263-270

[115] Kwon J, Lee K M. Visual tracking decomposition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 1269-1276

[116] Li Y, Huang C, Nevatia R. Learning to associate: Hybridboosted multi-target tracker for crowded scene//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, Florida, USA, 2009: 2953-2960

[117] Berclaz J, Fleuret F, Turetken E, et al. Multiple object tracking using k-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(9): 1806-1819

[118] Berclaz J, Fleuret F, Fua P. Multiple object tracking using flow linear programming//Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter). Utah, USA, 2009: 1-8

[119] Andriyenko A, Schindler K, Roth S. Discrete-continuous optimization for multi-target tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Rhode Island, USA, 2012: 1926-1933

[120] Breitenstein M D, Reichlin F, Leibe B, et al. Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(9): 1820-1833

[121] D'Angelo A, Dugelay J L. People re-identification in camera networks based on probabilistic color histograms//Proceedings of the SPIE. Bellingham, USA, 2011

[122] Chen X, Huang K, Tan T. Object tracking across non-overlapping views by learning inter-camera transfer models. Pattern Recognition, 2014, 47(3): 1126-1137

[123] Yu Yi-Nan. Image Understanding by Latent Variables [Ph. D. dissertation]. Institute of Automation, Chinese Academic of Sciences, Beijing, 2012(in Chinese)
(余轶南. 基于潜在变量的图像理解研究[博士学位论文]. 中国科学院自动化研究所, 北京, 2012)

[124] Huang Yong-Zhen. Object Classification and Detection Based on Visual Saliency [Ph. D. dissertation]. Institute of Automation, Chinese Academic of Sciences, Beijing, 2011 (in Chinese)
(黄永帧. 基于视觉显著性的目标分类与检测研究[博士学位论文]. 中国科学院自动化研究所, 北京, 2011)

[125] Csurka G, Dance C, Fan L, et al. Visual categorization with bags of keypoints//Proceedings of the European Conference on Computer Vision, Workshop on Statistical Learning in Computer Vision. Prague, Czech, 2004: 1-22

[126] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos//Proceedings of the 9th IEEE International Conference on Computer Vision. Wisconsin, USA, 2003: 1470-1477

[127] Van Gemert J C, Geusebroek J M, Veenman C J, et al. Kernel codebooks for scene categorization//Proceedings of the European Conference on Computer Vision (ECCV). Marseille, France, 2008: 696-709

[128] Olshausen B A, Field D J. Sparse coding with an overcomplete basis set: A strategy employed by V1. Vision Research, 1997, 37(23): 3311-3325

[129] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA, 2010: 3360-3367

[130] Huang Yongzhen, Huang Kaiqi, Yu Yinan, Tan Tieniu. Salient coding for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado, USA, 2011: 1753-1760

[131] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification//Proceedings of the European Conference on Computer Vision (ECCV). Crete, Greece, 2010, 6314: 143-156

[132] Zhou Xi, Yu Kai, Zhang Tong, Huang T S. Image classification using super-vector coding of local image descriptors//Proceedings of the European Conference on Computer Vision(ECCV). Berlin, Germany, 2010: 141-154

[133] Hubel D H, Wiesel T N. Receptive fields of single neurones in the cat's striate cortex. The Journal of Physiology, 1959, 148(3): 574

[134] Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics, 1988, 59(4-5): 291-294

[135] Smolensky P. Chapter 6: Information processing in dynamical systems: Foundations of harmony theory//Proceedings of the Parallel Distributed: Explorations in the Microstructure of Cognition, Volume 1: Foundations. USA: MIT Press, 1986

[136] Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527-1554

[137] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[138] Huang Yongzhen, Huang Kaiqi, Tao Dacheng, et al. Enhanced biologically inspired model for object recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2011, 41(6): 1668-1680

[139] Li Fei-Fei, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories//Proceedings of the Computer Vision and Pattern Recognition (CVPR), Workshop on Generative-Model Based Vision. Washington, USA, 2004: 178

[140] Griffin G, Holub A, Perona P. The Caltech 256. Caltech Technical Report

[141] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303-338

[142] Deng Jia, Dong Wei, Socher R, et al, ImageNet: A large-scale hierarchical image database//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). Miami, USA, 2009: 248-255

[143] Bai X, Liu W, Tu Z. Integrating contour and skeleton for shape classification//Proceedings of the IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops). Kyoto, Japan, 2009: 360-367

[144] Gao S, Tsang I W, Chia L T, et al. Local features are not lonely — Laplacian sparse coding for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA, 2010: 3555-3561

[145] Yang J, Li Y, Tian Y, et al. Group-sensitive multiple kernel learning for object categorization//Proceedings of the IEEE 12th International Conference on Computer Vision. Kyoto, Japan, 2009: 436-443

[146] Huang Yongzhen, Wu Zifeng, Wang Liang, Tan Tieniu. Feature coding in image classification: A comprehensive study. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(3): 493-506

[147] LeCun Y, Ranzato M A. Deep learning tutorial. Atlanta: International Conference on Machine Learning (ICML), Technical Report, 2013

[148] Yao Bangpeng, Li Fei-Fei. Action recognition with exemplar based 2.5 D graph matching//Proceedings of the European Conference on Computer Vision (ECCV). Firenze, Italy, 2012: 173-186

[149] Khan F S, Anwer R M, van de Weijer J, et al. Coloring action recognition in still images. International Journal of Computer Vision, 2013, 105(3): 205-221

[150] Yao Bangpeng, Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(9): 1691-1703

[151] Hogg D. Model-based vision: A program to see a walking person. Image and Vision Computing, 1983, 1(1): 5-20

[152] O'Rourke J, Badler N. Model-based image analysis of human motion using constraint propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1980, 2(6): 522-536

[153] Rogez G, Orrite-Uruñuela C, Martínez-del-Rincón J. A spatio-temporal 2D-models framework for human pose recovery in monocular sequences. Pattern Recognition, 2008, 41(9): 2926-2944

[154] Tautges J, Zinke A, Krüger B, et al. Motion reconstruction using sparse accelerometer data. ACM Transactions on Graphics, 2011, 30(3): 18:1-18:12

[155] Shotton J, Sharp T, Kipman A, et al. Real-time human pose recognition in parts from a single depth image//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 1297-1304

[156] Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 1385-1392

[157] Johansson G. Visual motion perception. Scientific American, 1975, 232(6): 76-89

[158] Bobick A F, Davis J W. The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(3): 257-267

[159] Ke Y, Sukthankar R, Hebert M. Spatio-temporal shape and flow correlation for action recognition//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[160] Laptev I, Lindeberg T. Space-time interest points//Proceedings of the IEEE International Conference on Computer Vision. Nice, France, 2003: 432-439

[161] Everts I, van Gemert J C, Gevers T. Evaluation of color stips for human action recognition//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2850-2857

[162] Willems G, Tuytelaars T, Van Gool L. An efficient dense and scale-invariant spatio-temporal interest point detector//Proceedings of the European Conference on Computer Vision. Marseille, France, 2008: 650-663

[163] Yuan C, Li X, Hu W, et al. 3D R transform on spatio-temporal interest points for action recognition//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 724-730

[164] Han D, Bo L, Sminchisescu C. Selection and context for action recognition//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 1933-1940

[165] Kovashka A, Grauman K. Learning a hierarchy of discriminative spacetime neighborhood features for human action recognition//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 2046-2053

[166] Campbell L W, Bobick A F. Recognition of human body motion using phase space constraints//Proceedings of the IEEE International Conference on Computer Vision. Boston, USA, 1995: 624-630

[167] Lv F, Nevatia R. Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost//Proceedings of the European Conference on Computer Vision. Graz, Austria, 2006: 359-372

[168] Huang K, Zhang Y, Tan T. A discriminative model of motion and cross ratio for view-invariant action recognition. IEEE Transactions on Image Processing, 2012, 21(4): 2187-2197

[169] Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 104-111

[170] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 2013, 103(1): 60-79

[171] Wang C, Wang Y, Yuille A L. An approach to pose-based action recognition//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 915-922

[172] Wang J, Liu Z, Wu Y, et al. Mining actionlet ensemble for action recognition with depth cameras//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 1290-1297

[173] Tran D, Sorokin A. Human activity recognition with metric learning//Proceedings of the European Conference on Computer Vision. Marseille, France, 2008: 548-561

[174] Wang H, Schmid C. Action recognition with improved trajectories//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 3551-3558

[175] Nguyen N T, Phung D Q, Venkatesh S, et al. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 955-960

[176] Shi Y, Huang Y, Minnen D, et al. Propagation networks for recognition of partially ordered sequential action//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, USA, 2004: II-862-II-869

[177] Damen D, Hogg D. Recognizing linked events: Searching the space of feasible explanations//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 927-934

[178] Tran S D, Davis L S. Event modeling and recognition using Markov logic networks//Proceedings of the European Conference on Computer Vision. Marseille, France, 2008: 610-623

[179] Ivanov Y A, Bobick A F. Recognition of visual activities and interactions by stochastic parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 852-872

[180] Joo S W, Chellappa R. Attribute grammar-based event recognition and anomaly detection//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. New York, USA, 2006: 107-107

[181] Zhang Z, Tan T, Huang K. An extended grammar system for learning and recognizing complex visual events. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(2): 240-255

[182] Vu V T, Bremond F, Thonnat M. Automatic video interpretation: A novel algorithm for temporal scenario recognition//Proceedings of the International Joint Conference on Artificial Intelligence. Acapulco, Mexico, 2003: 1295-1300

[183] Moore D J, Essa I A, Hayes III M H. Exploiting human actions and object context for recognition tasks//Proceedings of the IEEE International Conference on Computer Vision. Kerkyra, Greece, 1999: 80-86

[184] Gupta A, Davis L S. Objects in action: An approach for combining action understanding and object perception//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[185] Gong Shaogang, Xiang Tao. Recognition of group activities using dynamic probabilistic networks//Proceedings of the IEEE International Conference on Computer Vision. Nice, France, 2003: 742-749

[186] Zhang D, Gatica-Perez D, Bengio S, et al. Modeling individual and group actions in meetings with layered HMMs. IEEE Transactions on Multimedia, 2006, 8(3): 509-520

[187] Dai P, Di H, Dong L, et al. Group interaction analysis in dynamic context. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2008, 38(1): 275-282

[188] Vaswani N, Chowdhury A R, Chellappa R. Activity recognition using the dynamics of the configuration of interacting objects//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, USA, 2003: II-633-II-640

[189] Blank M, Gorelick L, Shechtman E, et al. Actions as space-time shapes//Proceedings of the IEEE International Conference on Computer Vision. Beijing, China, 2005: 1395-1402

[190] Schuldt C, Laptev I, Caputo B. Recognizing human actions: A local svm approach//Proceedings of the International Conference on Pattern Recognition. Cambridge, UK, 2004: 32-36

[191] Liu J, Luo J, Shah M. Recognizing realistic actions from videos "in the wild"//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1996-2003

[192] Marszalek M, Laptev I, Schmid C. Actions in context//Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 2929-2936

[193] Rodriguez M D, Ahmed J, Shah M. Action mach: A spatio-temporal maximum average correlation height filter for action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8

[194] Weinland D, Boyer E, Ronfard R. Action recognition from arbitrary views using 3D exemplars//Proceedings of the IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007: 1-7

[195] Niebles J C, Chen C W, Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification//Proceedings of the European Conference on Computer Vision. Heraklion, Greece, 2010: 392-405

[196] Reddy K K, Shah M. Recognizing 50 human action categories of web videos. Machine Vision and Applications, 2013, 24(5): 971-981

[197] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, 2012, abs/1212.0402

[198] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 2556-2563

[199] Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3D points//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. San Francisco, USA, 2010: 9-14

[200] Kurakin A, Zhang Z, Liu Z. A real time system for dynamic hand gesture recognition with a depth sensor//Proceedings of the European Signal Processing Conference. Bucharest, Romania, 2012: 1975-1979

[201] Huang T. Surveillance video: The biggest big data. Computing Now, 2014, 7(2)

[202] 973Project: Public safety oriented social perception data processing, IA, CAS, 2011(in Chinese)
(973 项目：面向公共安全的社会感知数据处理，IA，CAS，2011)

[203] Chen D M, Baatz G, Koser K, et al. City-scale landmark identification on mobile devices//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 737-744

[204] Song Y C, Zhang Y D, Cao J, et al. Web video geolocation by geotagged social resources. IEEE Transactions on Multimedia, 2012, 14(2): 456-469

[205] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2547-2554

[206] Tanenbaum A S, Van Steen M. Distributed Systems Principles and Paradigms. 2nd Edition. New Jersey: Prentice Hall, 2006

[207] Durmus Y, Ozgovde A, Ersoy C. Distributed and online fair resource management in video surveillance sensor networks. IEEE Transactions on Mobile Computing, 2012, 11(5): 835-848

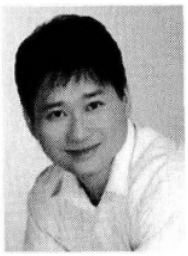
[208] Atrey P K, Hossain M A, El Saddik A, et al. Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 2010, 16(6): 345-379

[209] Bhatt C A, Kankanhalli M S. Multimedia data mining: State of the art and challenges. Multimedia Tools and Applications, 2011, 51(1): 35-76

[210] Villars R L, Olofson C W. Eastwood M. Big data: What it is and why you should care. White Paper, IDC, 2011

[211] Tian Y, Srivastava J, Huang T, et al. Social multimedia computing. Computer, 2010, 43(8): 27-36

[212] Konstantinou N, Solidakis E, Zafeiropoulos A, et al. A context-aware middleware for real-time semantic enrichment of distributed multimedia metadata. Multimedia Tools and Applications, 2010, 46(2-3): 425-461



HUANG Kai-Qi, born in 1977, Ph. D., professor. His research interests include computer vision, pattern recognition and visual surveillance.

CHEN Xiao-Tang, born in 1987, Ph. D., assistant professor. Her research interests include computer vision and pattern recognition.

KANG Yun-Feng, born in 1981, engineer. His research interest is visual surveillance.

TAN Tie-Niu, born in 1964, Ph. D., professor, member of Chinese Academy of Sciences. His research interests include biometrics, image and video analysis, information forensics and security.

Background

Due to the rapid increase of the number of cameras used in the video surveillance and the high needs of the smart city and public security, video surveillance by human beings is no longer suitable. Hence, intelligent video surveillance emerges and becomes one of the hottest research point. Intelligent video surveillance is an interdisciplinary research area that has abundant research interests and diverse applications. This paper summarizes the history, the state-of-the-art, and various applications of the intelligent videosurveillance. Firstly, this paper classifies the algorithms by low level, middle level and high level, then analyses their advantages and disadvantages, compares the performances on different datasets, and presents the outstanding issues; finally, we discuss the future research trends of intelligent video surveillance in the context of Internet of Things.

This work is funded by the National Basic Research Program (973 Program) of China (Grant No. 2012CB316302), the National Natural Science Foundation of China (Grant No. 61322209), the National Key Technology R&D Program (Grant No. 2012BAH07B01).