

多摄像机之间基于区域 SIFT 描述子的目标匹配

明安龙 马华东

(北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876)

摘 要 提出了一种多摄像机之间的目标匹配方法,摄像机可以带有云台.该方法是基于区域的方法,但是区域的特征以 SIFT 描述子而不是通常的颜色来描述,同时目标的检测使用减背景技术,目标跟踪则选用粒子滤波.该文的匹配方法不需要摄像机之间的合作,也不要求目标物体处于同一地平面上.文中方法的主要特点还表现在:(1)无几何约束的需求,同一目标物体的背景完全切换后,也可以进行匹配;(2)可以匹配各种类型的目标物体;(3)摄像机在目标跟踪期间可以简单运动(通过云台);(4)适合分布式计算,但也可以集中式处理;(5)容忍亮度的变化.实验结果证明作者的方法是有效的.

关键词 SIFT 描述子;多摄像机;匹配;多目标检测与跟踪

中图法分类号 TP391

Region-SIFT Descriptor Based Correspondence Between Multiple Cameras

MING An-Long MA Hua-Dong

(Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract This paper proposes a region-SIFT descriptor based target matching method for multiple cameras. This is a region based method. However, the region is represented by SIFT descriptor instead of traditional color features. In the new method, the background subtraction is used in multi-target detecting, and the particle filter is utilized in multi-target tracking. Non camera calibrations are required in the new method, neither the constraint that all objects stand in the same plane. The main features of the method are highlighted as follows: (1) Non geometry constraints are required to some extent. (2) Many types of objects are supported. (3) The cameras can simply move during object tracking. (4) It is more fit to distributed computing, but the traditional centralized computing is also supported. (5) It is robust to changes of the light intensity. Experimental results show that the method is effective.

Keywords region-SIFT descriptor; multiple cameras; correspondence; multi-target detection and tracking

1 引 言

在计算机视觉领域的一些应用中,如视频监控、

跟踪、智能环境监测等,经常会部署并使用多个摄像机.多个摄像机不仅扩展了采集信息的区域,而且从不同视角采集的视觉信息有助于解决一些特定问题,如遮挡、物体分类、3D 建模等.但是多个摄像机

的使用也会带来一系列新问题,包括多摄像机之间的目标匹配^[1-4]、摄像机协作、摄像机之间的自动切换和数据融合等。其中,多摄像机之间的目标匹配是指在不同的图像序列中找到目标间的对应关系,其结果直接影响后续的数据融合,因此它是多摄像机应用中最重要和基础的问题之一。本文主要研究多摄像机之间的目标匹配问题。

研究多摄像机之间的目标匹配的问题可以概括为:在保障目标匹配正确率的前提下,尽可能地减少约束条件。约束条件包括计算量、场景情况、重叠情况、遮挡情况、摄像机参数设置、摄像机标定等。近年来,相关领域的研究人员陆续提出了一些方法来尝试解决这个问题^[1-6]。这些方法一般都是基于宽基线立体视觉算法^[7](Wide-baseline Stereo algorithm),并且按照所使用视觉特征的类型^[1],可分为:基于区域的方法和基于点的方法。此外还有一些非主流的分类方法,如 Avidan^[6]等考虑到智能摄像机有一定的处理和通信能力,将多视角目标匹配方法分为集中式方法和分布式方法。

基于区域的方法将目标视为区域,利用区域的特征在多视角中进行匹配。区域的特征一般与颜色相关。Orwell 等^[8]提出了使用颜色直方图作为区域特征,Krumm 等^[3]也提出了类似的方法。Mittal 等^[4]使用高斯颜色模型来解决多个摄像机之间的匹配问题。Chang 等^[2]则结合对极几何(epipolar geometry)和颜色映射等来建立两个摄像机之间的匹配。然而,颜色特征的使用带来新的问题,主要有:(1)当目标的颜色相同或相近时容易产生“误配”;(2)视角的光线变化时,颜色随之产生差异;(3)不同类型的摄像机或不同的参数设置,导致对同一目标采集颜色不尽相同;(4)目标本身的颜色不单一,不同视角可能对应不同的颜色。因此,通常在基于区域特征的方法中使用的颜色特征是不可靠的,需结合其它的考虑。然而文献[2]结合对极几何的方法又使得匹配依赖于摄像机的参数设置,即产生新的约束条件。

基于点的方法是一个相对实用的方法。该方法的一般做法是根据几何约束对目标的特征点进行匹配,从而实现多摄像机的目标匹配。这里的几何约束是 3D 或 2D 的:

(1)依据 3D 几何约束的方法。Tsutsui 等^[7]和 Utsumi 等^[9]以目标物体的质心作为特征点,并将不同的特征点映射到同一个 3D 空间中,然后通过比

较这些特征点在同一个坐标系中的位置来建立目标物体的匹配关系。Kelly 等^[10]由摄像机之间的协作来估计目标在环境 3D 模型中的位置,并通过估计得到的位置信息进行匹配,但是这种方法需要事先知道关于环境的 3D 模型知识。Cai 等^[11]使用邻接的两个摄像机之间的协作来提取对极几何约束,然后通过人体的上半身中线上的特征点进行匹配。这些方法的共同之处在于需要较多的先验知识(如环境的 3D 模型)或者摄像机之间的协作。另外,从不同视角提取的同一目标的特征点并不能确保对应到 3D 空间中的同一点,以此为基础的匹配因此有一定的任意性。

(2)依据 2D 几何约束的方法。Khan 等^[12],Black 等^[13]均根据基于地表平面的单应矩阵(homography)约束进行匹配。Khan 等提出使用人体膝盖上的点进行匹配。但是除非用专用的传感器,否则人体膝盖点的精确定位本身就是一个很大的挑战。此外,在不同的视角中,人体膝盖可能会因遮挡而不可见。Black 等提出使用人体的质心点进行匹配,也有同样的问题。

此外,Hu 等^[1]提出了一种基于主轴线的多摄像机中目标人物的匹配方法。其做法是将人体区域以最小外接矩形框限定后取其主轴线,并以此为基础进行匹配。我们认为这种方法属于一种特殊的点的匹配,该“点”即主轴线与地平面交点(ground point)。Hu 等的方法无需摄像机之间的合作。通过选取人体区域主轴线,该方法一定程度上可以降低人体区域检测误差对匹配造成的影响,而且匹配的结果反过来有助于估计某些视角中人体被遮挡时的位置。但是,Hu 等的方法基于如下的一些假设或限定:(1)目标人物处于同一地平面,如不允许存在阶梯;(2)没有说明人物的影子对检测主轴的负面影响;(3)多个摄像机是从不同视角但却是同时对相同场景进行图像采集,需满足同时性和一定的几何约束;(4)只考虑了人物和车两种目标,或者可以扩展到基本对称的物体;(5)人体是直立的。此外,该方法还利用卡尔曼滤波器针对每个人物进行目标跟踪,但是依据文献[21],经典卡尔曼滤波只能处理线性、高斯、单模态的情况,而实际的视觉跟踪过程中,后验概率的分布往往是非线性、非高斯、多模态的,并且日常的监控用摄像机多带有云台而不是完全固定,因此多目标视觉跟踪不宜直接使用卡尔曼滤波算法。

针对已有的匹配方法的回顾和分析,本文提出了一种多摄像机之间基于区域 SIFT 描述子的目标匹配方法,为了验证本文提出的方法,我们主要在 PETS 系列监控数据库上进行效果测试,PETS 是国际著名的开放监控视频库.

2 方法概述

本文的目的是为了实现一个简单而有效的多摄像机目标匹配方法.为此我们采用基于区域的匹配方法,但在表示区域特征的时候引入 Lowe^[16]的 SIFT 描述子而不是通常使用的颜色特征. SIFT 描述子是对 SIFT 特征的上层描述,一般应用于物体的分类和识别中,也可用来构建全景图^[17]. SIFT 描述子有两个扩展:PCA-SIFT 和 GLOH^[18],但是因为扩展后投影矩阵需要一系列有代表性的图像,这

个矩阵只对这类图像起作用,为了减少多摄像机之间目标匹配的几何约束条件,我们采用原始的 SIFT 描述子,这样需要较少经验知识. SIFT 描述子具有一些很好的特点,如对旋转、尺度缩放、亮度变化保持不变性;对视角变化、仿射变换、噪声也保持一定程度的稳定性;而且经优化的 SIFT 描述子匹配算法甚至可以达到实时的要求. 又因为匹配时传递特征而不是原始数据,通信负载较小,适合分布式计算的需要.

本文方法的基本流程是:

(1)对每个摄像机所采集的图像序列,我们首先使用减背景技术进行运动分割,接着通过分割目标物体得到多个目标运动区域,然后使用粒子滤波来进行多个目标的跟踪,跟踪的结果用于修正后续图像序列中的多个目标区域(如发生部分遮挡时). 图 1 是目标区域检测和跟踪的过程.

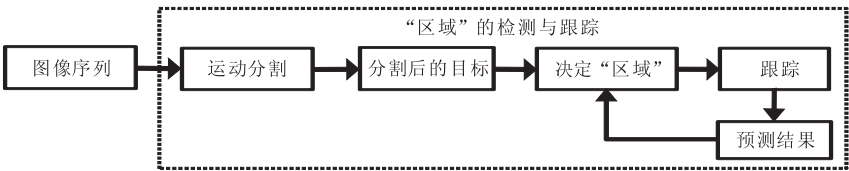


图 1 单个摄像机中的目标区域检测和跟踪

(2)多摄像机之间的目标匹配过程则可以分为两种情形:分布式处理模式和集中式处理模式.这两种情形中匹配的算法是完全一样的,其最大区别在于分布式处理模式的匹配是在待匹配的两个摄像机之一中进行,而集中式处理模式的匹配则是在后台的视频服务器中进行.

(a)分布式处理.两个摄像机之间通信,一个摄像机将自己的区域特征传给另一个摄像机.另一个摄像机接受前一个摄像机传来的区域特征并完成目

标匹配,然后将结果传回前一个摄像机,匹配的结果反过来优化这两个摄像机目标区域的检测和跟踪. 分布式处理的过程如图 2 所示,其中, Camera 1 和 Camera n 的角色是可以互换的,匹配的处理在摄像机内部完成. 当系统中存在大量的摄像机时,分布式处理显得尤为必要. 不过分布式模式要求摄像机有一定处理能力和通信能力. 这种摄像机被称为智能摄像机(smart camera^①),智能摄像机是一个当前研究的热点.

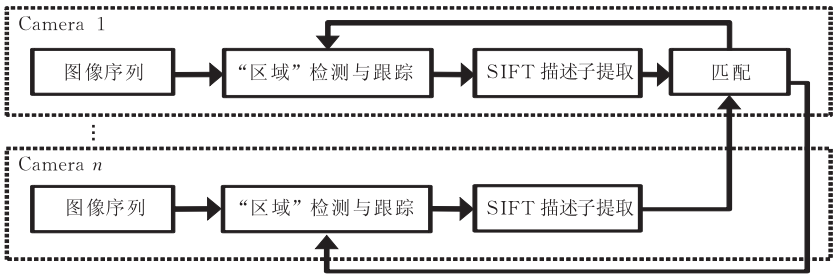


图 2 分布式处理中多摄像机之间的目标匹配

(b)集中式处理.每个摄像机与后台的视频服务器通信,采集的原始图像序列作为通信的内容传到后台进行目标区域的检测和跟踪、SIFT 描述子的提取和目标的匹配.匹配的结果用于优化目标区域

的检测与跟踪.集中式处理的过程如图 3 所示.

① Intelligent design with smart cameras, 2007. <http://www.vision-components.com>.

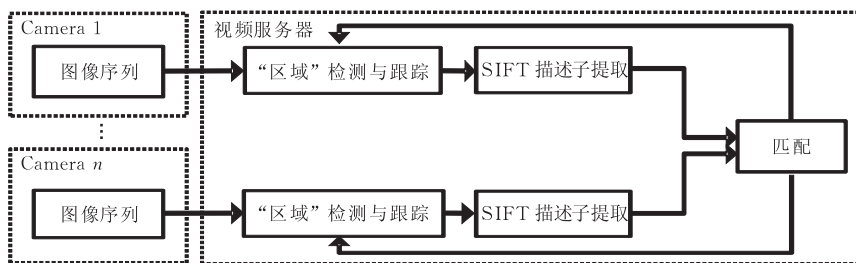


图 3 集中式处理中多摄像机之间的目标匹配

3 目标区域的检测和跟踪

通过减背景技术实现前后景分离,以此为基础进行运动分割,进而检测出目标区域,然后使用粒子滤波进行跟踪。

3.1 运动分割

通过减背景技术进行运动分割是一个常用的方法^[21],包括时间差分法(temporal difference)、中值滤波法(average\medium filter)、线性预测法(linear)、混合高斯法(mixture of Gauss)、基于均值替换的背景估计法(mean-shift based estimation)等.考虑到处理速度和背景扰动,本文中采用中值滤波法进行前后景分离,即建立一个视缓冲区用来缓存视频的 L 帧,然后把缓冲区中 L 帧的同位置像

素的平均值或中值作为背景中该处像素的值:

$$B_{t+1}(x,y) = \text{median}(I_t(x,y), \dots, I_{t-L}(x,y)) \quad (1)$$

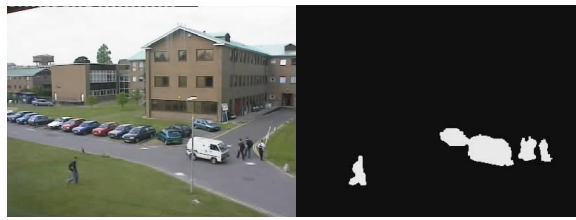
其中, $B(x,y)$ 和 $I(x,y)$ 是背景以及缓存中某帧在像素点 (x,y) 处的灰度值.显然算法需要 L 帧的缓存,为了降低这个需求,有必要引入学习率 λ ^[20] 来体现背景变化对场景变化的响应,通常取 $\lambda=0.05$, λ 值越小,前景的变化对背景的影响也越小.

$$\begin{cases} B_{t+1}(x,y) = B_t(x,y), & \text{当 } I_t(x,y) \text{ 是前景} \\ B_{t+1}(x,y) = \lambda I_t(x,y) + (1-\lambda)B_t(x,y), & \text{当 } I_t(x,y) \text{ 是背景} \end{cases} \quad (2)$$

我们选用 2 段视频来给出运动分割的例子,一个是室内监控,一个是室外监控.图 4(a),(b) 分别给出了室内监控的视频和 2058 帧的运动分割的例子.



(a) 大楼监控视频



(b) PETS2001 监控视频

图 4 2 个运动分割的例子(一个是室内监控,一个是室外监控)

3.2 跟踪

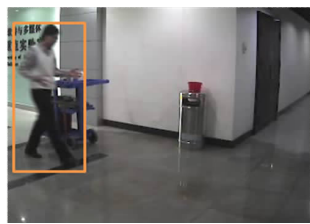
运动分割后我们需要对分割后的区域进行跟踪.考虑到运动分割后,特定情况下摄像机会主动通过云台来转动以扩大监控视野,再加上可能会发生其它的运动物体和群聚的背景,我们选取粒子滤波进行目标跟踪。

粒子滤波的主要思想是用一组具有权值的粒子来完全地描述后验概率分布,根据蒙特卡罗理论,当粒子的数目足够多,这组具有权值的粒子就能完全地描述后验概率分布,此时粒子滤波就是最优的贝叶斯估计.本文中我们使用隐马尔可夫模型来对跟踪进行预测,以图像数据为观察值,目标的位置和大小作为隐藏的状态。

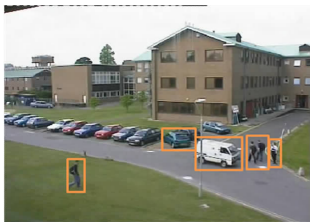
$$\underbrace{p(x_t | y_{1:t})}_{\text{当前目标状态}} = \underbrace{\alpha p(y_t | x_t)}_{\text{观察模型}} \int \underbrace{p(x_t | x_{t-1})}_{\text{转移模型}} \underbrace{p(x_{t-1} | y_{1:t-1})}_{\text{前一时刻目标状态}} dx_{t-1} \quad (3)$$

初始状态 $p(x_0)$ 表示目标初始的状态(位置和大小),本文中运动分割的结果作为 $p(x_0)$. $p(x_t | x_{t-1})$ 表示转移模型,描述了相临两帧图像之间目标的移动,一个简单的办法就是通过对当前状态附近的高斯窗口进行采样得到,更好的办法则是考虑前面状态的速度和加速等信息,本文中转移模型采用了一个二阶导的动态模型,考虑了前两帧的状态和噪声,设系数 A_1, A_2 和 B_0 , $X_t - \bar{X} = A_2(X_{t-1} - \bar{X}) +$

$A_2(X_{t-1} - \bar{X}) + B_0 w_t$, 其中 X_t 表示序列为 t 的一帧, w_t 为一些噪声. $p(y_t | x_t)$ 表示观察模型, 描述了目标处于某种状态(即位于某位置和大小)的相似程度, 本文中采用了一个简单的基于 HSV 直方图的观察模型. 图 5(a), (b) 分别给出了一个室内监控和一个室外监控视频片段的例子.



(a) 大楼监控视频



(b) PETS2001 监控视频

图 5 2 个目标跟踪的例子

3.3 区域检测

区域检测一般可分为三种情况: 单个目标、多个目标和遮挡. 因为我们匹配的目标是各种类型的物体, 所以我们不刻意区分多个目标和遮挡. 我们约

定: 一个从前面的视频帧跟踪到当前帧的目标称为 Tracked Object, 简称 TO; 而从当前帧开始运动分割出来的目标称为 Detected Object, 简称 DO, 那么区域检测可以表述如下: (1) 如果只存在一个 TO, 并且这个 TO 对应到一个 DO, 那么 DO 区域为一个独立的目标区域; (2) 如果存在多个 TO 对应到一个 DO, 说明发生群聚或遮挡, 那么对应到这个 DO 的多个目标以 TO 区域为目标区域; (3) 如果存在一个 DO 没有任何一个 TO 与之对应, 说明有新的目标加入, 设这个 DO 为 TO, DO 区域为目标区域.

4 SIFT 描述子的提取

SIFT 算法是一种提取局部特征的算法, 它在尺度空间寻找极值点, 提取位置、尺度和旋转不变量. 图 6 显示了 SIFT 描述子的提取步骤: (1) 检测尺度空间极值点; (2) 精确定位极值点; (3) 为每个关键点指定方向参数; (4) SIFT 描述子的生成. 尺度空间模拟图像数据的多尺度特征, 一幅二维图像的尺度空间定义为 $L(x, y, \sigma) = G(x, y, \sigma) \times I(x, y)$. 其中 $G(x, y, \sigma)$ 是尺度可变高斯函数且 $G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$, (x, y) 是空间坐标, σ 是尺度坐标, σ 值越小, 则表征该图像被平滑得越少, 相应的尺度也就越小. 大尺度对应于图像的概貌特征, 小尺度对应于图像的细节特征.

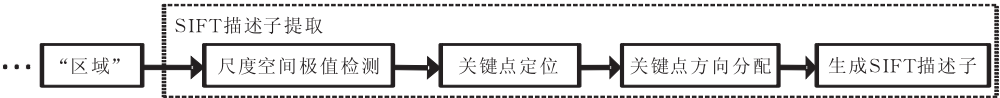


图 6 SIFT 描述子的提取步骤

4.1 空间极值点坐标检测

满足在图像二维平面空间和 DoG^[19] (Difference of Gaussian) 尺度空间中同时具有局部极值的点作为 SIFT 关键点. DoG 算子定义为两个不同尺度的高斯核的差分:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \times I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{4}$$

通过计算某采样点在每一尺度下 DoG 算子的值, 可以得到特征尺度轨迹曲线. 特征尺度曲线的局部极值点即为该采样点的尺度. 为了寻找尺度空间的极值点, 每一个采样点要和它所有的相邻点比较, 看其是否比它的图像域和尺度域的相邻点大或者小, 一般采样点要和它处于同一尺度的 8 个相邻点和上下相邻尺度对应的 9×2 个点共 26 个点比较,

以确保在尺度空间和二维图像空间都检测到极值点.

4.2 精确定位极值点

上文通过拟和三维二次函数确定了关键点的位置和尺度(达到亚像素精度). 然而因为 DoG 算子会产生较强的边缘响应, 所以 SIFT 算法需要舍弃低对比度的关键点和不稳定的边缘响应点以增强匹配稳定性和提高抗噪声能力. 舍弃关键点的依据是: 一个定义不好的 DoG 的极值在横跨边缘的地方有较大的主曲率, 而在垂直边缘的方向有较小的主曲率. 而 DoG 的主曲率通过一个 2×2 的 Hessian 矩阵 H 求出: $H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}$, DoG 的主曲率和 H 的特征值成正比, 令 α 为最大特征值, β 为最小的特

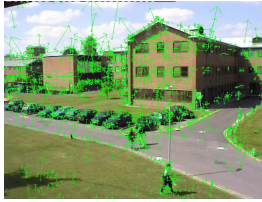
征值,则 $\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(\gamma\beta+\beta)^2}{\gamma\beta^2} = \frac{(\gamma+1)^2}{\gamma}$.
 $Tr(\mathbf{H}) = D_{xx} + D_{yy} = \alpha + \beta$ 和 $Det(\mathbf{H}) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta$ 分别表示 \mathbf{H} 的迹和行列式, $\alpha = \gamma\beta$. 因为 $\frac{(\gamma+1)^2}{\gamma}$ 的值在两个特征值相等的时候最小,随着 γ 的增大而增大,所以检测主曲率是否在某值 γ 下,只需检测 $\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} < \frac{(\gamma+1)^2}{\gamma}$. 一般 $\gamma=10$.

4.3 关键点方向分配

SIFT 算法利用关键点邻域像素的梯度方向分布特性为每个关键点指定方向参数,使算子具备旋转不变性. 设

$$m(x, y) = ((L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2)^{1/2}$$

表示采样点 (x, y) 处的模值,而采样点 (x, y) 处的方向 $\theta(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y)))$. 其中 L 的尺度为每个关键点各自所在的尺度. 实际计算中,一般在以关键点为中心的邻域窗口内采样,并用直方图统计邻域像素的梯度方向. 梯度直方图的范围是 $0^\circ \sim 360^\circ$,其



(a) Camera 1 的第1566帧图像 (b) 在(a)中检测出的2491个SIFT特征 (c) Camera 2 的第1566帧图像 (d) 在(c)中检测出的3520个SIFT特征

图7 PETS2001 中 2 个不同视角图像的 SIFT 特征提取的例子

SIFT 描述子是对一个 SIFT 特征区域的描述,其生成步骤如下:

(1) 首先将坐标轴旋转为 SIFT 特征区域的方向,以确保旋转不变性.

(2) 以这个 SIFT 特征区域的关键点位置为中心取 8×8 的窗口. 如图 8 左边部分所示,每个小窗口代表 SIFT 特征区域关键点邻域所在尺度空间的一个像素,箭头方向代表该像素的梯度方向,箭头长度代表该点的梯度模值,图中圈代表高斯加权的范围(越靠近关键点的像素梯度方向信息贡献越大). 在每 4×4 的小块上计算 8 个方向的梯度方向直方图,绘制每个梯度方向的累加值,即可形成一个种子点. 如图 8 右边部分所示,一个关键点由 2×2 共 4 个种子点组成,每个种子点有 8 个方向向量信息,这样对于一个关键点就可以产生 32 个数据,即最终形成 32 维的 SIFT 描述子. 此时 SIFT 描述子已经去除

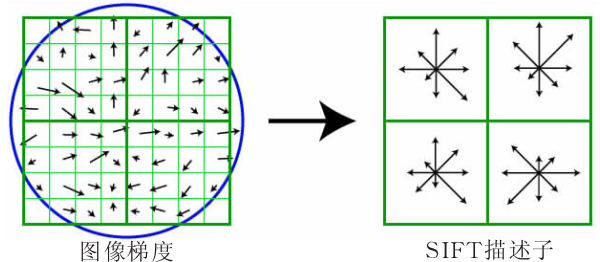
中每 10° 一个柱,总共 36 个柱. 直方图的峰值则代表了该关键点处邻域梯度的主方向,即作为该关键点的方向. 在梯度方向直方图中,当存在另一个相当于主峰值 80% 能量的峰值时,则将这个方向认为是该关键点的辅方向. 因此一个关键点可能会被指定具有多个方向(一个主方向和多个辅方向),这可以增强匹配的鲁棒性. 然而在本文中因摄像机处理能力的限制,在分布式模式中不考虑辅方向.

4.4 SIFT 描述子的生成

通过 4.1~4.3 的描述,我们可以检测出图像的 SIFT 关键点,每个关键点有三个信息:位置、所处尺度和方向,由此可以确定一个 SIFT 特征区域.

图 7 给出了 PETS2001 数据库的 dataset 3 中 2 个不同视角图像的 SIFT 特征提取的例子. 图 7(a) 是 Camera 1 拍摄的,图 7(c) 是 Camera 2 拍摄的,从图 7(a), (c) 可以看出两个摄像机视角明显不同. 图 7(b), (d) 显示了所检测出的 SIFT 特征,其中,交叉点表示关键点的位置,长度表示模值,箭头表示方向.

了尺度变化、旋转等几何变形因素的影响. 继续将 SIFT 描述子的长度归一化以进一步去除光照变化的影响. 这种邻域方向性信息联合的思想增强了算法抗噪声的能力,同时对于含有定位误差的特征匹配也提供了较好的容错性. 在本文的集中式模式中,因为视频服务器的处理能力较强,为了增强匹配的稳健性,可以考虑对每个 SIFT 特征区域的关键点



图像梯度

SIFT描述子

图8 由 SIFT 特征区域的关键点邻域梯度信息生成 SIFT 描述子

使用 4×4 共 16 个种子点来描述, 这样 SIFT 描述子的维度就是 128.

5 目标匹配

5.1 SIFT 描述子之间的匹配

两个 SIFT 描述子之间的匹配采用最短欧式距离的方法. 但是考虑到 32 维乃至 128 维的数据, 在进行欧式距离比较时, Lowe 使用了 BBF^[19] (the Best-Bin-First algorithm) 来对传统的 k - d 树算法做逼近, 因为 k - d 树算法一般解决不超过 10 维的数据. BBF 算法之所以更快速是因为仅考虑那些次最邻近点的距离 0.8 倍以内的最邻近点, 这样就规避了麻烦的多个距离相近邻接点的问题.

我们对图 7(b), (d) 的 SIFT 特征按照 4.1 的算法提取 SIFT 描述子, 并进行 SIFT 描述子的直接匹配, 共成功匹配 102 对, 如图 9 所示, 从中可以直观地看出, 误配的情况比较多, 而且我们关注的运动目标区域存在零匹配的情况. 此外, 图 9 还存在大量的我们不关心区域的匹配计算. 所以直接进行 SIFT 描述子的匹配是不合适的.

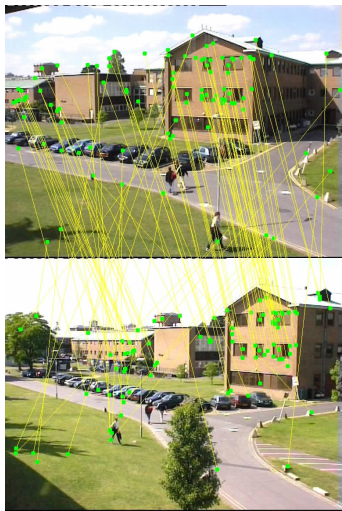


图 9 SIFT 描述子的直接匹配

5.2 多摄像机之间的匹配

图 7(b) 中有 2491 个 SIFT 特征, 那么 SIFT 描述子的种子个数至少是 $2491 \times 4 = 9964$, 每个种子有 8 个方向的向量信息, 那么 SIFT 描述子的数据总数是至少 $9964 \times 8 = 79712$, 对于视频中即使是准实时的处理其结果也是灾难性的. 所以我们可以得出: 直接使用一幅图像全部 SIFT 特征的 SIFT 描述子进行多摄像机之间的目标匹配是不合适的, 因此我们提出基于区域 SIFT 描述子的目标匹配算法,

可以大幅地缩减数据总数.

假设在 t 时刻, Camera i 观察到 M 个目标, Camera j 观察到 N 个目标, 我们的匹配算法基于如下定义.

定义 1. 从 Camera i 观察到的某个目标 p' 与从 Camera j 观察到的某个目标 q' 是匹配的, 当且仅当 Camera i 中 p' 对应的区域 p 与 Camera j 中 q' 对应的区域 q 是匹配的.

定义 2. Camera i 中目标区域 p 与 Camera j 中某个目标区域 q 的区域是匹配的, 当且仅当 p 的 SIFT 描述子与 q 的 SIFT 描述子相匹配的对数 (两个相匹配的 SIFT 描述子称为 1 对) 等于 $\max\{p$ 的 SIFT 描述子与 Camera j 中任意目标区域的 SIFT 描述子相匹配的对数 $\}$.

这样通过定义 1 和定义 2, 我们就把多摄像机之间目标匹配问题转换为底层的 SIFT 描述子之间相互匹配的问题, 即将一个较难解决的问题转换成已经解决的问题. 如图 10 所示, 本文的匹配算法的主要步骤表述如下:

1. 将属于 Camera i 的所有目标区域 Region 1~ M 中检测到的所有 SIFT 描述子加入链表 L_i ; 同样, 将属于 Camera j 的所有目标区域 Region 1~ N 中检测到的所有 SIFT 描述子加入链表 L_j . 初始化 Camera i 的任意一个目标区域与 Camera j 的任意一个目标区域之间的匹配度为 0.
2. 从 L_i 取出一个 SIFT 描述子 x , 如果可以在 L_j 中找到与之匹配的 SIFT 描述子 y , 则 x 所属的目标区域与 y 所属的目标区域之间的匹配度加 1.
3. 重复步 2 直至 L_i 中所有的描述子都被遍历到. 这样我们得到了 Camera i 的任意一个目标区域与 Camera j 的任意一个目标区域之间的匹配度.
4. 对 Camera i 的任意一个目标区域, 与之匹配的目标区域是 Camera j 中与之匹配度最大的一个, 如果最大的匹配度对应的 Camera j 中目标区域有多个, 取 SIFT 特征个数最接近的 Camera i 的任意一个目标区域的那一个. 如果 Camera i 的目标区域存在 2 个或 2 个以上匹配 1 个 Camera j 的目标区域, 保留匹配度最大的那一对区域为匹配结果.

同样以图 7(a) 和图 7(c) 的匹配为例, 如图 11 所示, 我们首先使用第 3 节中区域检测的方法进行目标区域的提取, 提取的结果如图 11(a) 所示, 各有 3 个目标区域. 然后按照步 2 和步 3, 供进行 $3 \times 3 = 9$ 区域之间的 SIFT 描述子匹配, 得到相应的匹配度. 最后通过步 4 得到匹配的结果. 因为匹配时可能图片大小不对称, 我们在图片周围都添加空白区域, 所以个别匹配点不在区域内是正常的. 从图 11 我们可以直观地看出比对的数据量得到了极大的压缩, 匹配结果也是正确的.

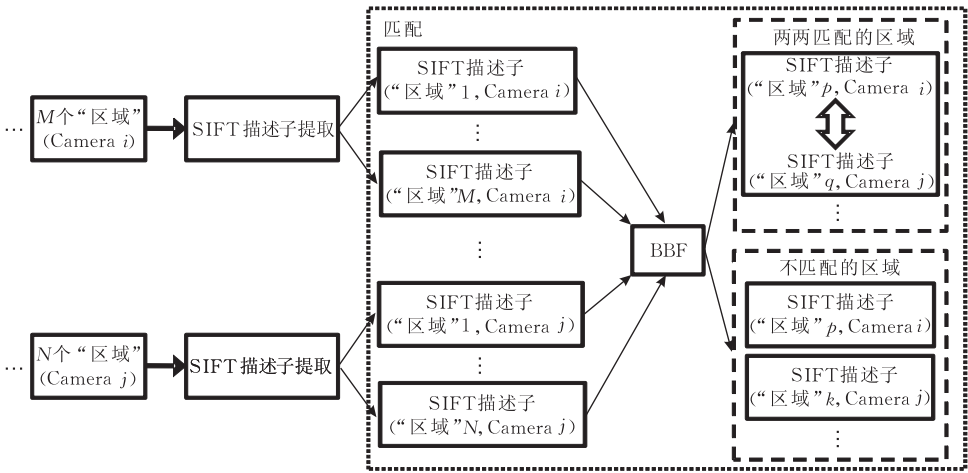


图 10 多摄像机两两之间的匹配

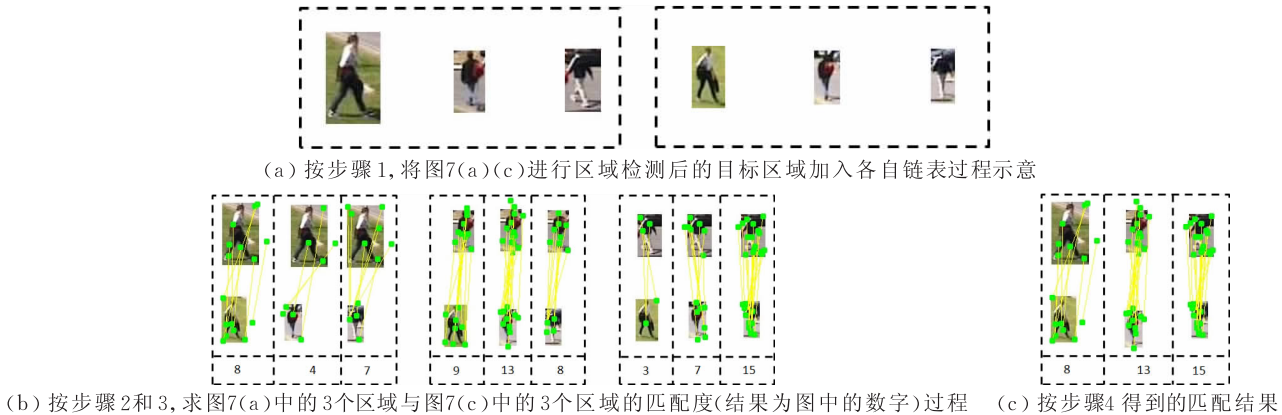


图 11 多摄像机两两之间的匹配

5.3 匹配后的融合

当匹配关系建立后,即一个摄像机中的 TO 可以对应到其它摄像机中的 TO,当跟踪的目标对象发生遮挡或与其它目标交错的时候,有可能发生跟踪失败的情况,这时可以利用其它摄像机中的 TO 来综合判断当前的这个 TO 是否正确.此外,当同一个目标在多个摄像机中建立了对应关系,如果这样的摄像机的个数大于 2 且处于不同的视角,我们就有可能根据对极几何和立体视觉的知识建立该目标的 3D 模型.

6 实验结果与分析

为了验证本文提出的方法,我们主要在开放的 PETS 系列监控视频库上进行测试,根据文献[1],很多算法都使用这个视频监控数据. PETS 系列中, PETS-ICVS 包含多个室内场景视频,但是没有多摄像机从不同视角拍摄的视频; PETS2000 仅仅是一个摄像机拍摄的停车场场景视频; PETS2002 是

室内拍摄的画面存在球面形变的监控视频,同样没有多摄像机从不同视角拍摄的视频; PETS2003 是从相差极大的不同视角拍摄的英国 Liverpool 足球队比赛视频,因为视角相差太大,远景拍摄人物很小,而且足球场上球员穿的衣服颜色样式完全相同,所以不适合进行匹配测试. 基于以上分析,在本文实验中我们主要在 PETS2001 数据库上进行匹配测试,它满足多摄像机对同一场景拍摄的要求,不同的摄像机视角明显不同但相差不是太大,比较贴近实际情况. 此外,我们以 SIFT 特征的个数作为计算复杂度的一个近似表示,对直接使用整图的 SIFT 描述子进行匹配和我们的基于区域 SIFT 描述子匹配进行了比较.

本文的多目标跟踪算法参考了 Okuma(<http://www.cs.ubc.ca/~okumak/research.html>)的 Matlab 代码,该代码实现了基于 boosted particle filter 算法的冰球比赛中的多目标跟踪. 本文的 SIFT 相关代码主要参照了 Lowe(<http://www.cs.ubc.ca/~lowe/keypoints/>)提供的代码,然而该代码的 GUI

有待改进,我们将它移植到 Visual C++. Net2003, 并进行了一些整合. 此外, 在 SIFT 描述子匹配中, BBF 算法搜索最近邻近的候选关键点的个数上限设为 100, 最近两次搜索的距离比率的平方设为 0.7.

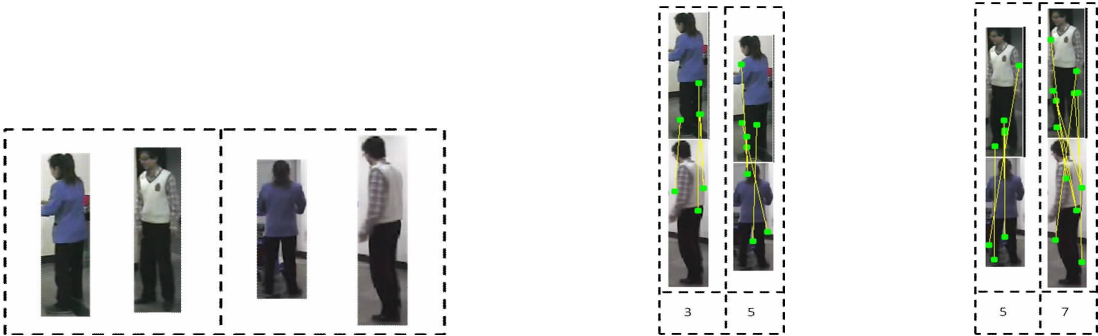
6.1 相同类型的目标物体

我们用两个 D-link DCS-5300 无线摄像机(带云台)从两个不同角度对某大楼室内大厅进行监控, 摄像机的分辨率为 352×288 (25 帧/s). 我们选用了某时段的一段视频来说明我们方法的效果. 因为室内监控一般被监控的目标与摄像机比较近, 再加上

无线摄像机的分辨率不高, 这类监控存在如下特点: (1) 在同一镜头中出现的目标比较少, 目标区域比较大; (2) 运动的目标类型单一, 一般都是人; (3) 为了扩大监控的范围, 摄像机往往带有云台. 我们给出这个序列中某帧的匹配结果, 如图 12(a)所示, Camera 1 视角有 2 个人, 而 Camera 2 视角有 2 个人. 图 12(b)显示了分别从 Camera 1 视角和 Camera 2 视角检测出的目标区域, 共 4 个. 图 12(c)显示了目标的匹配情况, 其中匹配度最高的分别是 7 和 8, 也是实际应该匹配的结果.



(a) 左边是 Camera 1 视角, 右边是 Camera 2 视角



(b) 左边是 Camera 1 的 2 个目标, 右边是 Camera 2 的 2 个目标 (c) 匹配结果, 上面是 Camera 1 的两个目标, 下面是 Camera 2 的两个目标

图 12 多摄像机之间不同类别物体的比较

6.2 不同类型的目标物体、亮度变化和遮挡

我们选用 PETS2001 数据库的 dataset 3 中 frame 2048 到 frame 3248 做不同类型目标物体的匹配测试, 画面中既有人, 又有自行车. 在这部分图像序列中先后有 8 个运动目标, 包括 6 个行人和 1 个骑自行车的车手. 这部分的视频序列中进行目标匹配至少存在以下困难:

(1) 不同类型的物体. 在分割出目标区域之后, 一般的多摄像机之间的目标匹配算法都要做一下物体类型的判别, 如文献[1]中所使用的颜色直方图的方法. 但是本文的方法中可以省略这一步骤, 因为我们不是基于形状特征或颜色特征, 而是本地纹理特征.

(2) 亮度变化. 这部分图像序列中, Camera 1 视角是背光的, 而 Camera 2 是逆光的, 光线变化比较

剧烈. 但是 SIFT 特征本身对光线的强度是鲁棒的, 本文的方法也受益于这种鲁棒性.

(3) 遮挡. 在这部分图像序列中, 一般的行人均存在遮挡现象. 文献[1]中的匹配算法需要单独区分个人、群体和遮挡三种情况. 本文的方法则省略了这一步, 将遮挡情况当作一般情况来处理, 只考虑不被遮挡的那部分区域.

我们给出这个序列中的某帧(第 3033 帧)的部分匹配结果, 这个例子中存在不同类型的物体、半数左右的遮挡、光线的明显变化, 镜头远近也不同, 是一个相对困难的匹配, 比较有代表性. 如图 13(a)所示, Camera 1 视角有 5 个人和 1 个自行车车手, 其中有 4 个行人存在不同程度的遮挡, 而 Camera 2 视角有 6 个人和 1 个自行车车手, 其中有 3 个行人存在不同程度的遮挡. Camera 1 的天空云彩部分清晰

可见,而 Camera 2 中则几乎曝光,说明光线变化剧烈.图 13(b)显示了分别从 frame 3033 的 Camera 1 视角和 Camera 2 视角检测出的目标区域,共 13 个.

图 13(c)显示了骑自行车的人与其他目标的匹配情况,其中匹配度最高的是 7,也是实际应该匹配的结果.

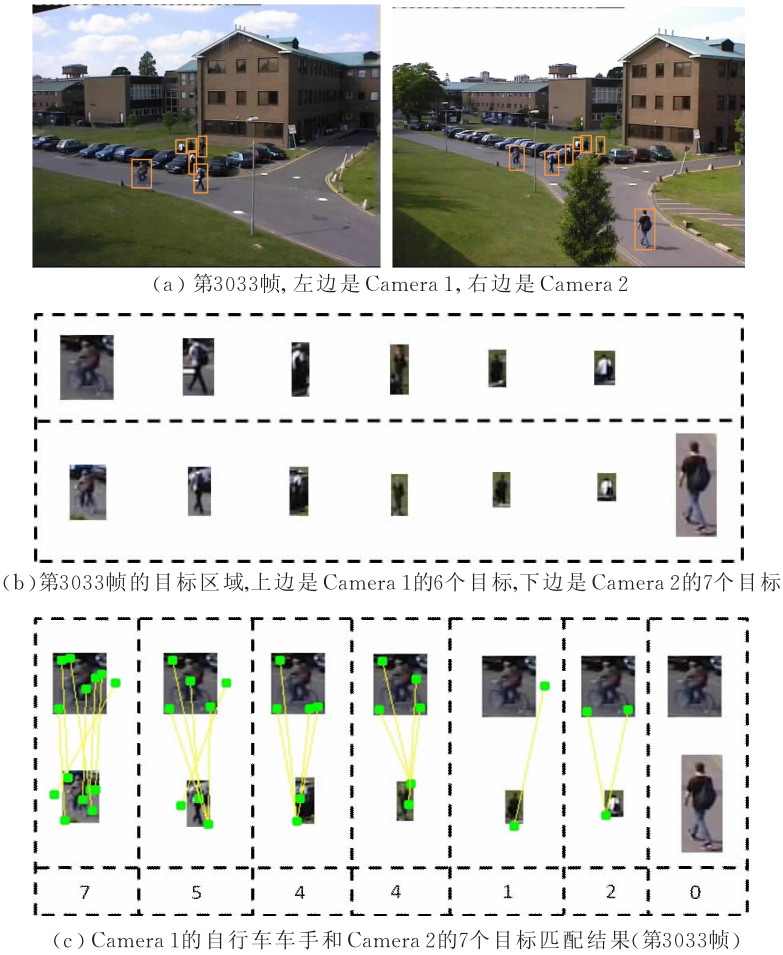


图 13 多摄像机之间不同类别物体的比较

6.3 比较

正如第 1 节中所讨论的, 现有的多摄像机匹配算法大致分为基于区域的或者基于点的, 当然还有非主流的集中式和分布式的分类. 基于区域的匹配方法大多是以颜色特征为表示的, 第 1 节中分析了颜色特征往往是不可靠的. 我们的方法虽然是基于区域的, 但是区域的表示却是基于本地纹理特征的 SIFT 描述子, 可以说我们的方法结合了基于区域和基于点这两类方法的优点. 但是本文的方法直接和其它的方法进行比较是比较困难的, 因为不同的方法对应于不同的假设, 如不同的几何约束和应用等.

然而我们可以和直接使用 SIFT 算子进行图像匹配的算法进行比较. SIFT 特征被广泛地用于宽基线匹配和全景图构建中. 我们以处理的 SIFT 特征的个数来粗略代表计算量, 以图 13 所举的例子为统计样本, 那么本文的方法与直接使用 SIFT 特征进

行全图匹配的比较结果如图 14 所示.

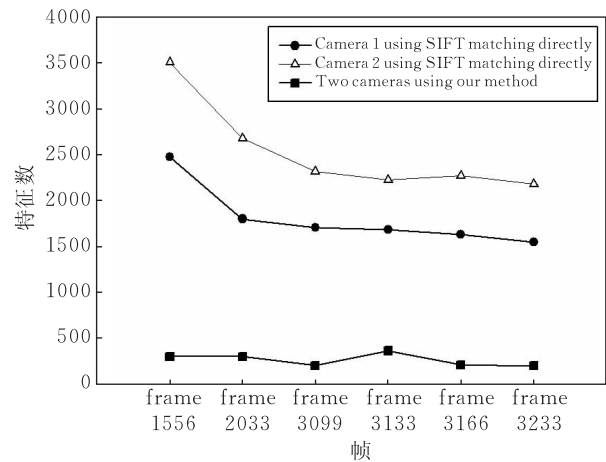


图 14 直接使用 SIFT 特征匹配与我们方法的比较

从第 5.2 节的分析我们知道直接使用 SIFT 特征进行匹配时, SIFT 描述子的数据量是巨大的, 需

要降维. 从图 14 可以看出, 我们的方法处理两个摄像机的数据也远小于直接匹配时一个摄像机需要处理的数据.

7 总结与将来的工作

本文提出了一种基于区域 SIFT 描述子的目标匹配方法把多摄像机之间目标匹配问题转换成为底层的 SIFT 描述子之间相互匹配的问题, 即将一个较难解决的问题转换成已经解决的问题. 本文的方法是基于区域的方法, 但是区域的特征以 SIFT 描述子而不是通常的颜色特征表示, 同时多目标的检测使用减背景技术, 多目标跟踪则选用基于序贯蒙特卡罗方法的粒子滤波. 本文的匹配方法不需要摄像机之间的合作, 也不要求目标物体处于同一地平面, 无几何约束的需求, 甚至同一目标物体的背景完全切换后, 也可以进行匹配, 而且可以匹配各种类型的目标物体, 适合分布式计算, 但也可以集中式处理, 并且容忍亮度的变化. 在 PETS 监控数据库上进行的效果测试显示我们的方法是有效的, 而且计算复杂度小.

本文的方法虽然无几何约束的要求, 然而随着视角差别的不断增大, 匹配的正确率是不断下降的或者说是不可靠的. 另外匹配的正确与否很大程度上将取决于目标区域分割的精度, 而且本文的方法在分布式模式下需要智能摄像机的硬件支撑.

致 谢 本文共享了 Okuma K 的部分多目标跟踪 Matlab 代码和 Lowe David G 的部分 SIFT 相关的 Matlab 代码, 在此表示感谢!

参 考 文 献

- [1] HU Wei-Ming et al. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(4): 663-671
- [2] Chang T H, Gong S, Ong E J. Tracking multiple people under occlusion using multiple cameras//*Proceedings of the British Machine Vision Conference*. Bristol, UK, 2000: 566-575
- [3] Krumm J et al. Multi-camera multi-person tracking for easy living//*Proceedings of the IEEE International Workshop Visual Surveillance*. Dublin, Ireland, 2000: 3-10
- [4] Mittal A, Davis L S. M2Tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo//*Proceedings of the European Conference on Computer Vision*. Copenhagen, Denmark, 2002: 18-36
- [5] Bollobas B, Thomason A G. *Random Graphics of Small Order*. North-Holland, Amsterdam, 1985: 47-97
- [6] Avidan S, Moses Y, Moses Y. Centralized and distributed multi-view correspondence. *International Journal on Computer Vision*, 2007, 71(1): 49-69
- [7] Tsutsui H, Miura J, Shirai Y. Optical flow-based person tracking by multiple cameras//*Proceedings of the IEEE Conference on Multisensor Fusion and Integration in Intelligent Systems*. Baden-Baden, Germany, 2001: 91-96
- [8] Orwell J, Remagnino P, Jones G A. Multiple camera color tracking//*Proceedings of the IEEE International Workshop Visual Surveillance*. Fort Collins, USA, 1999: 14-24
- [9] Utsumi A, Mori H, Ohya J, Yachida M. Multiple human tracking using multiple cameras//*Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. Nara, Japan, 1998: 498-503
- [10] Kelly P et al. An architecture for multiple perspective interactive video//*Proceedings of the ACM Multimedia*. San Francisco, USA, 1995: 201-212
- [11] Cai Q, Aggarwal J K. Tracking human motion in structured environments using a distributed-camera system. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 1999, 21(11): 1241-1247
- [12] Khan S, Javed O, Shah M. Tracking in uncalibrated cameras with overlapping field of view//*Proceedings of the IEEE International Workshop Performance, Evaluation of Tracking and Surveillance*. Kauai, USA, 2001: 84-91
- [13] Black J, Ellis T. Multi-camera image tracking//*Proceedings of the IEEE International Workshop Performance Evaluation of Tracking and Surveillance*. Kauai, USA, 2001: 68-75
- [14] Zhou S K, Chellappa R, Moghaddan B. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 2004, 13(11): 1491-1506
- [15] Hue C, Cader J, Perez P. Tracking multiple objects with particle filtering. *IEEE Transactions on Aerospace and Electronic Systems*, 2002, 38(3): 791-812
- [16] Lowe David G. Object recognition from local scale-invariant features//*Proceedings of the International Conference on Computer Vision*. Corfu, Greece, 1999: 1150-1157
- [17] Brown M, Lowe D. Recognizing panoramas//*Proceedings of the International Conference on Computer Vision*. Nice, France, 2003: 1218-1225
- [18] Krystian Mikolajczyk, Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(10): 1615-1630
- [19] Lowe David G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2): 91-110
- [20] Huo Zi-Qiang, Han Chong-Zhao. A survey of visual tracking. *Acta Automatica Sinica*, 2006, 32(4): 603-617(in Chinese)

(侯自强, 韩崇昭. 视觉跟踪技术综述. 自动化学报, 2006, 32(4): 603-617)

[21] Dai Ke-Xue, Li Guo-Hui, Tu Dan, Yuan Jian. Prospects and current studies on background subtraction for moving objects detection from surveillance video. Journal of Image and Graphics, 2006, 11(7): 919-927(in Chinese)

(代科学, 李国辉, 涂丹, 袁见. 监控视频运动目标检测减背景技术的研究现状和展望. 中国图象图形学报, 2006, 11(7): 919-927)



MING An-Long, born in 1979, Ph.D. candidate. His main research interests include multimedia systems, computer vision and grid computing.

MA Hua-Dong, born in 1964, professor, Ph. D. supervisor. His research interests include multimedia systems and networking, grid computing, sensor networks and formal method.

Background

This work is supported by the National High Technology Research and Development Program of China under grant No. 2006AA01Z304; the National Natural Science Foundation of China under Grant No. 90612013; the Specialized Research Fund for the Doctoral Program of Higher Education under grant No. 20050013010; the NCET of MOE, China.

Visual surveillance using multiple cameras has attracted much attention in the computer vision community in recent years. This is because by utilizing multiple cameras, the area of surveillance is expanded and information from multiple views is extremely helpful to handle many issues such as occlusion. However, visual surveillance using multiple cameras also brings a number of problems such as camera installation, calibration of multiple cameras, correspondence between multiple cameras, automated camera switching, and data fusion. Correspondence between multiple cameras involves finding correspondences at the same time between objects in the different image sequences. Only after correspondence between multiple cameras is well constructed can the information from multiple cameras be fused. Therefore, it is one of the most important and basic problems in visual surveillance using multiple cameras. Although correspondence of multiple

cameras is a newly emergent research topic, some attempts have been made to investigate this problem. The existing methods for establishing correspondences can be classified according to the types of employed features, whether the cameras are calibrated or not, and whether the correspondences are region-based or point-based. In this paper, we propose a region-SIFT descriptor based target matching method for multiple cameras. This is a region based method. However, the region is represented by SIFT descriptor instead of traditional color features. In our method, the background subtraction is used in multi-target detecting, and the particle filter is utilized in multi-target tracking. Non camera calibrations are required in our method, neither the constraint that all objects stand in the same plane. The main features of our method are highlighted as follows: (1) Non geometry constraints are required to some extent. (2) Many types of objects are supported. (3) The cameras can simply move during object tracking. (4) It is more fit to distributed computing, but the traditional centralized computing is also supported. (5) It is robust to changes of the light intensity. Experimental results show that our method is effective.