基于深度学习和时空约束的跨摄像头行人跟踪*

夏 天 李旻先 邵晴薇 管 超 陆建峰
（南京理工大学计算机科学与工程学院 南京 210094）

摘 要 目前，视频监控的布设十分广泛，如何从多个视频监控的数据中有效获取行人的轨迹信息，对于社会安防体系具有非常重要的价值。因此，跨摄像头行人跟踪已成为计算机视觉领域的一个重要研究内容。论文设计了一个基于深度学习的跨摄像头行人跟踪的方法，将跨摄像头行人跟踪任务划分为行人检测和行人检索两部分。在行人检测部分使用Faster R-CNN方法，在行人检索部分使用CNN特征来计算相似度距离并通过时间与空间关系对检索结果进行约束与优化，并在复杂的监控视频下进行了实验。

关键词 深度学习；行人检测；行人检索；Faster R-CNN；CNN；时空约束
中图分类号 TP301 **DOI**：10. 3969/j. issn. 1672-9722. 2017. 11. 039

Pedestrian Tracking Across Cameras Based on Deep Learning and Spatiotemporal Constraint

XIA Tian LI Minxian SHAO Qingwei GUAN Chao LU Jianfeng
（School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094）

Abstract Nowadays surveillance video is everywhere, how to track pedestrian across the multi-camera is valuable for social security system. So pedestrian tracking across cameras becomes an important research in computer vision. This paper proposes a method of pedestrian tracking across the multi-camera. The method divides the task into pedestrian detection and person re-identification. In the part of pedestrian detection, Faster R-CNN is adopted. In the part of person re-identification, CNN is used features to calculate the similarity score and optimize the results by the spatiotemporal constraint. Meanwhile, experiments in a complex surveillance video is applied to test the proposed method performance.

Key Words deep learning, pedestrian detection, person re-identification, Faster R-CNN, CNN, spatiotemporal constraint
Class Number TP301

1 引言

近些年来，监控视频大量普及，随之而产生的是大量的监控视频数据，由于单个摄像头的感知范围有限，为了监控特定的目标，通常需要查看多个不同位置的视频监控信息，在这一系列的监控视频中寻找特定的行人目标需要耗费大量的人力物力，因此，基于跨摄像头的行人自动跟踪已经成为了视频分析工作中亟待解决的重要课题。跨摄像头行人跟踪任务的目的是在多个不同的摄像头中找到某个特定的行人，由于每个摄像头的场景不同，光照不同，行人姿态可能也会发生变化，获取的行人图像质量往往较低，所以通过计算机来判断在不同摄像头下的目标是否是同一个行人，是一个挑战性的课题。

在智能监控研究中，跨摄像头行人跟踪一般分为行人检测和行人检索两个过程。行人检测在目前计算机视觉领域中较为成熟，比较经典的方法有HOG+SVM[7]，DPM[12]以及现在流行的深度学习方法，例如Faster R-CNN[1]，YOLO[13]，SSD[14]。而行人检索方面，虽然也有很多研究者提出过很多方法，但是目前方法的性能还没能达到行人检测的水平，

* **收稿日期**：2017年5月9日，**修回日期**：2017年6月27日
作者简介：夏天，男，硕士研究生，研究方向：人工智能与模式识别。李旻先，男，讲师，研究方向：人工智能与模式识别。邵晴薇，男，硕士研究生，研究方向：人工智能与模式识别。管超，男，硕士研究生，研究方向：人工智能与模式识别。陆建峰，男，教授，研究方向：人工智能与模式识别。

目前常用的方法一般采用是提取图像的特征向量来计算向量距离进行匹配,经典的特征有RGB、HOG[7]、SIFT[8-9]、ColorName[10]等以及深度学习中的Convolutional neural network(CNN)特征。

本文设计了一个跨摄像头行人跟踪的算法,在行人检测部分,本文采用了近两年性能比较优异的Faster R-CNN[1]方法。行人检索部分,本文使用深度学习的神经网络提取行人图像的特征,再计算相似度进行排序筛选得到初步检索结果,然后通过检索结果的行人出现的时间与所在的空间关系判断是否能够形成一条轨迹,最后对形成的轨迹进行优化得到跟踪结果。文本创新点主要在于将单摄像头行人跟踪引入检索系统、对CNN特征的优化和通过时空关系来优化行人检索结果。

2 行人检测

行人检测是跨摄像头行人跟踪的第一步。行人检测的结果作为行人检索的检索库,如果在检测阶段对查询目标漏检了,那么在行人检索时查询目标也必然匹配不到,所以可以说行人检测是行人匹配的基础,行人检测的性能直接影响到跨摄像头行人跟踪的性能。

Faster R-CNN[1]是以R-CNN[2],SPP-NET[3],Fast R-CNN[4]为基础改进而来的基于候选区域的深度学习目标检测方法,是目前目标检测领域中性能较好,效率最高的目标检测算法,在PASCAL VOC2007和2012上mAP值位居榜首(在VOC2007上的mAP值是73.2%,在VOC2012上的mAP值是70.4%)[1]。在效率上,在K40 GPU上ZF模型的一张图像的平均检测时间大约为0.06s,VGG16模型的一张图像的平均检测时间大约为0.2s。Faster R-CNN是由Fast R-CNN改进而来,使用深度网络来提取候选区域(Region Propose Network,RPN)[1]来代替Fast R-CNN中的Selective Search(SS)[15],可以理解为RPN+Fast R-CNN。其算法流程图1所示。

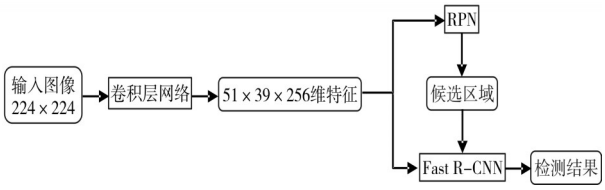

图1 Faster R-CNN算法流程图

本实验挑选了4个跟踪场景的监控视频中的部分片段并把其中的行人作为训练样本,训练了一个专门针对于该4个场景的Faster R-CNN行人检测模型,该模型检测结果较为理想,图2给出了Faster R-CNN使用该模型在全国研究生智慧城市视频分析技术挑战赛跨摄像头多类目标检测数据库上的部分行人检测结果。从图2中可以看到,在摄像头中近处以及中远距离的几乎不会出现行人的漏检与误检,在远处道路尽头由于行人目标十分微小,检测结果会出现漏检。

图2 Faster R-CNN部分行人检测结果

3 行人检索

通过行人检测得到一组庞大的行人池,本节的任务是在这个庞大的行人池中,检索出查询目标。本文匹配时采用的特征是AlexNet[5]中的pool5层特征。本实验挑选了视频中10个行人来进行测试,查询对象为10个行人在某个视频的某一帧处的位置。算法具体流程图如图3所示。首选训练了一个AlexNet的分类网络,将行人检测得到的行人池按照查询目标进行分类,缩小每个对象所对应的行人池大小。对行人池中的每个行人提取AlexNet中的pool5层特征。由于查询对象为一张单一的图片,所以并不能反应整个人的完整信息,本文通过对查询对象进行单摄像头行人跟踪得到该查询对象的轨迹序列,将单样本检索问题转化为多样本检索为题,然后再对序列中的每一个图像提取AlexNet的pool5层特征。在得到查询对象序列和行人池的pool5层特征之后,计算行人池中的每个行人与查询对象序列的相似度距离并进行排序,排序之后选择一个阈值对行人池中的行人进行过滤筛选。最后根据筛选得到的行人序列根据帧号进行排序,通过时间与行人框的位置关系,进行精修与修正得到最后结果。

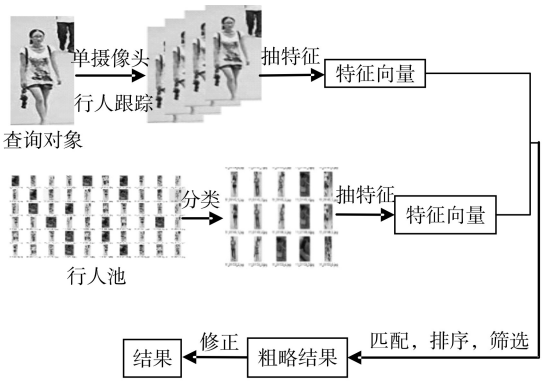

图 3　Faster R-CNN 行人匹配算法流程图

3.1　单摄像头行人跟踪

由于监控视频中远处行人较为模糊、行人在行走过程中经常出现姿态变化以及不同摄像头的光照不同导致的同一行人在不同摄像头下的差异等等原因，基于单样本的行人检索在监控视频中精度较低。为了提高精度，本文通过单摄像头行人跟踪算法 Kernelized Correlation Filters(KCF)[6]对查询目标进行跟踪，得到查询目标的轨迹序列，将单样本检索任务转换为多样本检索任务。

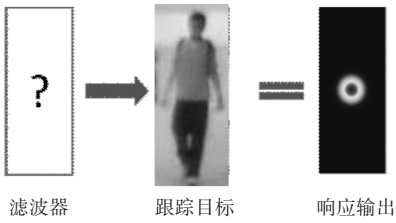

图 4　KCF 算法原理

KCF 方法是通过该相关滤波器来对目标位置进行预测更新来对目标进行追踪。其算法基本原理是通过相关滤波器对上一帧中目标周围的位置进行滤波，计算出一个响应输出，根据响应输出判断目标在下一帧中的位置，如图 4，该响应输出越高，则与我们要找的目标越相似。因此响应输出最高的框就是下一帧中目标的位置。通过该目标的新位置去更新滤波器的参数，更新后的滤波器用来预测再下一帧的目标位置。

通过单摄像头行人跟踪得到查询对象第一次在视频中出现的完整轨迹，即一个查询对象的行人图像序列，该序列一般包含了对象从远道近的图像以及在行走过程中的一些姿态变化还有在视频边缘消失时的一些半身图像，这些图像对跨摄像头行人检索来说，信息是较为完整的。

3.2　AlexNet 分类与 pool5 层特征提取

3.2.1　AlexNet 分类模型

将第二章的行人检测结果作为检索库，使用 3.1 章节中的行人追踪结果作为查询对象进行行人检索。但是由于第二章中的行人检测结果数量巨大，使得检索库过于庞大，本实验中查询对象为 10 个行人，而检索池中包含了几千个不同的行人。过于庞大的行人池导致了两个问题：第一，冗余信息太多，庞大的行人池中 90%以上的行人不是查询目标；第二，对行人池中的每个人都要进行特征提取并计算它与每个查询对象的距离，计算量太大。为了解决第二个问题，首先对庞大的行人池进行分类划分。实验中使用了 CNN 特征在图像分类上的经典网络 AlexNet[5]对行人池进行分类，其结构如图 5。将 10 个查询对象的轨迹序列作为训练样本，训练一个 AlexNet[5]的分类模型。该分类模型可将庞大的行人池进行分类，把行人池中的每个行人都分到它与之最像的一个查询对象中的行人池中去。将分类后的行人池中的某一类作为其对应查询对象的检索池，与初始行人池相比分类后行人池大约缩小了 10 倍。

3.2.2　AlexNet 的 pool5 特征提取

从图 5 中可以看出，pool5 层特征是一个 6×6×256 的三维向量，可以看做是 256 个 6×6 的矩阵，对于每个 6×6 的矩阵，计算它的均值与最大值，得到了 256 个均值与 256 个最大值，再将其串联起来，得到一个 512 维的特征向量，前 256 维为 256 个 6×6 矩阵的均值，后 256 维为 256 个 6×6 矩阵的最大值，如图 6。

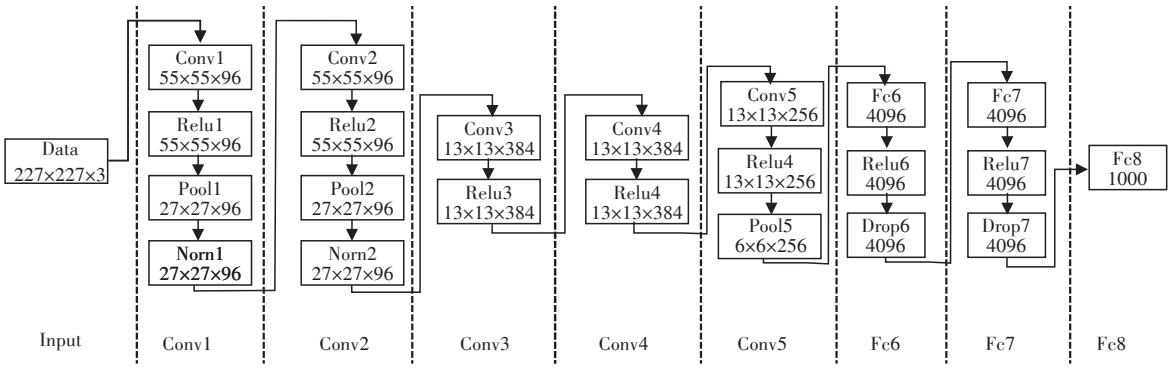

图 5　AlexNet 结构图

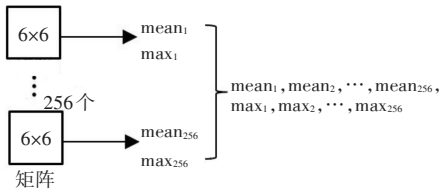

图6 Alexnet Pool5层特征提取示意图

3.3 相似度计算

提取行人池中每个行人与查询目标序列中每个对象的特征向量之后,对每个查询对象,计算该查询对象序列与分类后行人池中的每个行人的特征向量距离。本实验采用了余弦距离。余弦距离公式如下:

$$\cos(X,Y)=\frac{x_1y_1+x_2y_2+\ldots+x_{512}y_{512}}{\sqrt{x_1^2+x_2^2+\cdots+x_{512}^2}\cdot\sqrt{y_1^2+y^2+\cdots+y_{512}^2}} \tag{1}$$

由于我们的查询对象是一个序列,其中包含了查询对象的许多张图片,所以在计算相似度时需要做一个简单的处理,公式如下:

$$\text{distance}(p,Q)=\max_{q\in Q}\cos(p,q) \tag{2}$$

其中 p 为行人池的行人的512维特征向量,Q 为查询对象特征的集合。我们要在查询对象特征的集合 Q 中找到一个与 p 相似度最高的 q ,$\cos(p,q)$ 就是 p 与 Q 的相似度。

3.4 通过时空约束优化结果

3.4.1 基于时间关系的筛选

对行人池中的行人与查询对象的相似度进行排序,排名越靠前的行人认为与目标对象越相似。由于目标行人在视频中出现是集中在一段时间内且行人池中是目标行人的行人排名都很靠前,所以在目标行人出现的时间段内,正确的查询目标的相似度排名都普遍很低。在本节中我们通过时间轴上的局部行人平均相似度来反应该现象。

首先使用行人池中行人在视频中出现的时间信息对分类后的行人池进行一次过滤。由于一个行人在某个视频的某一帧不可能重复出现,所以将行人池划分成许多小块,属于同一帧的行人分为一个小块,在块内,取相似度排名最高的,即在某一帧图像中,只选一个与目标最像的行人作为候选人。通过过滤使得分类后的行人池再一次缩小,并且一些与查询对象同时出现的行人被过滤掉了。接着对按照时间顺序排列的行人相似度排名,做如下操作:

$$Rank(t)=mean(r_t,r_{t+1},\cdots,r_{t+n}) \tag{3}$$

其中 $Rank(t)$ 为行人池中第 t 帧到第 $t+n$ 帧的行人的平均相似度排名。r_t 为行人池中第 t 帧的行人的相似度排名。

平均相似度排名 $Rank(t)$ 反应了从第 t 帧开始是否大量出现了与目标相似的对象。当相似度平均排名低于某个阈值 m 时,说明在第 t 帧到第 $t+n$ 帧之间有与目标相似的人出现并形成了轨迹。将排名绘制成一条曲线,如图7,相似度排名在 m 以下的我们认为是该行人出现了,图7中红线代表阈值 m 的高度。最后将平均相似度排名低于阈值 m 的时间段作为查询对象在视频中出现的时间段,在过滤后的行人池中将改时间段的行人截取,作为初步检索结果,结果如图8。

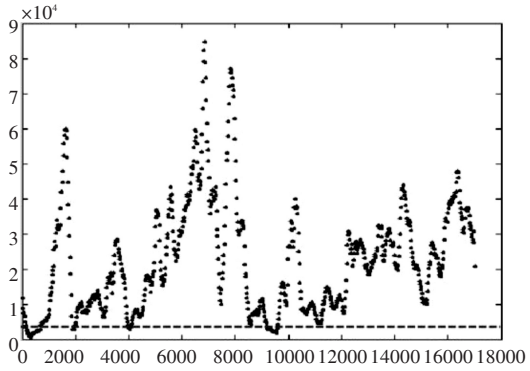

图7 行人池局部平均排名分布图

3.4.2 基于位置关系的修正

图8显示了检索结果中的一部分,从图8中可以看到绿色框为错误的检索结果,本节将通过矩形框的位置信息来对前两种原因导致的错误检索结果进行修正。

图8 通过局部平均相似度排名截取的行人检索结果

由于视频中前后帧的连贯性,相邻帧之间查询对象的正确矩形框可以看做是由一个初始矩形框滑动而来的,前后帧之间的矩形框应该是有重叠面积的。因此对3.4.1节的结果中的每一帧的行人框计算其与前 s 帧有重叠的矩形框数,记为 n_1 ,与后 s 帧有重叠的矩形框数记为 n_2 ,当 $n_1/s>0.5$ 且 $n_2/s>0.5$ 时,认为该帧的矩形框正确,否则认为该帧的矩形框错误。对于有问题的矩形框,取出在行人检测部分得到的行人池中属于该帧的所有行人,并依次计算每个行人与前后 s 帧的检索结果的矩形框重叠数量 n_1 与 n_2 ,将 n_1+n_2 最大的行人作为

该帧的正确结果。

4　实验结果与分析

4.1　数据准备

本文实验数据库来自于全国研究生智慧城市视频分析技术挑战赛。数据库中的视频均来自于北京大学的校园监控视频，该监控视频真实且较为复杂。本文从2016年智慧城市的跨摄像头行人跟踪比赛的数据中挑选了其中4个摄像头进行行人检索，部分摄像头截图如图9。行人检测的训练数据是从2015年智慧城市的单摄像头多类目标检测数据中的4个摄像头视频，与行人检索的4个摄像头是相同场景。行人检测训练数据的每个视频包含800张图像，共计3200张，包含12664个行人。行人检索每个摄像头9000到10000张图像不等，共计37934张，从中挑选了10个行人进行检索，10个行人的查询对象都为其中一个摄像头出现的第一帧位置。

图9　智慧城市的跨摄像头行人跟踪比赛的部分视频截图

4.2　评测标准

本文实验评测指标采用了智慧城市比赛的跨摄像头行人跟踪比赛中的评测指标。对所有 N 个摄像头中的每个摄像头，在时间轴上设置 $T_n(1,2,\cdots,N)$ 个匹配节点，每个匹配节点在时间轴上具有一定的范围（例如前后10帧）。对于某一个指定跟踪对象，在某个包含的匹配节点范围内，当提交的跟踪轨迹和目标真实轨迹有超过50%的帧都能匹配上时，则认定该跟踪对象在跨头跟踪时匹配到了该匹配节点。

对于第 i 个跟踪对象，评测指标为

$$\text{Recall}(i)=\frac{\sum_{n=1}^{N}C_n^i}{\sum_{n=1}^{N}G_n^i};\ \text{Precision}(i)=\frac{\sum_{n=1}^{N}C_n^i}{\sum_{n=1}^{N}D_n^i}$$

$$\text{Fscore}(i)=\frac{2\times\text{Recall}(i)\times\text{Precision}(i)}{\text{Recall}(i)+\text{Precision}(i)} \quad (4)$$

其中 N 是总的摄像头数，$G_n^i(G_n^i<T_n^i)$ 是第 i 个指定跟踪对象的跟踪Ground Truth轨迹在摄像头 n 里包含的匹配节点的数目。C_n^i 是指第 i 个指定跟踪对象在摄像头 n 里匹配到的匹配节点的数目。D_n^i 是第 i 个指定跟踪对象在摄像头 n 里包含的匹配节点的数目。

4.3　实验结果分析

本文在实验中使用CVPR2015的LOMO+XQDA[11]的行人检索方法在同一个数据集上进行了测试，并与本文的方法进行了对比。LOMO指的是Local Maximal Ocurrence，即局部最大出现次数，XQDA指的是Cross-view Quadratic Discirminant Analysis，即跨摄像头二次方程判别式。该方法通过LOMO特征与XQDA空间度量学习方法来训练并计算特征向量距离来进行行人匹配。图10给出了其中一个行人的跨摄像头跟踪可视化结果。

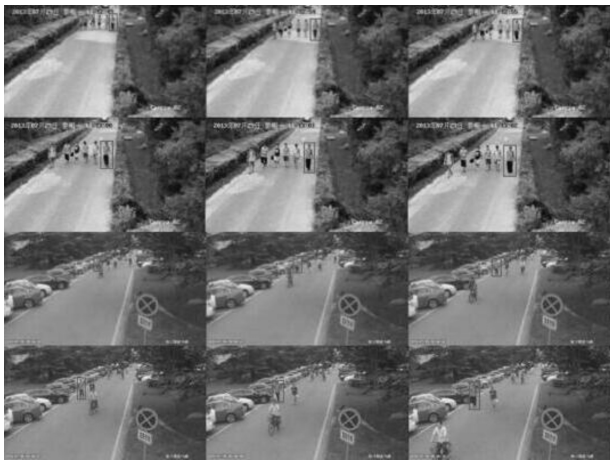

图10　跨摄像头行人检索部分可视化结果

实验对比结果如表1所示，表中单元格内3个参数依次为Recall、Precision、Fscore。从实验结果对比中可以看出，LOMO+XQDA方法在召回率上要比本文方法要高，但是在准确率上较低。本文方法在召回率上较低的原因是因为在行人匹配时虽然找到了查询对象出现的大致时间段，但是对于轨迹的开始与结束的定位不够精准。其次由于监控视频中视频中人流混杂，经常出现遮挡，行人检测漏检也导致召回率低。最后综合比较，本文方法的Fscore比LOMO+XQDA略高一筹。

5　结语

表1 本文方法与LOMO+XQDA方法的实验结果对比

	行人1(R/P/F)	行人2(R/P/F)	行人3(R/P/F)
本文方法	0.882/0.565/0.689	0.622/0.455/0.526	0.261/0.167/0.204
LOMO+XQDA	0.870/0.401/0.549	0.943/0.475/0.632	0.159/0.079/0.106
	行人4(R/P/F)	行人5(R/P/F)	行人6(R/P/F)
本文方法	0.335/0.203/0.253	0.186/0.145/0.163	0.914/0.343/0.499
LOMO+XQDA	0.537/0.213/0.305	0.336/0.138/0.196	0.951/0.191/0.318
	行人7(R/P/F)	行人8(R/P/F)	行人9(R/P/F)
本文方法	0.541/0.216/0.309	0.303/0.132/0.184	0.479/0.289/0.360
LOMO+XQDA	0.634/0.175/0.274	0.545/0.113/0.187	0.695/0.206/0.318
	行人10(R/P/F)	总计(R/P/F)	
本文方法	0.747/0.318/0.446	0.499/0.293/0.369	
LOMO+XQDA	0.834/0.212/0.338	0.617/0.226/0.331	

本文提出了一种基于深度学习的跨摄像头行人检索方法,并加入时间与空间信息来提高检索准确率。在2016智慧城市跨摄像头行人跟踪数据集上进行了测试,并与LOMO+XQDA方法进行了对比。最后分析了实验结果并提出了不足与缺点,在接下来的工作中将对不足与缺点进行改进并进一步完善这个方法。

参考文献

[1] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence,2015:1-1.

[2] Ross Girshick,Jeff Donahue,Trevor Darrell, Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [J].CVPR, 2014:580-587.

[3] K He, X Zhang, S Ren, J Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,37(9):1904-1916.

[4] Ross Girshick. Fast R-CNN[J]. CVPR,2015:1440-1448.

[5] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS,2012,25(2):2012.

[6] JF Henriques, C Rui, P Martins, J Batista. High-Speed Tracking with Kernelized Correlation Filters [C]//IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(3):583-596.

[7] Navneet Dalal, Bill Triggs. Hisograms of Oriented Gradients for Human Detection[J]. CVPR,2005,1:886-893.

[8] D.G.Lowe. Distinctive image feature from scale invariant keypoints[C]// IJCV,2004,60(2):91-110.

[9] J.Philbin, O.Chum, M.Isard, A.Zisserman. Object retrieval with large vocabularies and fast spatial matching[J]. CVPR,2007:1-8.

[10] F.Shahbaz Khan, R.M.Anwer, J.van de Weijer, A.D. Bagdanov, M.Vanrell, A.M.Lopez. Color attributes for object detection[J]. CVPR, 2012:3306-3313.

[11] S Liao, Y Hu, X Zhu, SZ Li. Person Re-identification by Local Maximal Ocurrence Representation and Metirc Learing[J]. CVPR, 2015:2197-2206

[12] Felzenszwalb, Girshick, McAllester, Ramanan. Object detection with discriminatively trained part-based models [C]//IEEE Trans.PAMI,2010,32(9):1627-1645.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi.You Only Look Once:Unified, Real-Time Object Detection[J].CVPR,2016:779-788.

[14] Wei Liu, Dragomir Anguelov. SSD:Single Shot MultiBox Detector[C]// ECCV,2016:535-549.

[15] J.R.Uijlings, K.E.van de Sande, T.Gevers, A.W.Smeulders. Selective Search for object recognition[C]//IJCV, 2013,104(2):154-171.