第53卷第5期
2021年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 53 No. 5
Oct. 2021

DOI:10.16356/j.1005-2615.2021.05.016

基于步行周期聚类的视频行人重识别关键帧提取算法

李梦静，吉根林，赵　斌

（南京师范大学计算机与电子信息学院/人工智能学院，南京 210023）

摘要：视频行人重识别旨在不同摄像头拍摄的视频中检索特定行人。但是，它面临着数据量庞大和视频数据存在时间冗余的问题，即视频数据耗费大量的存储空间且不同帧之间存在极强的相关性。因此，使用所有的帧进行识别会带来查询效率的下降，而且视频中大量的干扰和噪声也会给准确率带来不利影响。本文提出了基于步行周期聚类的视频行人重识别关键帧提取算法，首先利用行人步行时双脚距离变化的周期性规律提取候选步行周期，然后利用聚类的方法从候选步行周期中选出关键步行周期作为关键帧。最后，将该算法应用在视频行人重识别中，仅使用关键帧的信息进行识别以减少时间冗余的影响，从而提高准确率，并且在查询前对视频进行处理，减少视频数据量以提高查询效率。在视频行人重识别数据集MARS和DukeMTMC-VideoReID上的实验表明，本文算法能够减少59%～82%的视频数据量，并且累积匹配曲线Rank-1提高了1.1%～1.4%，平均精度均值提高了0.2%～5%。

关键词：视频行人重识别；关键帧提取；步行周期；聚类；视频分析

中图分类号：TP37　　**文献标志码：**A　　**文章编号：**1005-2615(2021)05-0780-09

Key Frame Extraction Algorithm for Video-Based Person Re-identification Based on Walking Cycle Clustering

LI Mengjing, *JI Genlin*, *ZHAO Bin*

(School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, Nanjing 210023, China)

Abstract: Video-based person re-identification aims to retrieve specific pedestrians from videos taken by different cameras. However, it faces the problem of huge data volume and time redundancy of video data, that is, video data consume a lot of storage space and have strong correlation between different frames. Using all frames for identification will reduce the query efficiency, and the interference and noise in the video will also adversely affect the accuracy. In order to solve such problems, this paper proposes a key frame extraction algorithm for video-based person re-identification based on walking cycle clustering. Firstly, the algorithm extracts the candidate walking cycle by using the periodicity of the distance between feet of pedestrians. Then, it selects the key walking cycle as the key frame from the candidate walking cycle by clustering. Interference and noise are removed and only key frame information is used for identification to reduce the impact of time redundancy and improve accuracy. Finally, the algorithm is applied to video-based person re-identification, and the data will be processed before querying to reduce the storage space and to improve the query efficiency. Experimental results on MARS and DukeMTMC-VideoReID datasets show that the algorithm can reduce storages space by 59%—82%, the cumulative match characteristic Rank-1 is improved

基金项目：国家自然科学基金(41971343)资助项目。

收稿日期：2020-10-11；**修订日期：**2021-01-09

通信作者：赵斌，男，博士，副教授，E-mail：zhaobin@njnu.edu.cn。

引用格式：李梦静，吉根林，赵斌. 基于步行周期聚类的视频行人重识别关键帧提取算法[J]. 南京航空航天大学学报，2021，53(5)：780-788. LI Mengjing, JI Genlin, ZHAO Bin. Key frame extraction algorithm for video-based person re-identification based on walking cycle clustering[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(5): 780-788.

by 1.1%—1.4% and the mean average precision is improved by 0.2%—5%.

Key words: video-based person re-identification; key frame extraction; walking cycle; clustering; video analysis

视频行人重识别是指在不同摄像头拍摄的行人视频中判断是否存在特定行人的技术，用来解决不重叠监控视野中行人身份匹配的问题[1]。这种针对特定人的视频检索具有重要的研究意义，在失踪者定位、犯罪跟踪和智能安防等方面有着广泛的应用[2]。随着近年来视频监控范围不断扩大、监控点数量增多，数据量持续猛增，给存储及使用带来了巨大限制[3]。据报道，中等城市的监控规模一般为数千到数万个摄像头，以 1080P 为例，在 8 M/s 的码率下，每只摄像头每天产生的视频数据约 84 GB，一般要求这些数据必须在系统中保存 30 d 以上，这对存储空间的大小要求很高；另一方面，这些海量视频监控数据中存在大量时间冗余，即不同帧的行人外观特征之间存在极大的相似性。使用所有的帧图像进行识别会降低查询效率，也会给准确性带来不利影响。如何去除视频时间冗余、提高识别准确率并保留关键的视频帧以减少视频数据的物理存储空间是一个值得研究的问题。

传统的视频行人重识别方法不考虑时间冗余，对视频内所有帧进行最大池化或平均池化以得到视频级特征。平等地对待所有帧不仅会耗费巨大的计算代价，也会因为大量噪声的存在导致算法性能的退化，所以提取关键帧尤为重要。视频行人重识别中的关键帧提取是在相似的特征组中只保留一个特征，仅使用部分具有鉴别力的特征进行识别以提高识别准确率和效率，一般来说关键帧是视频内不同视角或者不同行人姿态的帧。近年来，相关研究[3-7]致力于解决时间冗余的问题，他们从视频序列中选择有鉴别力的帧生成视频特征。虽然这些方法一定程度地解决了时间冗余的问题，但是依然存在一些不足。文献[3-5,6]切断了视频的时间连续性，文献[6] 过于依赖行人检测框的质量，并且对摄像头角度变化及行人姿态变化表现出较差的鲁棒性，而且这些工作都是在训练或测试中减少行人特征的冗余，无法减少实际的数据存储空间。针对上述不足，为了解决海量视频数据带来的查询效率低下和准确率下降的问题，本文提出一个关键帧提取算法，既可以保留视频时间连续性又在实际查询操作之前完成，减少视频数据的存储空间。在对行人步行姿态的观察中，可以注意到：(1)行人步行时双脚交替运动，具有明显的周期性，而时间特征就蕴含在这些步行周期内，这说明步行周期可以作为划分视频数据的最小单位；(2)在这样的周期性运动中，脚部运动最为明显，即行人步行时，双脚之间的距离具有周期变化，呈现由小变大再变小的规律。基于上述两点，本文设计了基于步行周期聚类的关键帧提取算法（Walking cycle clustering based key frame extraction, WCC-based KFE）：第 1 步利用预训练好的人体姿态估计模型获取视频序列中行人双脚距离，根据距离的周期性变化规律提取所有的候选步行周期；第 2 步获得所有候选步行周期的特征，再利用聚类方法选取核心特征，仅保留其对应的关键步行周期以减少时间冗余。未保留的步行周期存在两种情况：(1)与簇中心距离近，此时使用簇中心统一表示既减少了数据量又保留了重要特征；(2)与簇中心距离远，此时该步行周期属于干扰或噪声，应将其去除，否则会影响识别准确率。本文提出了一个新的行人重识别框架将 WCC-based KFE 算法与行人重识别网络结合起来。在查询之前对视频数据进行处理可以大量减少数据存储空间，更加适用于实际应用。此算法的优点是在保留视频时间连续性和行人特征多样性的情况下，减少了时间冗余，去除了干扰和噪声，提高了视频行人重识别的准确率，而且 WCC-based KFE 算法在行人重识别网络训练和测试之前完成，节省了 59%～82% 的数据量。

1　相关研究工作

视频行人重识别对图像质量要求不高，使用场景更广且赋含信息更多，包含了帧与帧之间的时间信息、运动信息等[8]，这更有利于提高行人检索的准确率。但是它不仅面临着物体遮挡、姿态变化和光照变化等问题带来的挑战；而且存在视频数据独有的问题，例如数据量更大、计算量更大且存在高度冗余。近年来，越来越多的学者关注视频行人重识别，针对目前该研究领域存在的各种问题提出了相应的方法。

(1)行人遮挡。学者们大多引入注意力机制，弱化遮挡图像给网络模型带来的负面影响。例如，2017 年，Zhou 等[9]提出时间注意模型来衡量视频序列中每一帧的重要性，认为严重遮挡的帧是“坏”帧，将其剔除，仅对质量好的帧进行特征提取。同样，Xu 等[10]设计了注意力时间池化使网络模型给予包含有效信息的帧更多的权重。与上述两个工作在时间层面利用注意力机制不同，2018 年，Li 等[11]首先通过空间注意模型自动发现不同的身体

部位，提取质量好的局部区域特征，再利用时间注意模型进行组合。Hou等[12]认为丢弃遮挡图像的方法并不理想，因为它中断了视频的时间信息，于是提出了时空补全网络（Spatial-temporal completion network，STCnet），根据行人的身体空间结构，利用可见的身体部分预测缺失的部分，然后基于视频的时间连续性，利用相邻帧的信息来恢复当前帧的行人外观，从而解决物体遮挡的问题。它们都取得了一定的效果，但是仍然存在不足：(1)文献[9-11]仅对质量好的帧进行处理会严重破坏时间特征；(2)视频本身存在大量冗余，过多信息会降低算法的查询效率，而文献[12]补全图像遮挡部分的操作会额外花费大量的计算成本。

(2)姿态变化。2019年，Chen等[7]提出了一种基于KFS（Key frame selection）训练策略，首先将视频分成长度相等的片段，选择与前一个片段姿态变化最大的一帧作为关键帧，将所有的关键帧作为训练数据提高网络模型对姿态变化的鲁棒性。同年，Wu等[13]提出了一种半监督的方法，将训练好的姿态估计模型直接应用到行人重识别数据集上，避免了在行人重识别数据集上标注姿态的麻烦。其中，他们根据行人不同姿态对图像进行定位和分割，提取对应位置的行人外观特征以解决姿态变化的问题。

(3)时间冗余。目前较少学者关注到视频行人重识别时间冗余、计算成本高的问题。现有解决方法通常是采用关键帧提取的方法减少时间冗余，其中根据关键帧的性质不同，可以将现有方法分为两类：第1类以帧作为最小单位，此时关键帧不一定是连续的；第2类以步行周期作为最小单位，此时关键帧至少包含1个连续的步行周期。

第1类方法。2018年，Zhang等[5]训练一个“代理”，每次只验证2个视频序列中的一对图像是否属于同一个人，若能得到肯定回答：相同或者不同，则输出结果，此时只使用了2幅图像；若无法得到肯定回答：不确定，则加入另一对图像进行验证。不断循环，直到得到肯定回答。此方法的优点是对于一些简单样本，使用极少量的图像就能判断2个视频序列是否属于同一个人，缺点是忽略了视频的运动特征，仅使用表观特征来进行识别。Chen等[14]首先将查询和候选视频序列划分为多个固定长度的短视频片段，将片段相似性最大的认为是该视频序列的相似性，从而最小化序列中行人的外观变化。此方法划分的片段不具有完整的时间特征，而且虽然减少了时间冗余，但需要计算所有片段对的相似度，计算量庞大。2019年，Song等[15]提出“主图像组”的概念，认为图像序列中与平均特征距离最小的3帧为该图像序列的“主图像”，从“主图像”中提取行人的空间上下文特征以减少时间冗余的不利影响。与上一个工作不同的是，Zhang等[16]首先利用FEP（Flow energy profile）信号划分步行周期，根据信号值的变化，每个步行周期选择4帧图像作为关键帧，缺点仍然是中断了视频的时间连续性。

第2类方法。2019年，Gao等[6]跟踪图像中行人脚部的超像素，根据其在行人检测框的水平位置来提取步行周期，认为超像素水平位置曲线最接近正弦曲线的周期是最佳步行周期，然后仅使用最佳步行周期来表示该行人。但是，该方法对行人检测框精度要求高且行人行走方向和摄像机拍摄角度可能存在一定夹角，以人体最低部位来划分步行周期不具有可靠性，而且该方法只使用一个最佳步行周期去代表一个人，没有考虑到行人姿态变化导致的外观多样性。

此外，这些工作共同的缺点是没有减少实际的物理存储空间，随着监控点越多、拍摄时间越长、图像质量越高，视频数据量呈几何指数增长，算法执行将耗费大量的计算成本。

本文将提出的基于步行周期聚类的关键帧提取算法用于视频行人重识别中，通过重识别准确率评价关键帧提取算法的有效性，因此需要对原有视频行人重识别算法框架进行调整，调整后框架结构如图1所示，旨在查询前对数据进行处理，减少数据量以提高查询效率。

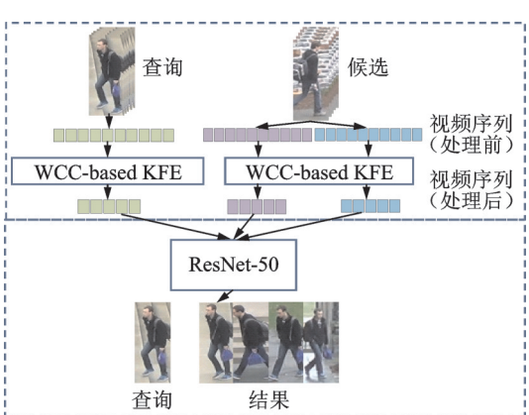

图1 框架示意图

Fig.1 Schematic diagram of the framework

框架主要有两步：(1)分别将查询集和候选集里的每一个视频序列输入到WCC-based KFE算法中，该算法是本框架的核心，具体见第2节。算法的输出是由关键步行周期组成的新序列，新序列是原视频序列的子集，且长度短很多。(2)将新查询集和新候选集一起输入到基础网络中进行识别，

并输出查询结果。只对查询集和候选集进行处理的原因是现实应用中大量占据存储空间的是测试集，因为只有测试集的数据由于不断拍摄而不断增加，而训练集不会发生变化。为了和别人统一比较，本文使用 ImageNet 预训练好的 ResNet-50 模型作为基础网络。

2　基于步行周期聚类的关键帧提取算法

视频行人重识别面临着数据量庞大、时间冗余等问题，这会严重影响重识别效率和准确率，本文提出基于步行周期聚类的关键帧提取算法来解决上述问题。算法输入是一个视频序列，输出是只包含关键步行周期的短的视频序列。如图 2 所示，主要分为两步：(1)提取候选步行周期，如图 2(a～d)所示；(2)提取关键步行周期，如图 2(d～f)所示。

2.1　提取候选步行周期

WCC-based KFE 算法通过提取关键步行周期在保留视频时间连续性的同时减少时间冗余，所以首先要得到关键步行周期的候选集，即候选步行周期。由图 2(a)可知，行人行走双脚交替运动，具有明显的周期性，而最能够从视频中反映的就是行人双脚之间的距离，距离满足由小变大再变小的变化规律，所以本文根据行人双脚距离来划分周期，提取候选步行周期的具体方法如下：

(1) 获得视频序列里每帧图像中行人双脚之间的距离。本文采用的是开源人体姿态识别项目 OpenPose。算法的输入是查询或者候选集中任意一个视频序列 $\text{tracklet}=\{\text{frame}_1, \text{frame}_2, \cdots, \text{frame}_n\}$，将每帧图像 $\text{frame}_i(i\in[1,n])$ 输入到预训练好的 OpenPose 模型中，分别提取图像中行人双脚的位置，计算出距离 d_i，此处计算的是双脚像素点之间的距离，所以 d_i 单位是像素。如图像存在严重遮挡，提取不到双脚位置，则定义 $d_i=-1$，表示无效值，得到距离序列 $\text{Distance}=[d_1, d_2, \cdots, d_n]$。

(2) 根据式(1)中获得的距离序列 Distance 划分周期。距离序列可视化如图 2(c)所示，曲线上每一个点的值是对应帧图像中行人双脚之间的距离。该曲线具有明显的周期性，不同周期之间行人的外观特征和运动特征都具有高度的相似性，这也是视频数据时间冗余的原因。

曲线中的红色三角形代表的是极小值，每一个极小值是上一个周期的结束，也是下一个周期的开始，所以将相邻两个极小值之间的帧划分为一个周期。极小值定义为

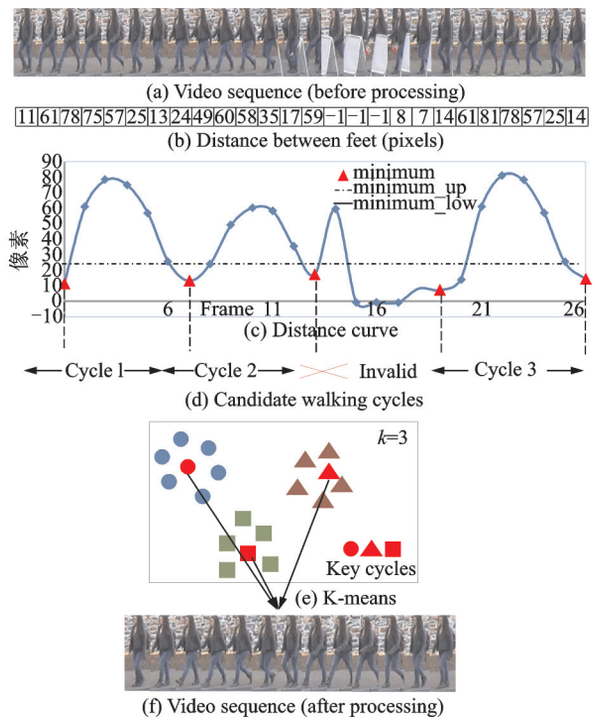

图 2　WCC-based KFE 算法示意图

Fig.2　Schematic diagram of WCC-based KFE algorithm

$$\text{minmum}=\left\{d_i \middle| (d_i<d_{i-1})\wedge(d_i<d_{i+1})\wedge (\text{mini\_}l\leqslant d_i\leqslant \text{mini\_}u)\right\} \tag{1}$$

式中：$\text{mini\_}l$ 和 $\text{mini\_}u$ 是极小值取值的下界和上界，如图 2(c)中 minimum_low 和 minimum_up 两条虚线所示。极小值必须为有效值，即不等于“−1”，而规定极小值取值上界的原因是 OpenPose 模型输出的结果有小范围误差，因为偶然性的误差，可能导致距离值较大的点符合数学上极小值的定义，但它显然不能看作是一个步行周期的结束。

得到所有的极小值之后，将相邻两个极小值之间的对应帧图像提取出来作为一个周期，例如图 2(c)中，极小值集合为 $\{d_1, d_7, d_{13}, d_{20}, d_{26}\}$，则第 1 个周期为 $\{\text{frame}_1, \text{frame}_2, \cdots, \text{frame}_7\}$，第 2 个周期为 $\{\text{frame}_7, \text{frame}_8, \cdots, \text{frame}_{13}\}$，以此类推。

(3) 判断前一步提取的周期是否有效。有效即该周期内的所有帧具有完整的行人外观特征，不存在严重遮挡问题。在式(2)中，只考虑了极小值对应的帧不能存在严重遮挡问题，并没有考虑一个周期内其他帧的情况。设某个周期长度为 m，表示为 $\text{Cycle}=\{\text{frame}_i, \text{frame}_{i+1}, \cdots, \text{frame}_{i+m}\}$，则其对应的距离序列 $\text{Distance}_{\text{cycle}}=[d_i, d_{i+1}, \cdots, d_{i+m}]$，$d_i, d_{i+m}\in\text{minmum}$，则其为有效周期的条件是

$$\sum_{t=i}^{i+m}\text{count}(d_t=-1)\leqslant\delta \tag{2}$$

式中：$1\leqslant i\leqslant n-m$，$\text{count}(\cdot)$ 为计数函数；δ 为无

效阈值,即一个周期内,距离无效值允许出现的最大次数。没有严格要求周期内所有帧的行人双脚距离都为有效值的原因是:行人检测框存在一些误差,当行人步幅很大时,行人检测框不能将完整的行人框出,可能缺少行人的脚部,从而导致出现距离无效值。但是这样的问题不会导致出现连续多帧图像中行人双脚距离均为无效值的情况,所以定义无效阈值避免因这类问题而导致的错误判断。

综上所述,提取候选步行周期这一步的输入是原始的视频序列,首先利用OpenPose人体姿态估计模型获得每帧图像中行人双脚的位置,然后根据得到的距离序列划分周期,并判断每个周期是否有效,候选步行周期就是该视频序列所有的有效周期,最后输出所有的候选步行周期。

2.2 提取关键步行周期

在获得一个视频序列里的所有候选步行周期之后,下一步是判断这些候选步行周期是否为关键步行周期,此时必须要考虑周期间特征的关系,去除特征提取模型输出的特征与关键步行周期的特征相似的冗余周期,最后将剩余的关键步行周期合并作为新的视频序列。

(1)获得所有候选步行周期的特征,本文使用预训练好的Resnet-50网络作为特征提取模型。设一个视频序列中提取出j个候选步行周期,则经过特征提取后得到的特征表示为$\{\text{feature}_1, \text{feature}_2, \cdots, \text{feature}_j\}$。

(2)从所有特征中选择核心特征,认为其对应的候选步行周期即为关键步行周期。使用K-means聚类的方法在j个候选步行周期中选出k个核心特征对应的关键步行周期,核心特征是K-means结果的每一个簇中最靠近簇中心的特征,可以代表整个簇的情况。K-means聚类采用距离作为相似性的评价指标,即认为两个对象的距离越近,其相似度就越大。该算法采用贪心的迭代方法直到得到紧凑且独立的簇。k个簇集合C定义和簇中心μ定义为

$$C^p = \arg\min\left\{\sum_{u=1}^{k}\sum_{\text{feature}_l \in C_u^p}\left\|\text{feature}_l - \mu_u^{p-1}\right\|_2^2\right\} \tag{3}$$

$$\mu_u^p = \frac{1}{n_u}\sum_{\text{feature}_l \in C_u^p}\text{feature}_l \tag{4}$$

式中:$1 \leqslant l \leqslant j$;$p$为K-means算法的当前迭代次数,当前迭代次数的C是由上一次迭代中心点结果μ根据式(3)计算得来。初始μ是在所有样本中随机选择的。$n_u = |C_u|$是簇C_u中样本的个数。如图2(e)所示:K-means将特征相似的候选步行周期聚成一个簇,总共有k个簇。关键步行周期定义为每个簇中最靠近聚类中心的特征对应的周期,即有

$$\text{key cycles} = \left\{\text{candidate cycle}_l \middle| \arg\min\left\|\text{feature}_l - \mu_u\right\|_2^2\right\} \tag{5}$$

若$j > k$,保留k个关键步行周期,其他候选步行周期为冗余周期去除。若$j < k$,则认为j个候选步行周期均为关键周期。最后,将所有关键步行周期内的所有帧合并作为新的视频序列。这时视频序列中只有具有代表性特征的帧,去除了时间冗余和质量差的帧,且保留了行人行走的时间连续性。

k的取值与数据集摄像头安装位置和行人行走方向有关,行人的不同角度之间的图像特征差距很大,尽可能保留不同角度的行人图像可以提高识别的鲁棒性。所以k个关键步行周期可以理解为k个不同的角度的行人图像序列。当k=1时,仅保留一个角度的行人图像序列。

如图3所示,行人拍摄角度可以大致分为8个,分别是正北、正南、正东、正西、东北、西北、东南、西南。当行人在固定位置的摄像头前走过时,行人被拍摄到的只是身体的一侧,角度为5个,即一般情况下,一个视频序列中存在5个角度,所以k的最佳取值在5左右。

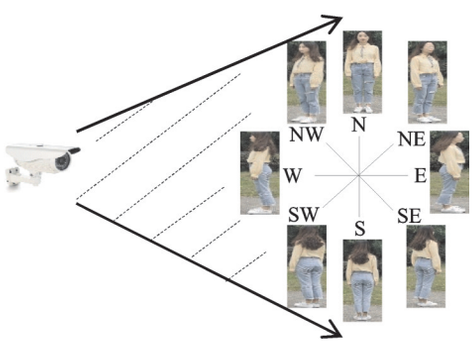

图3 k取值示意图

Fig.3 Schematic diagram of k value

3 实验结果与分析

3.1 数据集与评价指标

本文在公开数据集MARS和DukeMTMC-VideoReID上开展实验,与国际前沿进行对比,评价算法的有效性。MARS数据集[17]于2016年发表,拍摄于清华大学校园,是第1个可以用于深度学习的大型视频行人重识别数据集。MARS由6台摄像机拍摄,总共有1 261个不同的行人,625个用于训练,636个用于测试,总共有超过20 000个图像序列,每个行人至少被2个摄像机捕获。测试集大小一共有3.8 GB,其中干扰项有0.9 GB。DukeMTMC-VideoReID数据集拍摄于杜克大学,

是多摄像头跟踪数据集DukeMTMC[18]的子集，包括702个用于训练的身份，702个用于测试的身份，以及408个干扰项。总共有21 96个视频用于训练，2 636个视频用于测试。测试集大小一共为2.2 GB。本文使用累积匹配曲线（Cumulative match characteristic，CMC）和平均精度均值（mean Average precision，mAP）[19]作为评价标准。其中，使用Rank-1，Rank-5，Rank-20代表CMC曲线。CMC更关注准确率，而mAP同时关注准确率和召回率。

3.2　实验说明

（1）Resnet-50网络。本文实验使用的GPU是NVIDIA TITAN Xp且在PyTorch框架下实现，使用ImageNet预训练的Resnet-50网络进行训练和测试。在训练时，训练集不变，为了节省GPU内存，随机采样16帧作为输入，所有视频序列中的帧级特征经过平均池化形成视频级特征。本文采用动量为0.5，权值衰减为0.000 5的随机梯度下降法进行参数优化，迭代次数为70，批大小为8，学习率初始值为0.1，在最后15次迭代时，调整为0.01。

（2）WCC-based KFE。极小值下界mini_l设置为0，上界mini_u设置为20，比0小意味着提取不到行人双脚位置，该帧有严重的遮挡问题，不能作为极小值。通过观察输出结果，本文发现当距离值小于20时，行人双脚位置几乎重合，可以作为极小值，所以定义20像素是极小值取值的上界。

无效阈值δ设置为2，即有效周期中距离无效值出现次数不能大于2次。例如图2（c）中第3个周期，距离序列为{17，59，－1，－1，－1，8，7}，其中距离无效值“－1”出现了3次，不满足有效周期定义，将其去除，其余3个均为有效周期。

k的取值决定每个视频序列保留几个关键步行周期。本文k设置为6，在具体实现时，若一个视频序列没有候选步行周期，则对它不进行操作，保留原视频序列，但是这样的情况是极个别的。特别地，MARS数据集候选集有0.9 GB的干扰项，这些图像不包含完整的行人，所以不进行处理。

3.3　实验结果

本节探讨WCC-based KFE算法的两步处理操作对模型准确率和效率的提升。结果如表1所示，Baseline方法是测试集不使用WCC-based KFE算法处理，直接输入到ResNet-50网络进行识别的结果。Baseline＋WCC-based KFE（1）方法是测试集只经过WCC-based KFE算法的第1步结果，即仅提取候选步行周期。Baseline＋WCC-based KFE（1）＋（2）方法是测试集完整经过WCC-based KFE算法第2步的结果，即提取了关键步行周期。

从表1中可以看到，Baseline＋WCC-Based KFE（1）方法的结果比Baseline方法好，对于MARS数据集来说，CMC Rank-1提高了0.3%，mAP提高了5%。对于DukeMTMC-VideoReID数据集来说，CMC Rank-1提高了0.4%，因为虽然只得到了候选步行周期，但是这一步操作也过滤掉了大量的严重遮挡的帧，提高了准确率。而Baseline＋WCC-based KFE（1）＋（2）方法进一步过滤了质量不好的帧，以及去除了时间冗余，所以CMC Rank-1又比Baseline＋WCC-based KFE（1）方法上升了。MARS数据集的Rank-1提高了0.4%，DukeMTMC-VideoReID数据集的Rank-1提高了1%。

在评价指标CMC和mAP都提高的同时，测试集数据量却在不断的变小。对于MARS数据集，WCC-based KFE第1步操作处理后，数据量减少了24%，两步都处理后，数据量减少了58.6%。对于DukeMTMC-VideoReID数据集，WCC-based KFE第1步操作处理后，数据量减少了31.8%，两步都处理后，数据量减少了81.8%。

3.4　与其他方法对比

本文在MARS和DukeMTMC-VideoReID两个数据集上进行实验，将本文提出的方法与其他行人重识别方法进行比较，结果如表2、3所示。其中，K-reciprocal和See the Forest方法关注时间池化，Latent Parts、SRM＋TAM、QAN和DSAN是基于注意力机制的方法。

从表2、3中可以看出，本文提出的方法在使用更少数据量的同时，CMC和mAP也比现有的方法有所提高。准确度提升主要有两个原因：（1）WCC-based KFE算法的第1步过滤了严重遮挡的帧，例如汽车遮挡、垃圾桶遮挡等，避免了物体遮挡导致的错误识别；（2）算法第2步去除了时间冗余，过滤了图像质量差的帧，仅使用关键步行周期所在的帧代表该行人，进一步提高了准确率。

3.5　k取值分析

本节讨论k的不同取值对测试集数据量以及评价指标CMC和mAP的影响，实验结果见表4。其中MARS中测试集的数据量未计算干扰项。从表4中可以看到，随着k值从1增加到8，测试集的数据量在不断增加。对于MARS数据集来说，当k=1时，测试集数据量只有0.5 GB，k值每增加1，测试集数据量增加100～200 MB，当k=8时，测试集数据量增加到了1.4 GB。对于DukeMTMC-VideoReID数据集来说，当k=1时，测试集数据量仅有108.7 MB，k值每增加1，测试集数据量增加60～90 MB，当k=8时增加到了

表 1 WCC-based KFE 算法两步操作对数据量和准确率的影响

Table 1 Influence of two-step operation of WCC-based KFE algorithm on data volume and accuracy

处理方法	MARS						DukeMTMC-VideoReID					
	Rank 1	Rank 5	Rank 10	Rank 20	mAP	测试集数据量/GB	Rank 1	Rank 5	Rank 10	Rank 20	mAP	测试集数据量/GB
Baseline	77.7	89.2	92.0	93.8	64.2	2.9	87.6	96.6	97.9	98.4	83.9	2.2
Baseline+WCC-based KFE(1)	78.0	91.7	94.5	**96.2**	**69.2**	2.2	88.0	96.3	97.7	98.1	83.7	1.5
Baseline+WCC-based KFE(1)+(2)	**78.8**	**91.8**	**94.7**	96.1	68.5	1.2	**89.0**	**96.7**	**98.3**	**98.7**	**84.1**	0.4

表 2 MARS 数据集中各方法比较

Table 2 Comparison of methods in the MARS dataset

处理方法	MARS			
	Rank 1	Rank 5	Rank 20	mAP
QAN[20]	73.7	84.9	91.6	51.7
K-reciprocal[21]	73.9			68.5
See the Forest[9]	70.6	90.0	97.6	50.7
Zhang et al.[5]	71.2	85.7	94.3	
DSAN[22]	73.5	85.0	97.5	
Ours	**78.8**	**91.7**	**95.7**	**68.5**

表 3 DukeMTMC-VideoReID 数据集中各方法比较

Table 3 Comparison of methods in the Duke MTMC-VideoReId dataset

处理方法	DukeMTMC-VideoReID			
	Rank 1	Rank 5	Rank 20	mAP
EUG(supervised)[23]	83.6	94.6	97.6	78.3
Ours	**89.0**	**96.7**	**98.7**	**84.1**

642.5 MB。但是总体而言,随着 k 值的变大,数据量增幅在不断变小,因为会有更多候选步行周期个数小于 k 的视频序列。

随着数据量的不断增加,CMC 和 mAP 并没有随之不断增加。表 4 中红色标记为该评价指标最高值,蓝色标记为第二高的值。可以看到当 k=5 时,几乎所有的评价指标排名均在前两位。而对于最重要的 Rank1 和 mAP 来说,MARS 数据集 Rank1 达到了 78.8,mAP 达到了 68.5,而 DukeMTMC-VideoReID 数据集 Rank1 达到了 89.0,mAP 达到了 84.1,这些都是所有 k 取值当中最高的。

表 4 不同 k 值时的数据量和评价指标

Table 4 Data volume and evaluation indexes at different k values

数据集	MARS						DukeMTMC-VideoReID					
	测试集数据量/GB	Rank 1	Rank 5	Rank 10	Rank 20	mAP	测试集数据量/GB	Rank 1	Rank 5	Rank 10	Rank 20	mAP
k=1	0.5	74.9	90.4	93.4	95.3	64.6	108.7	80.1	91.0	94.3	96.3	73.0
k=2	0.7	76.2	90.9	94.0	96.3	66.7	199.1	86.3	96.0	97.3	98.3	80.8
k=3	0.9	76.8	91.3	93.9	96.1	67.7	281.5	86.3	95.6	97.2	98.1	82.3
k=4	1.0	78.0	91.7	94.1	95.6	68.2	359.5	88.3	96.0	97.0	98.0	83.8
k=5	1.2	78.8	91.8	94.7	96.1	68.5	435.2	89.0	96.7	98.3	98.7	84.1
k=6	1.3	77.5	91.6	94.7	96.2	67.7	508.3	87.6	96.6	98.0	98.3	83.4
k=7	1.4	77.6	91.5	93.9	96.1	68.4	576.9	87.5	96.4	97.7	98.6	83.2
k=8	1.4	77.2	91.5	94.3	96.4	68.0	642.5	87.5	96.9	98.0	99.0	83.2

3.6 数据量分析

本节分析 WCC-based KFE 算法对测试集数据量的影响。如图 4 所示,MARS 数据集中处理前的测试集共有 2.9 GB,处理后剩余 1.2 GB,比例为原来的 41.38%。DukeMTMC-VideoReID 数据集处理前的测试集共有 2.2 GB,处理后仅剩 0.4 GB,比例为原来的 18.18%。虽然都减少了大量的数据量,但是 MARS 处理后数据比例仍然较高。如图 5 所示,两个数据集处理后的长度都为 30 帧左右,但是 MARS 数据集中每个视频序列原始长度较短,平均只有 50 多帧,而 Duke-

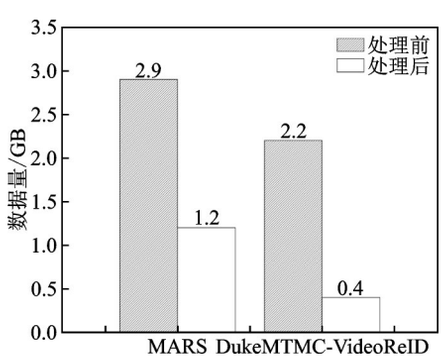

图 4 处理前后测试集数据量

Fig.4 Data volume of test set before and after processing

MTMC-VideoReID 数据集中每个视频序列原始长度很长，在 160 帧左右，所以 DukeMTMC-VideoReID 数据集节省的存储空间比例更大，这也说明本文提出的算法对更长的视频序列处理效果更好。

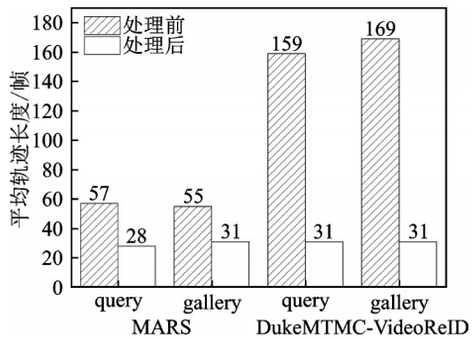

图 5　处理前后测试集平均轨迹长度

Fig.5　Average track length of test set before and after processing

4　结　　论

本文针对视频行人重识别研究问题中存在的数据量庞大、时间冗余等问题设计了基于步行周期聚类的关键帧提取算法，并提出一个新的框架将该算法与视频行人重识别网络结合起来。该算法在查询之前完成，可以减少大量的数据存储空间，同时因为去除了时间冗余及噪声，准确率也得到了提高。

参考文献：

[1] 李梦静，吉根林．视频行人重识别研究进展[J]．南京师大学报(自然科学版)，2020，43(2)：120-130.
LI Mengjing, JI Genlin. Research progress of video-based person re-identification [J]. Journal of Nanjing Normal University (Natural Science), 2020, 43(2): 120-130.

[2] 李幼蛟，卓力，张菁，等．行人再识别技术综述[J]．自动化学报，2018，44(9)：1554-1568.
LI Youjiao, ZHUO Li, ZHANG Jing, et al. Overview of person re-identification technology [J]. Chinese Journal of Automation, 2018, 44(9): 1554-1568.

[3] 黄凯奇，陈晓棠，康运锋，等．智能视频监控技术综述[J]．计算机学报，2015，38(6)：1093-1118.
HUANG Kaiqi, CHEN Xiaotang, KANG Yunfeng, et al. Overview of intelligent video surveillance technology [J]. Chinese Journal of Automation, 2015, 38(6): 1093-1118.

[4] YOUSRA H H, WALID A, TAREK O, et al. Multi-shot person re-identification approach based key frame selection [C]//Proceedings of the Eighth International Conference on Machine Vision. Barcelona, Spain: [s.n.], 2015, 98751H: 1-6.

[5] ZHANG J F, WANG N Y, ZHANG L Q. Multi-shot pedestrian re-identification via sequential decision making [C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: [s.n.], 2018: 6781-6789.

[6] GAO C X, WANG J, LIU L Y, et al. Superpixel-based temporally aligned representation for video-based person re-identification [J]. Sensors, 2019, 19(18): 3861-3881.

[7] CHEN Y Z, HUANG T D, NIU Y Z, et al. Pose-guided spatial alignment and key frame selection for one-shot video-based person re-identification [J]. IEEE Access, 2019, 7: 78991-79004.

[8] 罗浩，姜伟，范星，等．行人重识别研究进展[J]．自动化学报，2019，45(11)：2032-2049.
LUO Hao, JIANG Wei, FAN Xing, et al. Research progress of person re-identification based on deep learning [J]. Chinese Journal of Automation, 2019, 45(11): 2032-2049.

[9] ZHOU Z, HUANG Y, WANG W, et al. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification [C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: [s.n.], 2017: 6776-6785.

[10] XU S J, CHENG Y, GU K, et al. Jointly attentive spatial-temporal pooling networks for video-based person re-identification [C]//Proceedings of the International Conference on Computer Vision. Venice, Italy: [s.n.], 2017: 4743-4752.

[11] LI S, BAK S, CARR P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification [C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: [s.n.], 2018: 369-378.

[12] HOU R B, MA B P, CHANG H, et al. VRSTC: Occlusion-free video person re-identification [C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: [s.n.], 2019: 7183-7192.

[13] WU J J, JIANG J G, QI M B, et al. Independent metric learning with aligned multi-part features for video-based person re-identification [J]. Multimedia Tools and Applications, 2019, 78(20): 29323-29341.

[14] CHEN D P, LI H S, XIAO T, et al. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding [C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA:

[s.n.], 2018: 1169-1178.

[15] SONG W R, WU Y H, ZHENG J Y, et al. Extended global-local representation learning for video person re-identification [J]. IEEE Access, 2019, 7: 122684-122696.

[16] ZHANG W, HU S N, LIU K, et al. Learning compact appearance representation for video-based person re-identification [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(8): 2442-2452.

[17] ZHENG L, BIE Z, SUN F Y, et al. MARS: A video benchmark for large-scale person re-identification [C]//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: [s. n.], 2016: 868-884.

[18] ERGYS R, FRANCESCO S, ROGER S, et al. Performance measures and a data set for multi-target, multi-camera tracking [C]//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: [s.n.], 2016:17-35.

[19] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: A benchmark [C]//Proceedings of the International Conference on Computer Vision. Santiago, Chile: [s.n.], 2015: 1116-1124.

[20] LIU Y, YAN J J, OUYANG W L. Quality aware network for set to set recognition [C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: [s. n.], 2017: 4694-4703.

[21] ZHONG Z, ZHENG L, CAO D L, et al. Re-ranking person re-identification with k-reciprocal encoding [C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. HI, USA: [s. n.], 2017: 3652-3661.

[22] WU L, WANG Y, GAO J B, et al. Where-and-when to look: Deep siamese attention networks for video-based person re-identification [J]. IEEE Transactions on Multimedia, 2019, 21(6): 1412-1424.

[23] WU Y, LIN Y T, DONG X Y, et al. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning [C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: [s.n.], 2018: 5177-5186.

（编辑：刘彦东）