摘要

行人重识别任务旨在从海量监控摄像头拍摄的视频中精准识别出给定的查询行人。作为细粒度检索任务，行人重识别相较于传统的检索任务具有更大的类内差异和更小的类间差异，因此检索难度更高。对行人重识别的研究有助于推动图像检索任务及相关研究领域的发展，具有重要的科研意义。当前主流的行人重识别研究专注于设计高判别力的特征用于检索行人。然而，现实场景中行人的姿态变化和检测器的检测误差会导致行人在图像中表现出的外观发生变化，降低了行人特征的判别力。此外，海量的监控视频中存在外观极为相似的不同行人，难以通过图像中的视觉信息进行区分。将人体姿态信息用于行人特征学习有望缓解以上问题。然而，监控视频中行人尺度变化大，增加了人体姿态估计的难度，限制了姿态信息在行人特征学习中的应用。因此，当前行人特征设计面临着三个难题：行人尺度变化影响姿态估计准确率；姿态变化和检测误差降低了图像特征的判别力；外观相似的行人难以使用图像特征区分。本文针对上述三个问题展开研究，分别从提高人体姿态估计准确率、增加行人图像特征对姿态变化和检测误差的鲁棒性、高判别力的视频特征学习三个方面提出相应的方法。本文的主要创新点如下：

第一，针对行人尺度变化影响人体姿态估计准确率的难题，本文提出了一种基于极坐标回归的人体姿态估计方法，用于降低人体关键点回归的难度。该方法将现有方法中笛卡尔坐标系下二维偏移值回归任务转化为极坐标下角度预测和一维长度预测任务的结合。对于角度的预测，由于关键点角度不随行人尺度变化，本文的方法直接使用固定间隔将角度量化为离散值用于分类。对于长度的预测，本文设计了尺度不变的长度交并比损失函数用于长度优化。相比于现有方法，本文的方法对行人尺度变化更加鲁棒。此外，本文还设计了多中心回归策略用以降低角度量化过程中产生的量化误差，进一步提高了关键点预测的准确率。相比于之前方法，本文的方法更加简洁且不需要额外的多阶段精炼操作，因此可以更加高效地与下游任务结合。在 *COCO* 和 *CrowdPose* 数据集上的实验结果表明，该方法可以准确地回归人体关键点，在准确率和速度上均优于之前的最佳算法。

第二，针对行人姿态变化和检测误差影响行人特征判别力的难题，本文提出了一种姿态指导的行人特征学习方法，用于提升行人图像特征对姿态变化和检测误差的鲁棒性。该方法分别学习姿态不变性特征和局部描述特征用于缓解姿态变化和检测误差。姿态不变性特征的学习使用姿态估计算法预测的关键点对行人的身体部件进行归一化，并使用归一化的行人图像辅助全局特征的学习，用于消除姿态变化的影响。局部描述

特征的学习基于关键点定位多个行人的身体区域，通过多任务学习的方式监督网络关注到更广泛的行人身体区域，避免过拟合到单一身体区域，增加了特征对检测误差的鲁棒性。本文的方法在模型推理阶段不需要额外的姿态估计，可以直接从行人的原始图片提取对行人姿态变化和检测误差鲁棒的全局特征，保证了特征提取的效率。在多个数据集上的实验结果表明该方法能够准确识别姿态变化和检测误差较大的行人。

第三，针对海量监控视频中不同行人外观相似的问题，本文提出了一种高判别力的行人视频特征学习方法，用于从视频中学习动态姿态特征区分外观相似的不同行人。该方法包含用于视频特征提取的多尺度三维卷积神经网络和用于视频特征融合的时间自注意力模型。其中多尺度三维卷积神经网络包含多尺度三维卷积层用于特征提取和残差注意力模块用于特征精炼。时间自注意力模型通过学习全局的时间注意力掩码用于融合视频中的多视角信息，并降低视频中噪音的影响。与现有方法相比，本文提出的方法具有更强的动态姿态特征学习能力且对视频中噪音更加鲁棒。在多个公开数据集上的实验结果表明本文的方法能够有效地区分外观相似的行人。

综上所述，本文面向行人重识别任务中的高判别力行人特征设计，从提高姿态估计准确率、提高图像特征对姿态变化鲁棒性、高判别力视频特征学习三个方面展开研究。本文的研究有效提升了行人重识别的准确率和效率，提高了行人重识别技术的实用性，有望推进行人重识别任务的相关方法的实际落地应用，对智慧城市的建设提供支持。

关键词：行人重识别，人体姿态估计，图像特征，姿态不变特征，视频动态特征

Study on Pose Analysis based Person Re-Identification

Jianing Li (Computer Applied Technology)
Directed by Prof. Shiliang Zhang

ABSTRACT

The person Re-IDentification (ReID) task aims to accurately identify and retrieve the given query person from the videos taken by massive surveillance cameras. As a fine-grained retrieval task, person ReID has larger intra-class differences and smaller inter class differences than traditional retrieval tasks, which makes person ReID more difficult. The research of person ReID is of great significance and value to promote the development of retrieval tasks, and can drive the development of other retrieval related research fields.

The current research on person ReID task focuses on designing high discriminant features to retrieve person from large scale dataset. However, the pose variation and detection errors in real scenes will significantly change person's appearance and reduce the discrimination of person features. In addition, there exists different persons with similar appearance in a large number of surveillance videos, which are difficult to distinguish by image features. Apply human pose information to feature design may alleviate above challenges. However, the person's scale changes reduces the accuracy of pose estimation, limits the application of pose estimation in person feature design. In summary, the research on person ReID faces three challenges: the change of person scale affects the accuracy of pose estimation; the pose variation and detection error reduce the discriminant power of image features; The persons with similar appearance are difficult to distinguish only by image features. This thesis studies high-discriminative feature designing in person ReID task, and proposes corresponding methods from perspectives of improving the accuracy of person pose estimation, increasing feature robustness to pose variation and detection error, and high discrimination video feature extraction. Innovations of this thesis can be summarized as follows:

1) This thesis proposes an efficient regression based human pose estimation algorithm to handle the difficulty in long-distance regression. The proposed method transforms the 2D regression task in Cartesian coordinate into orientation classification task and 1D length regression task in polar coordinate. For orientation prediction, the proposed method directly uses fixed intervals to quantify the orientation into finite discrete values for classification. For

length prediction, this method designs a scale-invariant length intersection and union (IoU) loss, which is more stable to scale change. In addition, a multi center regression strategy is further proposed to reduce the quantization error during angle quantization. The proposed method is able to regress the keypoint offsets in a more reliable way, and achieves more accurate keypoint localization. Experimental results on widely used *COCO* and *CrowdPose* datasets show the effectiveness of the proposed method.

2) This thesis proposes a pose-guided representation learning method for person ReID to alleviate the pose variations and misalignment errors exhibited by person images. The proposed method crop human body part based on pose estimation result, and then normalizes the human body parts to a unified direction and scale through learned affine transformation by the pose transform network to obtain pose normalized image. The pose normalized image is used to assist global feature learning to relieve the influence of pose variation. During inference, the pose invariant can be extracted directly from original image without extra pose estimation. Experimental results on five widely used public datasets demonstrate the competitive accuracy and efficiency of the proposed method.

3) This thesis proposes a high discrimination video feature learning method to learn the motion cues from video sequence to distinguish similar persons. The proposed method consist of Multi-scale 3D convolution network (M3D CNN) for video feature extraction and Temporal Self-Attention (TSA) to video feature fusion. The proposed M3D network consists of M3D layer and Residual Attention Layers (RAL). The M3D convolution layer factorizes the traditional 3D convolution into 2D spatial convolution and 1D temporal convolution, and enhances the temporal cues learning ability through introducing multiple dilation temporal convolutions in parallel. And the RAL is inserted to reduce the weight of occlusion frames and improves the stability of video features. The TSA takes frame features as input and fuses the multi-viewpoint cues in video sequences. Experimental results on five public datasets verify the effectiveness of proposed method in distinguishing visually similar persons.

In conclusion, this thesis studies discriminative person feature learning in person ReID task from perspectives of improving the efficiency of human pose estimation, increasing the robustness of person features to pose variation, and high discrimination video feature extraction. It's expected to promote the practical application of person ReID, and provide support for the construction of smart cities.

KEY WORDS: Person Re-Identification, Pose Estimation, Image Feature, Pose Invariant Feature, Motion Feature