摘要

人体动作识别是计算机视觉领域中非常重要的研究任务之一。随着高科技信息化时代的到来，越来越广泛的实际应用对该技术提出了更高的要求。首先，现有的动作识别方法多针对于视角固定的场景。而在实际应用中，由于摄像机的固定位置和拍摄视角变化很大，动作视频的视角差异会非常显著。因此，跨视角动作识别成为亟待解决的问题。即如何利用某一视角下的运动数据实现对另一视角运动样本的分类识别。同时，随着质优价廉的深度摄像机的出现，基于深度数据的视频分析受到越来越多的关注。与传统颜色数据相比，深度数据直接反映运动特性，并且对环境变换的敏感度较低，非常适用于对人体动作的识别。设计实现针对多源数据的动作识别系统也尤为重要。另外，现有的跨视角动作识别方法多需要较多的参考信息，但在实际应用场景中，能得到的信息非常有限。因此，需要设计对数据要求更低、监督性更弱的识别方法。

本文致力于解决针对多源数据的跨视角动作识别问题。主要从基于时序信息的跨视角动作识别方法和基于深度学习的动作识别方法两个方面进行研究和讨论，具体的研究内容和贡献总结如下：

第一，提出了运动强度时序累积特征（SMA）。该特征利用视频序列时序结构来描述运动特性。由于时序信息的视角不变性，SMA~特征对视角变换具有很强的鲁棒性。同时，为了适用于不同的应用场景，我们提出了三种相互独立的提取运动强度的方法。其中，基于Cuboid~检测器的方案和基于 Harris 3D~检测器的方案可应用于颜色和深度数据视频。基于韦伯定律的感知密集采样方法专门针对深度数据视频。该方案设计时序运动筛选策略提纯密集采样得到的时空感兴趣点，以提高特征的描述力。该 SMA~特征不仅对不同的动作类别有很好的判别力，而且对显著的视角变换有很强的鲁棒性。

第二，提出了弱监督分类学习方法。针对于在目标视角下没有标注信息的跨视角动作识别问题，我们的方法以弱监督的形式进行学习。该方法首先利用 SMA~特征的视角不变性，通过构建粗粒度类别来生成目标视角样本的标注信息。然后，利用源视角和目标视角的视频样本数据学习跨视角度量方法。该方法能够滤除不同视角下相同运动类别视频样本的数据差异性，同时保留足够的运动特性信息，以提高特征的描述力和判别力，进一步提高跨视角动作识别的效果。

第三，构建了基于局部深度特征的深度网络架构，提出了基于比较编码特征的深度学习特征（DA-CCD）。该方法利用深度架构将局部特征映射为高层语义特征表达。其深度结构由多层去噪自动编码机构建而成，有很强的特征学习能力。由此构造的 DA-CCD~特征的判别力强、维度低，且对视角变换具有鲁棒性。 另外，该方法不需要场景结构信息和人体关节骨架信息，这样可以更灵活地应对实际应用场景。

关键词：动作识别，跨视角，深度数据，时序结构，深度学习

Abstract

Human action recognition is one of the most important research areas in computer vision. With the coming of the era of information and technology, this technology should be more effective and efficient to meet the increasing requirement of practical application. Firstly, most of the current methods resort to restricted and single-view scenarios. But in the real applications, significant view variations may be caused by the changing and frequently unknown positions of the camera. So the task of cross-view action recognition is a serious problem. In other word, we should transfer action model learned in one view to another completely different view. Additionally, the emergence of the cheap but high-quality depth sensors triggers significant attentions to video analysis using the depth data. Compared to the color ones, the depth data directly reflect the motion characteristics and are less sensitive to the change of environment, which can be properly applied to human action recognition. It is very important to build an action recognition system for multi-source data. Moreover, the existing cross-view action recognition methods require much more prior information. However, the prior information we can get is very limited in real scenarios. Therefore, recognition technology with weaker supervision should be proposed to meet the demand of practical applications.

Our work addresses the problem of cross-view action recognition for multi-source data. It consists of two parts: the cross-view action recognition based on temporal information and the action recognition based on deep learning. Concretely, the main contents and contributions of this dissertation are as follows:

Firstly, a Sequential Motion Accumulation descriptor is proposed for action representation. We represent action using the temporal structure of motion property, which is view invariant and makes the SMA descriptor be robust to significant view variations. Additionally, we propose three schemes to extract motion intensity. The Cuboid Detector Based Scheme and the Harris 3D Detector Based Scheme can be used for color and depth data respectively. The Perceptual Dense Sampling Based Scheme is designed for depth data. It utilizes temporal motion refinement strategy to purify the dense sampled spatial-temporal interest points for more descriptive action representation. The SMA feature is not only discriminative among different action types, but also robust to significant view variations.

Secondly, a weakly supervised categorization method is presented. It addresses the task of cross-view action recognition without label information in the target view. This method takes advantage of the view invariance of the SMA descriptor to generate target-view categorical prior. Then it utilizes the data properties of both views to learn a cross-view metric in a weakly supervised manner. This method can filter out fluctuations among actions of different views while retaining sufficient descriptiveness and discriminability of the feature, which can further improve the precision of cross-view action recognition.

Thirdly, we build a deep architecture based on local depth feature and propose a novel

action representation called Deep Architecture of Comparative Coding Descriptor(DA-CCD). This method utilizes deep network to reflect local descriptor to high-level semantic representation. The deep architecture is built by marginalized Stacked Denoising Autoencoder with strong capability of feature learning. The proposed DA-CCD feature has the advantage of enhanced discriminability and lower dimensionality, and is robust to view variations. Additionally, it does not require prior information of scenario structure and skeleton data, which can be flexibly applied to practical applications.

Keywords: Action recognition, Cross view, Depth data, Temporal structure, Deep learning