

摘要

人体动作识别是计算机视觉领域的关键问题，在多个领域有着广阔的应用前景。由于存在背景复杂、摄像机运动、遮挡和物体变化等因素，使得如何提取“好”的特征以及获取鲁棒的特征表达，对动作识别至关重要。时空局部特征和“视觉词袋”（BoWs）的表示方法在人体动作识别领域得到了广泛的应用，但是这个框架通常忽略了视频单词之间的时空关系，导致在动作识别过程中的模糊性，尤其是对真实自然场景下拍摄的动作视频，这种性能上的退化尤为明显。

为了解决上述问题，本文的研究从对人体动作模式的建模和特征的表述入手，在人体动作识别中引入时空上下文信息，寻求人体动作鲁棒、有效的表达方式，在特征构造中融入时空局部特征之间潜藏的结构信息及时空约束，以提高动作识别的准确率。在此技术路线的指导下，本文对动作局部特征描述子时空上下文的建模方法进行了分析与研究。本文的主要研究内容与贡献总结如下：

(1) 在特征构造上，从时空兴趣点形成的视频单词在三维时空中的几何关系出发，考虑他们之间的时空邻近性以及共生性，将一个视频单词时空邻域范围内其他视频单词的分布信息作为该单词的时空上下文（context），并通过统计和信息检索里常见的 TF-IDF（Term Frequency-Inverse Document Frequency）加权机制，得到了两种紧凑的动作描述方式——代表性时空视频词组（ST-DVPs）和代表性时空视频单词团体（ST-DVCs），在一定程度上融入了动作中的时空结构信息。

(2) 在特征构造上，为了克服时空局部特征时间信息缺失的问题，采用 KLT 跟踪器对时空局部特征进行跟踪，将得到的时空特征跟踪轨迹作为基本的处理、描述单元。与 ST-DVPs 和 ST-DVCs 相比，它能在更长的时间尺度上对运动进行描述，进而更好地捕获运动的动态变化与转变过程。至于轨迹之间关系的建模，因不同动作在特征分布上存在一些比较稳定的模式，表现在特征点的位置和速度等之间存在一定关系，因而提出轨迹相对位置、相对速度关系元来捕获这类时空关系。

(3) 在动作识别上，采用向量量化技术和 BoWs 的表示方法，将(1) 与(2) 中构造的不同特征与传统的动作描述方式置于统一的处理流程下，并通过直方图级联、多通道核函数学习等方式实现不同特征之间的融合，在本领域较具挑战性的公共测试数据库上，如：KTH 人体动作数据库、YouTube 动作数据库以及 UT-Interaction 交互动作数据库等，均取得了较好的结果，动作识别准确率得到不同程度的提高。

(4) 开发、实现了一个基于时空上下文建模，面向视频监控智能分析技术与开发的动作分类演示系统，该系统界面友好、易于操作，在人体动作数据库和非人体动作数据库（比如，Mouse 数据库）上均取得了较满意的结果。

关键词：动作识别；时空局部特征；特征点轨迹；时空上下文；TF-IDF

Research on Spatial-Temporal Context based Human Action Recognition

Qiong Hu (Computer Applied Technology)

Directed By Prof. Qingming Huang

Action recognition has been a hot spot issue in computer vision field with broad applications in multiple areas in the past few years. Due to background clutter, camera motion, occlusion, object scale and illumination condition changes, how to extract “good” features and acquire robust feature descriptions are crucial to human action recognition. In the state-of-the-art, spatial-temporal local features and Bag-of-Words (BoWs) model gradually become popular, however, this framework neglects the spatial-temporal layout information of local features and causes ambiguity in action recognition task, especially for videos captured in unconstrained scenarios, *i.e.*, for videos “in the wild”. The performance degradation on these videos is huge.

In order to solve this problem, the paper begins with human action modeling and feature expression, leverages spatial-temporal context information to action recognition, and tries to seek some robust and effective action feature representations. The proposed approaches incorporate internal structural information and spatial-temporal constraints in feature descriptors, thus we can expect performance improvement in action recognition by these features. Bearing this idea in mind, this paper does lots of analysis and research on spatial-temporal context modeling for local features. The main work and contribution of this paper are summarized as follows:

- A) Spatial-Temporal Descriptive Video Phases (ST-DVPs) and Spatial-Temporal Descriptive Video Cliques (ST-DVCs) are proposed. These features try to capture the geometrical relationships between video words formed by clustering on Spatial-Temporal Interest Points (STIPs) in 3-D spatial-temporal space. The generation process takes into account of the proximity and co-occurrence of local features and views the local feature distribution in the volumetric neighbourhood of an STIP as its spatial-temporal context. The selection scheme is based on Term Frequency-Inverse Document Frequency (TF-IDF) weighting strategy widely used in information retrieval field.
- B) To make up for temporal information loss of local features, the paper uses the KLT feature tracker to track each spatial-temporal local feature and treats the

tracked feature trajectory snippets as the basic processing and describing unit. Compared with ST-DVPs and ST-DVCs, it can capture the motion information of an action pattern in a longer time scale and better describe the dynamic characteristics and transitions of motion. As to the relationship modeling among feature trajectory snippets, we believe that there exist some stable feature distribution patterns in an action video clip, which lie in the interconnection of position and velocity between local features, so we propose the relative position relation and relative velocity relation descriptors to capture this kind of relation.

- C) As to action recognition, we take the advantage of vector quantization and bag-of-words model to set the proposed features in A) and B) as well as other traditional descriptors in a unified process. Besides, histogram concatenation, Multiple Kernel Learning and so on are utilized to fuse different features. The effectiveness of the proposed features are validated on challenging benchmark datasets, such as, KTH human action dataset, YouTube action dataset and UT-Interaction dataset, *etc.* The results are promising and the performance improves greatly compared with other traditional features.
- D) Finally, we developed and implemented a demonstrative action classification system, which is oriented at research on relevant techniques for video surveillance. The interface of this system is friendly and easy to use. And it gets satisfactory results on both human action and non-human action, *i.e.* Mouse dataset, classification task.

Keywords: action recognition, spatial-temporal local feature, feature trajectory, spatial-temporal context, TF-IDF