基于视觉的视频分类传统方法

传统的视频分类主要研究人体行为识别。人体行为识别兴起于 20 世纪 90 年代，当时主要的研究对象为简单场景下的人体行为或动作的识别。21 世纪以后，随着视频录入设备的普及和互联网的兴起，人体行为识别的研究进入了快速发展的阶段。当前，随着深度网络的发展，越来越多的研究者开始将深度学习的理论和方法应用到视频分类的研究中，研究内容也从人体行为识别扩展到大规模通用多标签视频分类，并且取得了很大的进展。

一、 研究难点

当前，基于视觉的视频分类的主要难点在于：

1. 从视频中提取强有力的特征：即如何从视频中提取出能更好的描述视频的时空（spatio-temporal）特征，特征越强，模型分类识别的效果越好。
2. 特征的编码和融合方法：包括空域（spatio）特征和时域（temporal）特征两方面，在空域，需要编码和融合多种空域特征；在时域，由于一些动作通过单帧的图像无法判断，只能通过时序上的变化进行判断，需要将时序上的特征进行编码和融合，获得对视频的整体描述；在时空域上，需要将空域和时域特征综合利用融合，以获得更好的效果。
3. 高效的算法：需要考虑模型的大小、训练时间和识别的速度等因素，算法越高效越有可能应用到实际场景中。

一般来说，人体行为识别可以分为**特征提取、行为表示和分类**三个阶段。特征提取即从视频中提取出与人体行为相关的具有模式不变性和模式间判别力的特征；行为表示指从统计特征在每个视频中的分布或状态变化得到视频的行为表示；分类是将视频的行为表示分类至某一类人体行为。

早期的人体行为识别主要针对简单场景下的人体行为识别。采用的方法主要是提取全局特征，比如人体轮廓、人体骨架或人体的运动场等，然后跟踪这些全局特征的变化或是计算这些特征在视频中的三维形状作为行动的表示，最后使用隐马尔科夫模型、模型匹配或机器学习中的分类器进行分类。

随着应用领域的不断拓展，需要识别越来越复杂的人体行为，而且视频录制的背景也越来越复杂，从视频中提取可靠的全局特征越来越困难，使得基于全局特征的人体行为识别方法难以满足应用的性能要求。随后，基于局部特征的人体行为识别方法受到越来越多的关注。相比于全局特征，局部特征通过具有一定模式不变性的特征描述符对局部时空区域进行描述，对视角、光照、尺度变化等复杂背景更加鲁棒，而且在特征提取时一般不需要复杂的预处理，计算相对简单。

局部特征提取后，通常使用词袋模型（bog-of-visual-word, BOVW）来描述人体行为。词袋模型最先应用于自然语言处理领域，通过统计“单词”在文档中出现的次数作为文档的表示。研究人员将词袋模型拓展到人体行为识别领域，用于视频中人体行为的表示，相对于早期的方法，词袋模型计算简便、对于遮挡和复杂背景较为鲁棒，在人体行为识别中取得了较大的成功。基于词袋模型生成视频特征向量后，使用强大的判别分类器（如支持向量机）对特征向量进行分类，能够降低人体行为巨大的类内距离对识别精度的影响。由于这些优点，使得局部特征词袋模型已经成为人体行为识别的一种主流方法。

二、 特征提取

特征提取一般可分为局部特征提取和全局特征提取。局部特征提取是指视频中的局部兴趣点或者兴趣区域，比如灰度变化剧烈的局部时空区域。全局特征是指人体轮廓、人体骨架等人体行为整体特征。相比于全局特征，局部特征对视频中的光照、视角、摄像机抖动以及复杂背景等更加鲁棒。

一般来说，人体行为识别中的局部特征提取可分为两个步骤：第一，在视频中确定一个区域；第二，使用特征描述符对这个局部区域进行描述。

局部特征区域既可以是二维的局部空间平面，也可以是三维时空立方体，区域的确定一般有两种方法：局部特征检测和稠密采样。局部特征检测通过一个激励函数遍历视频中的时空区域，当该激励函数在某个时空区域的值大于给定阈值时，该区域即为特征区域，比如 Harris3D 检测器、Cuboid 特征检测器和 Hessian 检测器等，这些方法往往只保留了与人体行为相关的特征，而舍弃与人体不相关的特征，因此又称为稀疏特征。稠密采样则是以固定的步进密集地采样视频中的时空区域，会获得大量的局部特征。实验表明稠密采样方法比特征检测方法的识别精度更高。

特征描述符使得提取的特征对光照、尺度、旋转等非相关因素的变化具有一定的模式不变性，同时又兼具较强的判别力。早期提出的局部特征提取方法多使用基于灰度变化的时空区域检测和特征描述，如梯度直方图（Histogram of Oriented Gradients，HOG）描述表观（Appearance）信息、光流直方图（Histogram of Optical Flow, HOF）描述运动信息、HOG3D 描述符使用三维梯度方向直方图、ESURF 描述符描述三维视频空间。通过对早期常用的局部特征检测器和特征描述符的全面评估，实验结果表明没有一种局部特征检测器或描述符能够在所有的人体行为识别数据集上全面胜出，表现不分伯仲。

解决复杂场景下的人体行为识别问题，仅仅通过检测时空区域内的灰度变化远远不够，因此研究人员提出了很多基于特征点跟踪的特征提取方法。这些方法

首先检测视频中的时空区域内的特征点，然后逐帧跟踪这些特征点并联接形成特征点的轨迹，之后使用特征描述符对轨迹及其时空邻域进行描述。在众多基于特征点跟踪的特征提取方法中，识别精度最好的是稠密轨迹（Dense Trajectories，DT）特征提取方法[1]。DT 方法按照固定的步进以多个尺度稠密采样视频中的每一帧图像，然后检测出空间特征点，并在各个尺度单独跟踪这些特征点形成固定长度的轨迹，最后对每一条轨迹及其时空邻域使用四种特征描述符进行描述，分别是用于描述轨迹本身的轨迹形状描述符（Trajectory Shape Descriptor，TSD）、描述轨迹邻域信息的运动边界直方图（Motion Boundary Histograms，MBH）、描述表观信息的 HOG 和描述运动信息的 HOF 描述符。考虑到摄像机运动导致视频中提取出与人体行为无关的 DT 特征，进一步对 DT 特征进行改进，提出了改进的稠密轨迹（Improved Dense Trajectories，IDT）方法[2]。IDT 方法的改进之处在于通过匹配前后两帧间的 SURF 描述符和稠密光流特征点，来估计相机的运动，消除相机运动带来的影响。特征提取后，DT/IDT 方法利用 FV（Fisher Vector）方法对特征进行编码，再基于编码特征向量训练支持向量机（Support Vector Machine， SVM）分类器实现人体行为识别。DT/IDT 的缺点在于算法的速度很慢。

三、 行为表示

行为表示包括特征编码、池化和归一化等一系列操作，最终形成描述视频的归一化特征向量。

特征编码就是将连续特征空间中的特征量化，得到特征编码向量。如词袋模型中将每个特征量化至词典中的一个词条。为降低量化误差，提出了很多将每个特征量化至一个以上的词条，如量化至基于核函数构建的全部词条上、基于稀疏编码的特征编码方法、基于位置约束的线性特征编码方法和基于 Fisher 核的特征编码方法等。通过对人体行为识别数据集上各种特征编码的评估，实验结果显示稀疏编码方法和 Fisher 核方法在不同的数据集上分别取得了最高水平的识别精度。

池化是根据视频中提取的所有特征编码计算视频的特征向量，即人体行为的表示。常用的池化方法有两种：和池化（Sum Pooling）和最大池化（Max Pooling）。和池化相当于累计所有特征编码，而最大池化相当于统计最显著的特征编码。在实际使用中，使用的池化方法取决于所选择的特征编码方法。采用稀疏编码方法时，一般使用最大池化方法统计量化至各个词条最显著的值构建特征向量；采用“硬指定”特征编码方法时一般使用和池化计算词条在视频中出现频率作为特征向量。视频池化得到特征向量后，还需要对特征向量进行归一化，常用的方法有

l_1归一化、l_2归一化和幂归一化等。

四、 分类

获得视频归一化向量以后，人体行为识别问题转化为分类问题。分类方法可以分为两类：直接分类和基于时间状态模型的分类。直接分类方法包括 K 近邻分类、随机森林、SVM 等方法。基于时间状态的分类器包括马尔科夫模型、条件随机场等。通过对人体行为识别数据集上对几种常用的分类器的识别精度研究，结果表明 SVM 的识别精度最高。因此，在应用词袋模型时，通常采用直接分类的 SVM 方法。

五、 当前研究水平和发展趋势

当前，**基于轨迹的方法（尤其是 DT 和 IDT）[2]是最高水平的人工设计特征算法的基础**。许多研究者基于 IDT 算法进一步深入研究，如对描述符使用不同的池化策略如 FV（Fish Vector）[3]和 Rank-Pooling[4]等，在 HMDB51 等数据集上取得了不错的性能。

方法	HMDB51	时间
IDT+FV[2]	57.2	2013
FV+SFV[3]	**66.79**	**2014**
VideoDarwin[4]	63.7	2015

然而，随着深度神经网络的兴起，特别是 CNN、LSTM、GRU 等深度网络在视频分类中的成功应用，其分类性能逐渐超越了基于 DT 和 IDT 的传统方法，使得这些传统方法逐渐淡出了人们的视野。值得注意的是，深度学习方法与 IDT 的组合通常能进一步提升准确度，这几年很多论文都是采用“Our method+iDT”的形式达到最高水平（state-of-the-art）。

[1] Wang H, Kläser A, Schmid C*, et al.* Dense Trajectories and Motion Boundary Descriptors for Action Recognition[J]. International Journal of Computer Vision, 2013, 103: 60-79.

[2] Wang H, Schmid C. Action Recognition with Improved Trajectories[C]. ICCV, 2013,

[3] Peng X, Zou C, Qiao Y*, et al.* Action recognition with stacked fisher vectors[C]. ECCV, 2014, 581–595.

[4] Fernando B, Gavves E, M. J O*, et al.* Modeling video evolution for action recognition[C]. CVPR, 2015, 5378–5387.