行为检测与识别方法研究

华俊豪[*]，　林上耀[†]

浙江大学信息与电子工程学系

2014 年 4 月 17 日

摘要

本文先综述了典型的几种行为检测与识别算法，然后在前人基础上做了扩展性的工作。首先复现了 Dollár 的工作；然后用 SVM 比较在不同核下的分类结果；最后提出一种简化的多动作视频的检测识别算法。其基本思路是在提取时空特征点并后建立其特征向量后，对其聚类建立词库, 再分配每个词属于某个动作类别的概率，确定测试样本中的兴趣点属于哪个词，由此确定所属类别。通过每一帧图像的兴趣点及其所属类别进行动作的分类与定位。

1 引言

人体行为检测与识别是指对人的行为模式进行分析和识别, 并用自然语言等加以描述, 这种技术包含从视频序列中抽取相关的视觉信息、用一种合适的方法进行表达, 然后解释这些视觉信息以实现识别和学习人的行为。

通常情况下，人体行为识别对应的基本工作流程是：选用各类传感器获取人体行为数据信息，并结合传感器特性及人的行为特性建立合理的行为模型，在此基础上从原始采集数据中提取出对行为类型具有较强描述能力的特征，并采用合适的方法对这些特征进行训练，进而实现对人体行为的模式识别。

行为识别的方法主要分为两种，基于可穿戴传感器和基于计算机视觉两种途径。其中基于可穿戴传感器具有无视地形遮挡，可自由活动，对环境容忍度高。但同样具有一定的不便捷性，需要携带大量传感器，而且需要定时更换电源，而且传感器获得的数据量有限，容易受到外界干扰，为后续的数据挖掘和模式识别造成了不小的困难。而基于计算机视觉的行为检测和识别是近年以来非常热门的研究领域，视觉是人类感知客观世界和获取信息的主要通道。利用成像设备采集场景中的图像序列，进而采用计算机视觉技术实现对人体目标的自动检测与跟踪及其行为识别，是当前人体行为识别研究的主要技术手段。

2 行为识别的研究背景

研究人员依据行为的复杂程度，将人体行为分为四个层次：姿态、个体行为、交互行为、群体行为。行为分析最基本的两个问题是行为的描述（表征）、行为的识别（分类）。

行为分析相关研究的发展历史、研究现状及目前存在的主要问题。行为分析的相关研究起始于 20 世纪的 70 年代,80 年代有了初步的进展,90 年代是行为分析的逐步发展阶段, 在这个时期提出了一些影响较大的研究方法。2000 年之后, 由于智能监控等方面的迫切需求, 行为分析的描述方法和识别算法以及行为理解都取得了快速而深入的发展。

行为分析有着广泛的应用背景, 如智能监控、人机交互、运动分析、运动员辅助训练、视频编码、虚拟现实等。近年来, 在这些应用的驱动之下, 行为分析已经成为图像分析、心理学、神经生理学等相关领域的研究热点。

[*]华俊豪: 实验, huajh7@gmail.com

[†]林上耀: 研究背景综述 512304584@qq.com

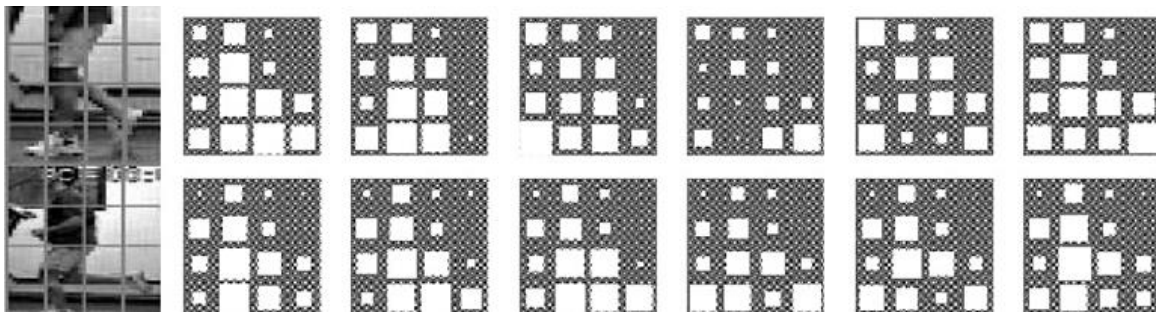

图 1: 细分网格内的光流幅度

3 行为的描述和表征

行为表示有多种方法，按照不同的标准，可以划分为不同的种类。按照行为的构成方式，主要可分为全局表示方式和局部表示方式。

3.1 局部特征描述

局部特征的行为识别算法可以分为三种: 光流时空兴趣点法, 基于网格的表示法, 和兴趣点轨迹描述法。相对于全局特征的提取, 提取的局部特征对于噪声, 光照变化以及部分遮挡等有很好的鲁棒性。并且随着高清摄像头的出现和计算机处理能力的增强, 很多细节点可以用来表达行为的特征并进行分析, 因此基于局部特征的识别方法受到了广泛的关注。

3.1.1 基于光流的方法

早在 1994 年，Polana 和 Nelson[23] 就提出将光流场信息用于行为识别，如图 (1) 所示，首先计算出连续图像帧间的光流场，然后将每个光流帧按照二维栅格形式进行细分，分别计算出每个单元格内光流场的总幅度，获得用于行为表征的高维特征向量，再采用最近邻分类器对行为分类。随后，Cutler 和 Turk 等人 [9] 将光流场聚类形成一系列运动块，并提取出这些运动块所对应的位置和运动速度等特征信息。

Efros[11] 则将光流场进行了拓展，新定义的光流场由水平分量的正、负场和垂直分量的正、负场等四个通道组成，这一拓展可以获取更为详尽的运动方向信息并增强行为识别对图像噪声的鲁棒性。Toby[15] 提出在光流场基础上建立光流场变化能量图（GFI），可降低携带外物状态对行为识别的影响。

由上述可知，基于光流场的行为表征不需要对采集图像进行背景减除运算，因此，在场景较复杂的情形下，可以回避背景分割这一难题。然而，也必须注意到，光流场的计算量比较大，不利于实现实时的行为识别；且光流法假设图像帧间的差异仅由目标的运动所致，而忽略了光照条件变化等外在因素的影响，因而，光流场的提取对各类图像噪声较为敏感。光流法对于运动物体的检测有比较好的适应性, 但是它的计算复杂度比较高, 运行时间会比较长, 但是检测效果并不尽如人意。

3.1.2 基于时空特征点 (Spatio-Temporal Feature)

时空特征点是近年来受到广泛关注的一类行为特征。时空特征点的定义并不唯一，故对应有不同的提取方法。是 Laptev 和 Lindeberg[16] 最早提出了用于提取时空特征点的方法, 他们将 2D-Harris 角点检测方法推广到了时空域，形成了 3D-Harris。但是其提取的特征点数较少，因此,Dollar[10] 等人对此进行改进并提出了一种新的提取时空特征点的方法，能够获取更多的特征点，且运算速度相比 3D-Harris 有了很大的提升, 目前应用非常广泛，图 (2) 描述了提取时空域的 cuboids。Scovanner[26] 在原有 2D-SIFT 的基础上进行拓展, 构造了 3D-SIFT 算子对特征点进行描述, 虽然计算时间相比起 Dollar 所提出的 Cuboid 较长，但是匹配效果要优于 Cuboids。

时空特征点的提取对于光照条件变化、目标序列对准偏差等因素不敏感，不需要对目标进行分割，但由于时空特征点仅对应运动变化较为剧烈的局部身体区域，而未能全面地反映出人体运动的姿态信息，因而识别效果会受到一定影响。

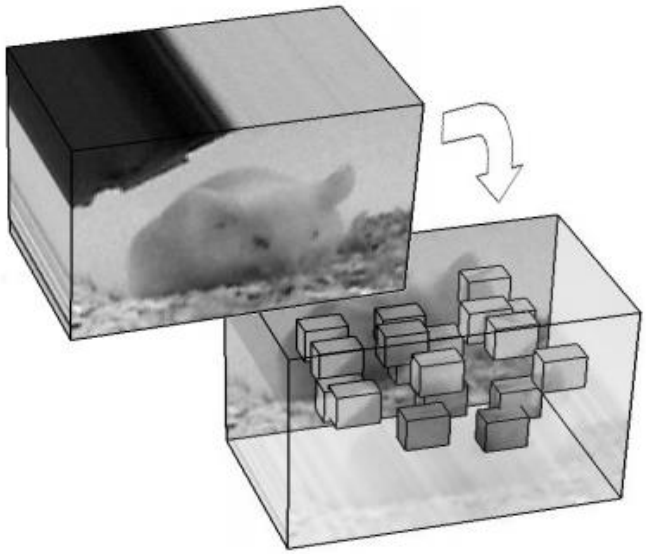

图 2: 白鼠的时空特征 cubiods

3.1.3 基于网格的方法

Nowozin 等人 [21] 在时间域上进行网格的划分为了能够克服噪声的影响, 划分的时候使得网格有重叠, 然后在时空兴趣点的周围提取特征描述子, 用 PCA 进行降维, 最终映射为 Codebook 的索引。Bregonzio[5] 等人在提取兴趣点后, 没有计算特征描述子, 计算的是特征点的数量, 在特征点中包含了不同尺度的时空网格组, 这种方法有很高的计算效率, 能够区分运动区域很大和运动区域很小的行为, 但是它没有在实质上反映行为的本质, 检测效果一般。

3.1.4 基于兴趣点轨迹的方法

时空轨迹法是把人的行为看做是在时间 -空间的一系列轨迹, 它首先用人的关节点 (比如射部, 膝部等) 来对人体进行抽象建模, 然后在运动过程中, 关节点在 3D(XYT) 或者 4D(XYZT) 的运动轨迹来可以用来表达一个行为。Rao 和 Shan [24] 跟踪人的皮肤特征得到人体在 3D-XYT 空间中运动轨迹, 然后提取轨迹曲线的峰值特征来作为特征进行匹配, 他们还证明了这些峰值特征是视觉不变的, 该方法应用到办公室中行为的识别 (例如开柜子, 捡起一个箱子等) 中,Sheikh 等 [27] 在 4D 空间中跟踪了 13 个关节点的运动轨迹，然后他们提出了一种空间映射的方法来对两个行为的轨迹进行比较, 从而做出判决。

3.2 全局特征描述

3.2.1 基于高层人体结构的方法

人的高层结构信息是指人身体结构所呈现的姿势, 与低层图像信息相比, 它可以更精细地描述人的行为。

根据提取特征过程中利用的人体模型不同, 可以将这类描述行为的算法分为三种: 基于人体点模型的方法、基于 2 维人体模型的方法和基于 3 维人体模型的方法。如图 (3)

这种描述行为的点模型方法对后来基于人体结构的行为描述算法起到了很重要的指导作用。但是由于需要自动估计关节点的位置, 所以要受限于姿势估计算法的发展。到了近几年随着姿势估计算法的快速发展, 人体点模型才被应用于行为的描述。

基于模型的表示方法能够准确地描述人的运动, 尤其是涉及到肢体的动作, 能够较为容易地解决遮挡问题, 但是其特征空间的维数很高, 在进行非线性优化的时候非常困难. 另外, 在图像分辨率低的情况下, 对模型参数进行估计也很困难.

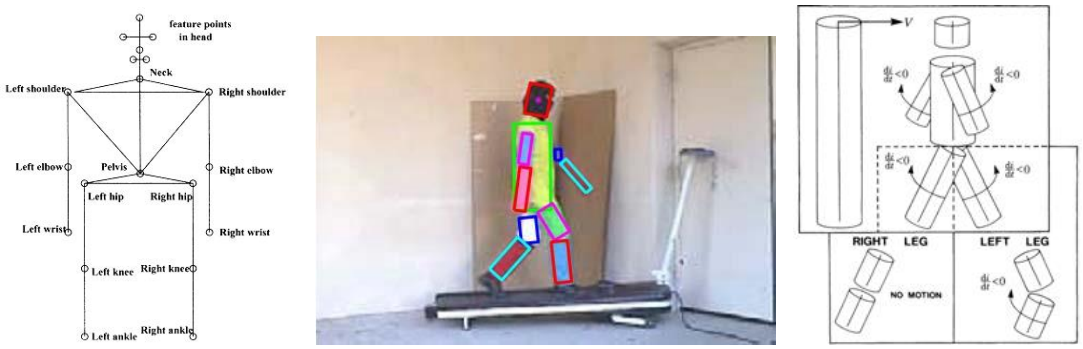

图 3: 人体结构模型: (a)stick 模型；(b) 矩形模型；(c) 三维圆柱体模型

3.2.2 基于人体外观轮廓模型的方法

基于人的轮廓来表示运动是一种很好的方法, 能够比较精确地描述人的运动细节. Kale 等人 [12] 采用基于人体轮廓的方法来解决步态识别问题, 首先将人体的轮廓提取出来, 然后求出图像中每一行的轮廓宽度, 以此宽度向量作为特征向量进行识别.Veeraraghavan 等人 [29] 采用轮廓上的标记点来分析人的运动, 对于有限的标记点集合, 通过 Kendall 形状理论来对其进行分析, 最后对人的步态进行了分析. 基于有限的标记点的好处是对运动的表示比较精确, 而且特征空间的维数较低, 计算复杂度小。

基于人体外观模型的行为表征，并不试图从采集视频中提取人体的结构，而是直接从图像中提取出目标的外观信息，如人体轮廓、剪影、各种时空模板图、方向梯度直方图等，以备进一步提取行为特征。

3.2.3 时空体积法

时空体积法的主要思想是: 首先去除背景提取运动区域, 然后把视频流的一系列帧进行累加, 形成一个时空的体积块, 并对其进行分析。在时空体积块的分析方法中, 会有一些提取局部信息的算法, 但是它总体上是对运动整体的描述, 我们把它归结为全局表示法。

Blank [3] 通过叠加已知序列的轮廓构成一个时空的体积块，然后通过泊松解获得局部时空拐点和方向特征。通过计算这些局部特征的权重向量，获得已知时间范围的全局特征。通过在不等的时间间隔内应用该方法，Achard 等人 [1] 改进了这个方法, 在不等的时间间隔内提取全局特征取得了不错的效果。

3.2.4 基于网格的全局特征表述法

如果能够把感兴趣区域转化为固定的空间域上或者是时间域的网格, 用网格中的每一个小部分来表示视频中运动人体的信息, 可以解决在轮廓描述中由于噪声, 遮挡, 尺度变化等引起的算法失效。Kellokumpu 等人 [14] 首先时间轴方向上计算局部的二值模式, 然后在空间网格中存储感兴趣区域的直方图。Lu 和 Little 等 [18] 首先用梯度直方图 (Histograms of Oriented Gradients, HOG) 提取了运动人体的边缘, 然后用 PCA 对特征描述子进行降维得出结果。另外, 光流纹理等信息以及其组合在网格表示法中也有相关的应用, 它们的组合克服了单一表达的缺陷, 使得提取的特征更加稳定。

4 行为的分类和识别

行为识别算法也可分为两类: 一类是基于模板匹配的算法, 一类是基于状态空间的算法行为基于模板匹配的算法计算量较少, 但是对行为时间间隔敏感; 基于状态空间的算法可以避免行为时间间隔建模的问题, 但是模型训练复杂。

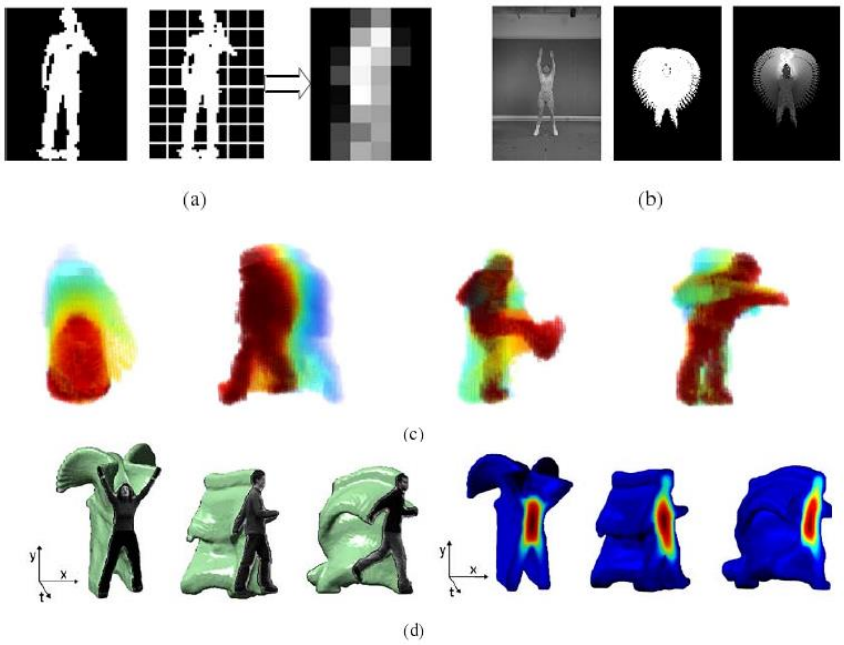

图 4: 人体结构模型: 人体外观模型：(a) 轮廓网格像素统计；(b) 运动能量图 (左) 和历史能量图 (右)；(c) 运动历史卷；(d) 时空轮廓图

4.1 基于模板的方法

基于模板的方法是将运动图像序列转化成一个或者一组静态的模板, 通过将待识别样本的模板和已知的模板进行匹配而获得识别的结果. 基本的模板匹配方法是将待识别模板和已知的模板标本进行直接匹配, 取距离最小的已知模板所属的类别作为识别结果. Bobick 和 Davis [4] 将图像序列转化成运动能量图像 (MEI) 和运动历史图像 (MHI) , 采用马氏距离 (Mahalanobis distance) 来度量模板之间的相似性. 其中 MEI 反映了运动所覆盖的范围及其强度, 而 MHI 在一定程度上反映了运动在时间上的变化. 该方法计算量小, 但是鲁棒性不够好, 尤其对时间间隔的变化比较敏感。

4.2 基于状态空间的行为识别算法

状态空间法也是传统行为识别常用的研究方法之一, 它的基本思想是把行为看作是一系列姿势 (状态) 有顺序的合集, 用动态的转移概率来把状态串联起来, 行为的执行就相当于是姿势集合的一次遍历, 最后用各个姿势联合概率的最大值用于行为的决策。这种方法对于时空变化不大的行为序列有着较好的检测性能, 常用的模型是隐马尔科夫模型 (Hidden Markov Models), 例如 Yamato [30] 和 Nguyen 等都分别用 HMM 或者是其改进模型进行了行为识别，但是这种方法计算量大, 不能检测突发状况。

2003 年,Luo [19] 将 DBN(动态贝叶斯网络) 引入行为识别，并对 HMM 和 DBN 进行了比较在一个时间切片上,HMM 只能含有一个隐含节点和一个观测节点; 而 DBN 在一个时间切片上是一个贝叶斯网络, 可以包含多个有因果关系的节点; HMM 在一个时刻需要将所有的特征压缩到一个节点中, 那么所需要的训练样本将是巨大的 (相当于联合概率密度函数); 而 DBN 用多个节点描述, 即用条件概率来形成联合概率, 训练相对要简单; 但是 DBN 的设计要比 HMM 复杂得多。更多的研究人员把重点放到了姿势表达上, 首先, 建立人体姿势的 2D 或者 3D 模型库, 对行为中人体的姿势, 角度以及周围的背景情况进行估计和建模, 最后形成行为描述的自然语言文本, 该方法也处于研究阶段。

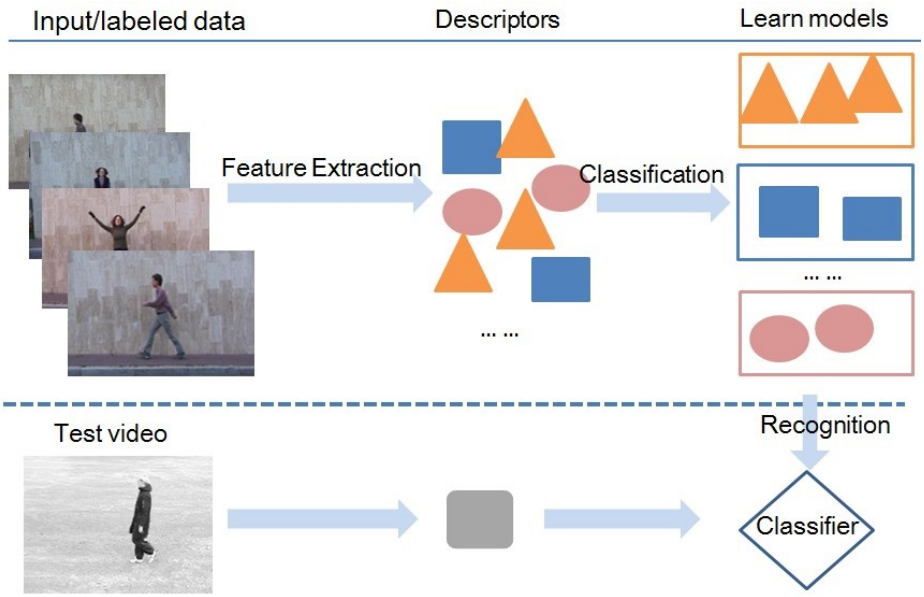

图 5: 动作检测与识别的基本框架

5 动作检测与识别实验

5.1 概要

行为检测与识别可通过多种方法，本文是基于时空兴趣点的。此处不对文献进行综述，仅讨论本文所涉及到的相关论文。Dollár[10] 通过提取时空特征点，进而形成动作的特征描述子，通过对特征描述子的学习与分类，对视频中行为或动作进行检测与识别。然而，该方法只能处理单动作的视频序列。Niebles [20] 提出一个无监督的视频动作学习方法，其同样采用上诉方法提取时空特征点，但并不建立单个动作的特征描述子，而是对兴趣点聚类，构成类似于文本分析中的词包模型 (bag of words)。每一个类对于一个词 (word), 每一个动作类型对应一个主题 (topic)，每一个视频序列就是一个文档 (document)。用图模型表示视频的生成模型，主要包括概率潜在语义分析 (pLSA) 和潜在狄利克雷分配 (LDA)。对 pLSA 或 LDA 概率推断，得到每个 word 在不同 document 里属于某个 topic 的概率。采用这种无监督学习算法能够对视频中的多个动作进行分类和定位。

本文的工作正在建立在上述 [10, 20] 两篇论文的基础上。首先复现了 Dollár 的工作；然后用自己写的 LSSVM[28] 与林智仁的 libsvm[7] 在不同的核下跑识别结果；最后，借鉴 [20] 的思想提出一种简化的多动作视频的检测识别算法。其基本思路是在提取时空特征点并后建立其特征向量后，对其聚类建立词库 (codebook), 再分配每个词属于某个动作类别的概率，确定测试样本中的兴趣点属于哪个词，由此确定所属类别。通过每一帧图像的兴趣点及其所属类别进行动作的分类与定位。

5.2 算法描述

对一堆输入的训练样本先检测并提取特征，再将特征转化为动作描述子，用分类器对动作描述子学习得到相应模型。对于测试数据，采用与训练样本相同的方法提取特征，包括相同的参数设置于基空间，得到动作描述子后，用分类器识别分类。如图 (5) 所示。

5.2.1 特征检测与提取

图像兴趣点检测最广为使用方法之一是角点检测，比如 Harris 角点，直观讲就是要求在水平和竖直方向上变化都比较大。通过对高斯平滑的图像 $L(x,y,\sigma) = I(x,y) * g(x,y,\sigma)$ 求一阶梯度，由协方差矩阵的特征值确定每个点的响应强度，从而提取出兴趣点。另一种常用方法是用 Laplacian of Gaussian (LoG) 构造响应函数，比如 Lowe [17] $D = L(x,y,;k\sigma) - L(x,y;\sigma)$. 然而，这些方法都是在空间维度上的 (因为处理对象是图像)，那么针对一个视频序列 $I(x,y,t)$，就需要将其扩展到时空维上。

图 6: 一个 walk 视频的兴趣点可视化 ($\sigma = 1.5, \tau = 1.5$)

按 [10]，响应函数定义为

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

其中 $g(x, y; \sigma)$ 为作用在空间的二维高斯滤波核，h_{ev} 和 h_{od} 为作用在时间上的一维 Gabor 滤波，

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$$
$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$$

其中采用 $\omega = 4/\tau$, σ 和 τ 分别表示空间和时间上的尺度。

然后用非极大值抑制 (Non-max suppression) 方法搜索局部极大值，即判断该点是否为其时空窗 (大小为 $(x, y, t) = (2\lceil 3\sigma \rceil + 1, 2\lceil 3\sigma \rceil + 1, 2\lceil 3\tau \rceil + 1)$) 内的最大值且满足一定的阈值条件。按照这种方法可以检测到相当多的极值点，但太多特征点反而会使结果变坏，因此控制兴趣点个数是很有必要的，可以只取最大或最小程度最大的前几百个。如图 (6) 所示为每一帧中的可视化的兴趣点。

通过以上步骤得到兴趣点，将包含兴趣点的时空窗定义为 cuboids. 如图 (7) 所示为每一帧中的可视化的 cuboids。对于一个长方体（cuboids), 数据量是比较大的，且直接定义相似度函数来用于比较是不合适的。因此更进一步地，需要创建一个 cuboids 描述子。转换方法是计算 cuboids 中每个点的亮度梯度 (brightness gradient), 为了得到更丰富的特征，应先在 cuboid 上加以不同尺度的高斯滤波 (实验用了 2 不同尺度)。这里得到的依然是一个三维向量，要转化为一维特征向量，可直接将其拉直，或者采用局部直方图（local histogram)，这里采用前者。

得到的特征向量维度是相当高的 ($x \times y \times t \times 3 \times 2$)，采用 PCA 降维。显然，需要保持每个 cuboid 主成分的一致性，否则无法比较。因此 PCA 降维是这样做的：所有训练样本提取出 cuboids 后，随机在其中取一定数目的 cuboids 特征向量，然后 PCA 降维，取特征值最大的前 K 个主成分, 如图 (8)。实际上就是提取了 k 个基 basis)。然后将每个 cuboids 投影到这些基上，构成了 $K \times 1$ 的 cuboid 特征描述子。这种描述子可看成了 PCA-SIFT 描述子 [13] 的推广。

5.2.2 特征描述

Cuboid Prototypes 虽然每个动作可能千差万别，但可发现许多兴趣点是具有相似性的。由此可见，虽然可能的 cuboid 有很多，但不同的 cuboid 类型数却不多。于是可通过对训练样本的 cuboid 描述子聚类得到 cuboid prototypes 字典库。这里聚类采用简单的 k-means 算法，在实验中聚类个数设为 50 个。

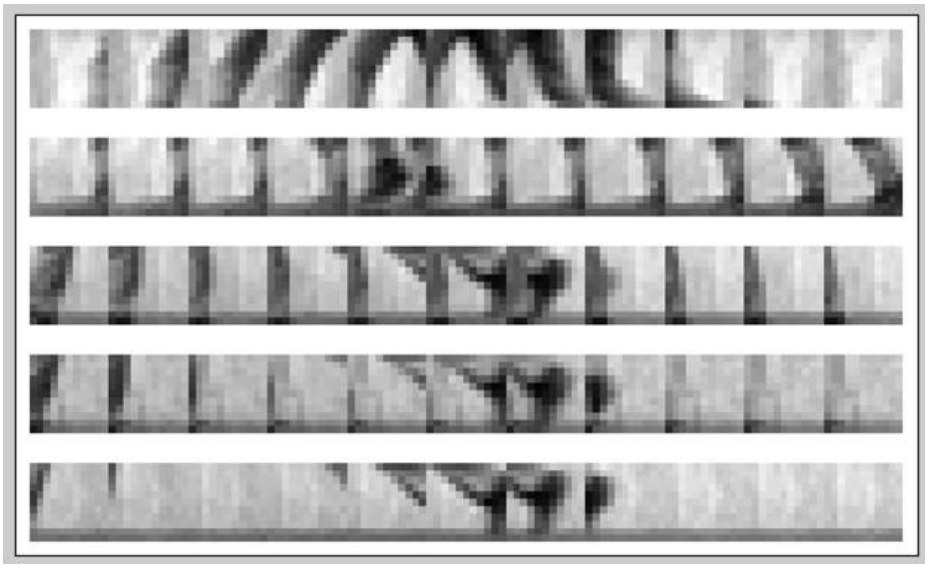

图 7: 一个 walk 视频中极值程度最强的 5 个 cuboids ($\sigma = 1.5, \tau = 1.5$)

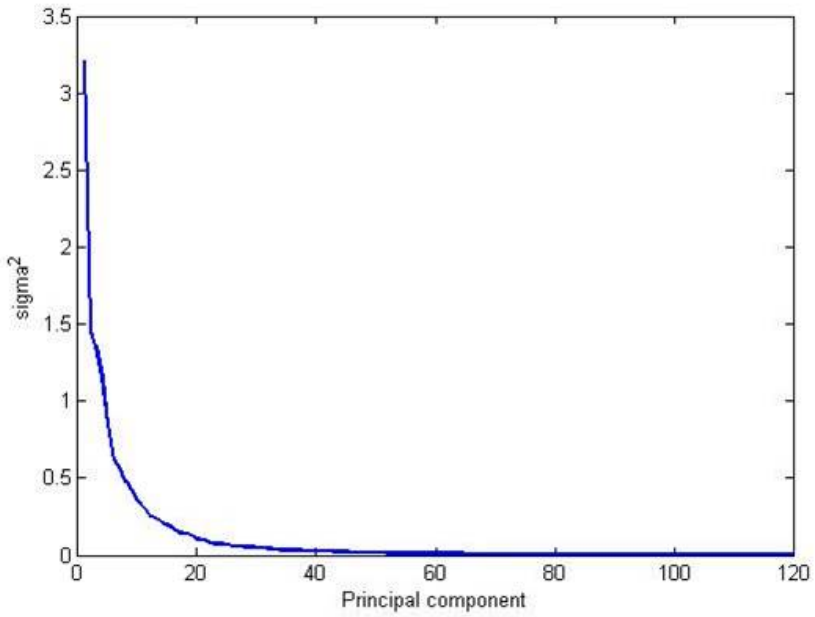

图 8: Weizmann 人类动作数据集 (见下文) 降维结果 ($\sigma = 1.2, \tau = 1.2$)

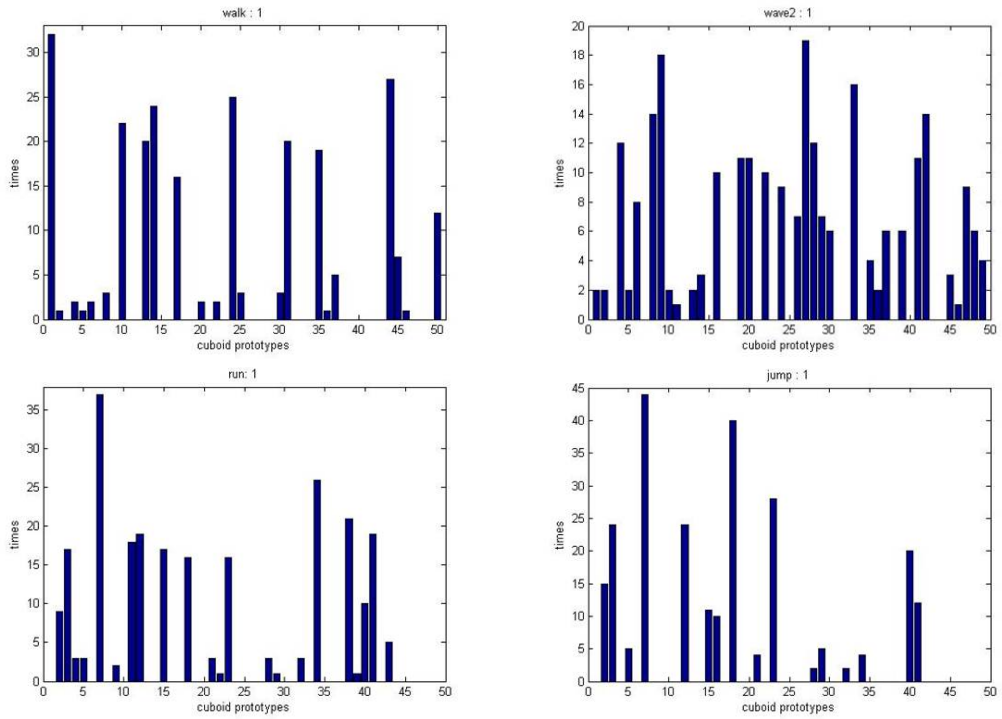

图 9: 左上：walk；左下：Run; 右上:wave2; 右下：jump.(Weizmann 人类动作数据集)

Behavior Descriptor 有了 Cuboid 类型库，再提取出动作描述子作为一个视频序列的特征向量。这个采用的方法是用 cuboid 类型的直方图表示。一个视频的兴趣点对于的 cuboids 属于某个 cuboid prototype 的个数构成的向量。图 (9) 举了 4 个例子。直方图的比较通常用卡方距离 (χ^2) 来表示相似度，定义为，

$$d_{\chi^2}(x,y) = \sum_i \frac{(x_i - y_i)^2}{2(x_i + y_i)} \tag{2}$$

5.2.3 动作分类算法

将视频数据转为相应的动作描述子后，便可以通过对训练样本的描述子得到分类器，然而用分类器对测试数据进行分类。可以采用最简单的分类器为 KNN(K-nearest neighbors)，也可以采用相对复杂一点的支持向量机 (SVM) 分类器。

KNN KNN 是一种无参分类器，对于一个测试样本，在训练样本的特征空间中搜索与之最近的 k 个样本点，这 k 个样本中属于某一个最多，那么就认为该测试数据属于哪一类。实验采用 1NN. 这里对距离的度量便采用的是卡方距离。

SVM 就两类分类问题而言，简单得讲，支持向量机就是用两个 $N-1$ 维的超平面分割 N 维的线性空间，使得在两个超平面两侧的半闭空间分别属于两类。SVM 优化求解就是使两个分隔面之间的距离最大。当然，引入松弛变量后，上一句就不那么严格了。采用 Lagrange 乘子法将其转化为求解它的对偶问题，该对偶问题是一个二次规划问题，直接求解比较复杂，可以采用 Chunking 简化 (Vanik,1982;Burges,1998 [6])，但更好更广为使用的 SMO 算法 (Sequential minimal optimization,Platt,1999 [22])。此外，还有一种 Least square SVM [28], 将目标函数中的松弛变量改成二次，使得最终的问题变成了一个线性方程求解，只需要几步矩阵运算。由于其为 2 范式, 最后的 Lagrange 乘子不再稀疏，使得所有的训练样本都为支持向量。当然, 作为改进的 Sparse LSSVM。

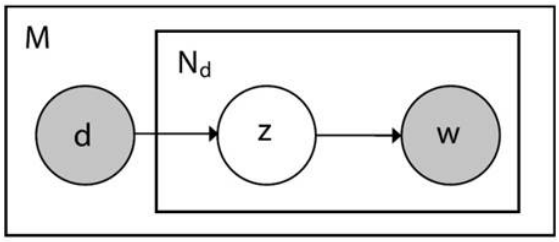

图 10: pLSA 图模型

```
% how many action categories ?
categ_idx = [];
k = 0;
for j=1:ntype
    if type_prob(j)>0.5 ||...
        (type_prob(j)>0.4 && type_cnts(j) > 5) ||...
        (type_prob(j)>0.3 && type_cnts(j) > 14) ||...
        (type_prob(j)>0.2 && type_cnts(j) > 13) ||...
        (type_prob(j)>0.1 && type_cnts(j) > 15)
        categ_idx = [categ_idx,j];
        k = k + 1;
    end
end
```

图 11: 部分代码截图：判断一帧图像有多个类

此外，考虑到原特征空间非线性，使用 kernel 方法，将低维的非线性空间映射到高维的 (线性) 特征空间。这里对直方图的比较，[8] 提出采用所谓的卡方核 (chi-squared kernel) 相对于高斯 RBF 具有更好的效果。对于一个多分类器，可以采用多种策略，本实验用最基本最简单的 one-versus-the-rest 方法。关于 SVM 的详细介绍见 PRML 的第 6,7 章节 [2]。简单起见，实验中仅实现了最基本的 LSSVM 算法以及使用 SMO 优化的 libsvm[7] 作为分类器。

5.2.4 多动作识别框架:**Voting**

以上算法框架，将整一个视频序列序列做个分类对象，因此单视频中包含多个动作时，便无法处理。针对此类情况，我们根据 [20] 的词包模型，提出简化版的多动作识别框架。

训练样本依然用单动作的视频，视频的时空特征提取与以上完全一致，但不再通过聚类得到数量较少的 cuboid prototypes, 而是聚类得多数目较多的时空词 (spatial-temporal words)。其实除了聚类数量级不同外 (比如前者 50，后者 1000)，聚类方法一模一样, 得到的类中心点，称之为 CodeBook。训练样本的每一个 cuboid 属于某一个 word，而 cuboid 所在的视频动作的标签（即属于哪个动作）是已知的。这样，通过每个 word 中的 cuboid 标签比例，计算 word 属于某个 topic(动作) 的概率 $P(z_k|w_i)$, 取概率最大的 topic 为 word 的标签。该模型类似于主题模型，如 pLSA(如图 10) 和 LDA.

对一个具有多动作的测试视频，将检测得到的兴趣点分配给某个 word，从而分配其相应的 topic. 然后对每一帧图像进一步动作分类和定位。首先确定该帧图像中有多少比较显著的动作的个数，判断方法是 topic 占用比例，以及对于的特征点个数，其值是需要微调的，一种可用的方案如图 (11)。然后在对兴趣点 k-means 聚类，得到各类所在的时空位置，而该类属于哪里 topic, 由盖类中的兴趣点投票确定。这种方法这本文中取名为：**Voting**. 多动作检测与视频框架如图 (12) 所示。

5.3 实验结果

5.3.1 实验一:**Toy SVM** 分类

第一个实验做了简单的 SVM 分类实验，测试了多类的 LSSVM 分类结果。如图 (13) 所示生成数据为各个圆环上的点，加圆圈的为测试数据. 使用径向基核函数，最终三类分类精度为 0.98, 四类分类精度为 0.94.

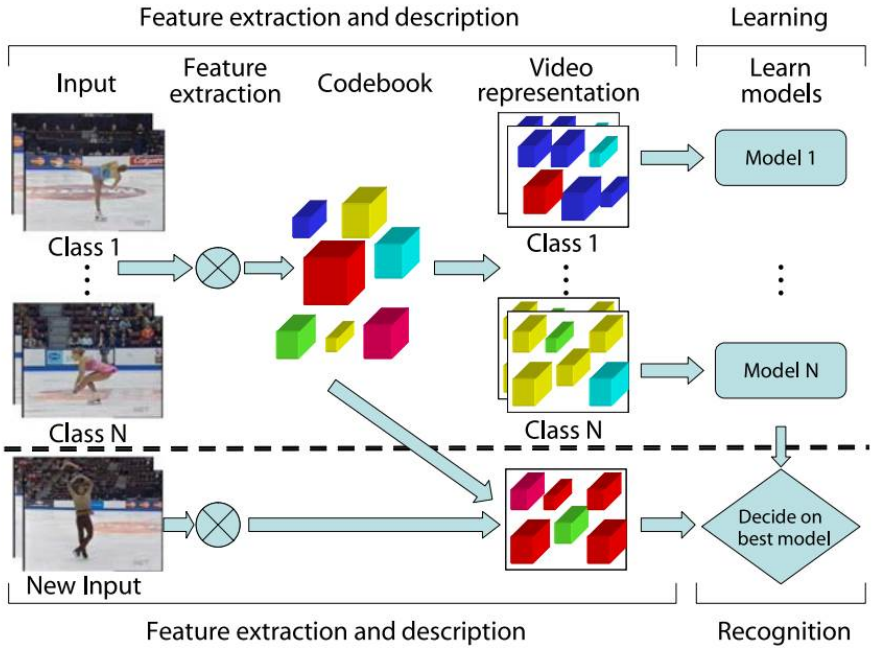

图 12: 多动作检测与识别的基本框架，图来自于 [20]

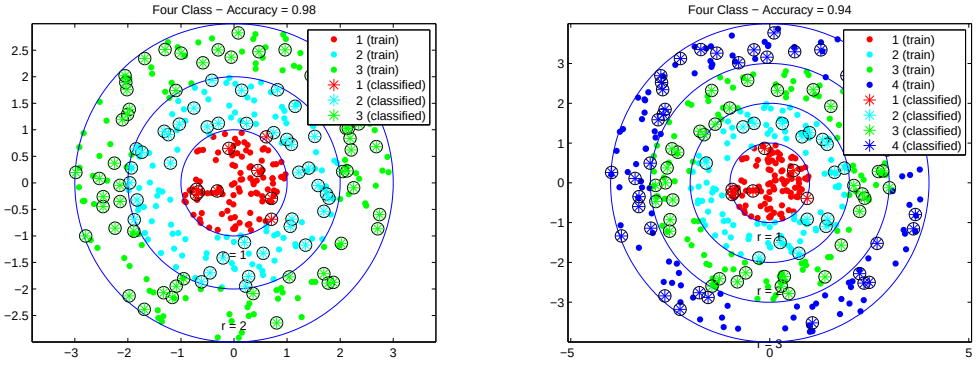

图 13: 左：3-class; 右:4-class

图 14: weizmann 视频数据例子

5.3.2 实验二：**Weizmann human action dataset**

第二个实验使用 Weizmann 人体动作数据集 [3]，总过包含 90 个低分辨率 (180×144,50fps) 视频，含有 10 种不同的动作（bend, jack, jump, pjump, run, side, skip, walk, wave1,wave2, 如图 (14)），分别由 9 个人完成。

用以上所述的方法检测和提取时空特征点，其参数设定为 $\sigma = 1.2, \tau = 1.2$, 兴趣点个数最大值为 200,PCA 降到 100 维。由于数据样本数不多，采用 Leave-one-out 交叉验证，将 90 个数据份分成 9 组，每组 9 个数据，每次取 9 组训练另 1 组测试。重复计算 5 遍，取平均值。这个包含三种分类方法：

- 1NN 用卡方距离度量特征向量的相似度，最终分类准确率为:**0.7667** 如图 (15). 抽取了各个动作的正确分类结果，见图 (18).
- LSSVM 用 LSSVM 且分别采用线性核 (linear)，多项式核 (Polynomial)，径向基 (RBF) 和卡方核 (Chi-squared) 分类. 结果表明，采用径向基核（高斯或 χ^2) 的方法并没有得到理想的效果, 见图 (16)。在实验过程中，由于首先考虑的是 χ^2 核，且一直达不到好的效果，本以为是由于数据量不够的原因。因此又做了具有更大数据量的数据集实验，即实验三。但后面会发现，并非是由于数据量的原因。
- Voting 采用投票的方法，得到的总精度为 **0.7333**, 主要的错误是将 wave1 识别为 wave2, 这是因为两个具有局部一致性。这也暴露了该方法的缺点，没有利用时空特征的相对位置信息。

5.3.3 实验三：**KTH** 数据集

上文已经提到 weizmann 数据量比较小，可能导致 SVM 训练不充分。这里使用了 KTH 数据集 [25], 包含 6 类人体动作 (walking, jogging, running, boxing, handwaving 和 handclapping), 由 25 个不同的人在 4 个不同场景下 (室内, 室外, 尺度变化)，总共 598 个 160×140 的视频序列（两个视频丢失），如图 (19)。

用以上所述的方法检测和提取时空特征点，其参数设定为 $\sigma = 1.5, \tau = 1.5$, 兴趣点个数最大值为 300,PCA 降到 200 维。由于数据样本数不多，采用 Leave-one-out 交叉验证，将 598 个数据份分成 25 组，每组 24 个数据，每次取 24 组训练另 1 组测试。重复计算 5 遍，取平均值, 用 1NN,LSSVM 和 Voting 共 4 种方法得到分类结果见图 (20)(21)(22)。由于数据集的数据量更多，且环境比较简单，得到的分类精度都有所提高，并且在 SVM 的使用中，基于 RBF 的高斯核与卡方核都没有得到有的效果，实际上，我们用开源的 libsvm 同时训练一遍结果与用我们的 LSSVM 一样，因此可排除代码错误，至于为什么没有好的结果，目前还找到不好的解释。另外，采用 Voting 方法，由于各个动作之间没有了局部一致性，因此达到最佳分类精度:**0.9047**。

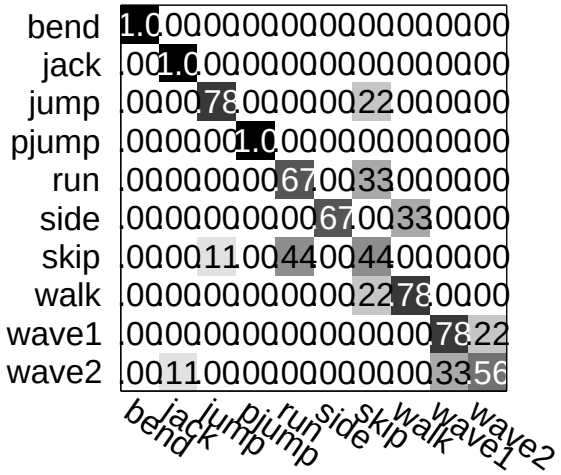

图 15: weizmann dataset by 1NN, 总分割精度为 0.7667

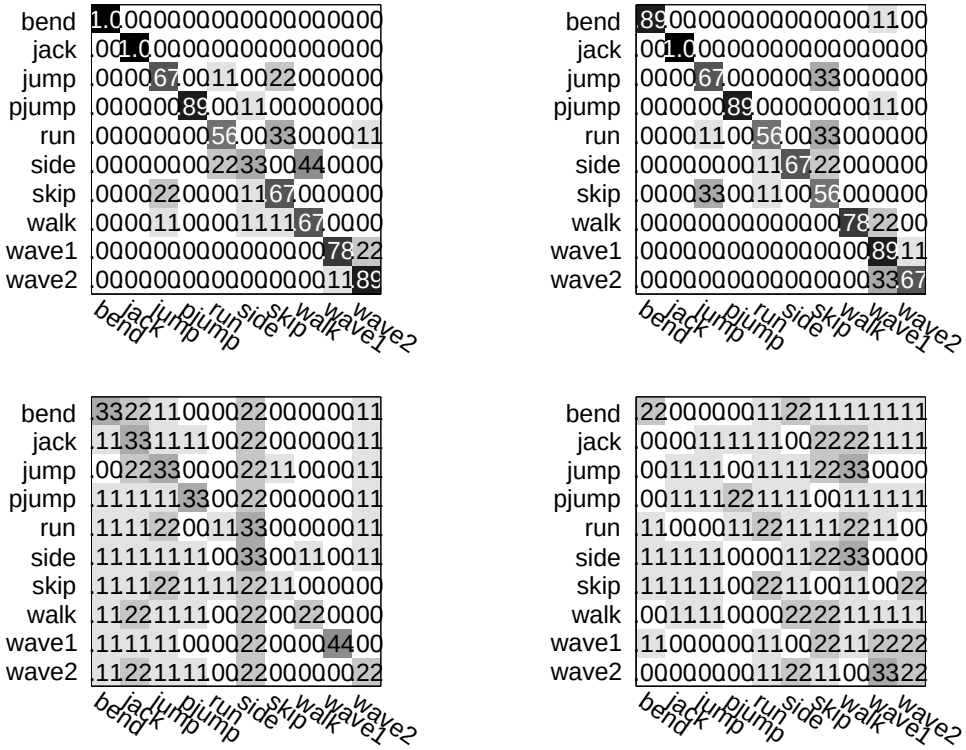

图 16: weizmann dataset by SVM. 左上：linear, 分割精度 0.7444；左下：rbf,0.2778; 右上:Polynomial,0.7556; 右下：Chi-squared,0.1444

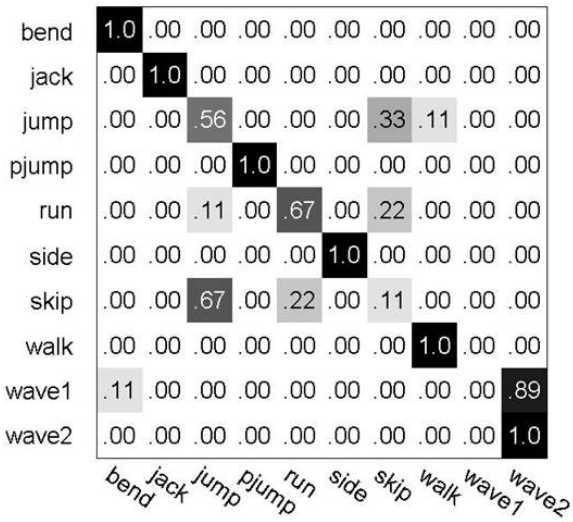

图 17: weizmann dataset by Voting, 总分割精度为 0.7333

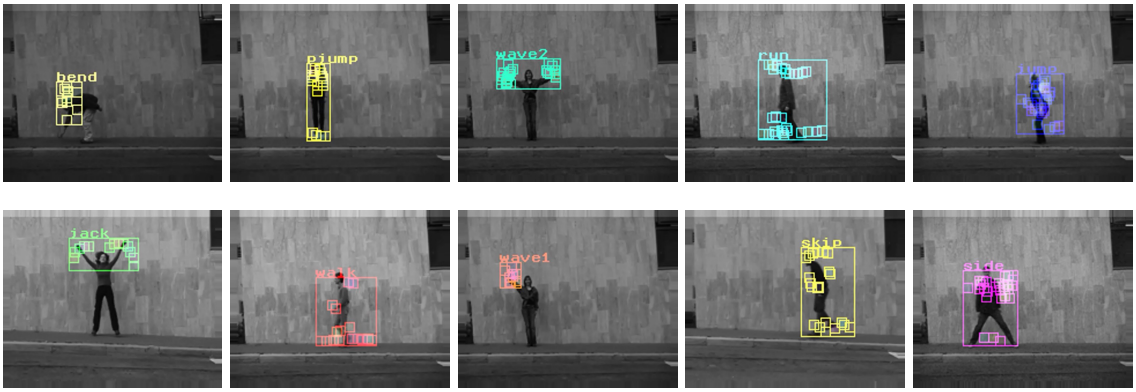

图 18: 分类结果截图 by 1NN

图 19: KTH 人体动作数据集

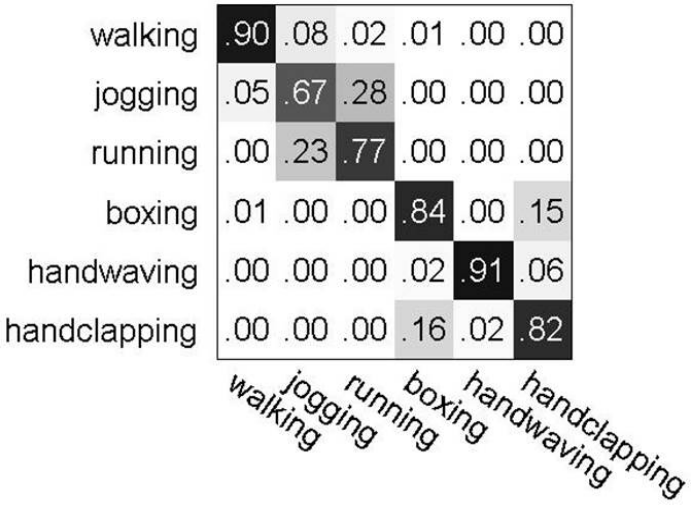

图 20: KTH dataset by 1NN, 总分割精度为 0.8179

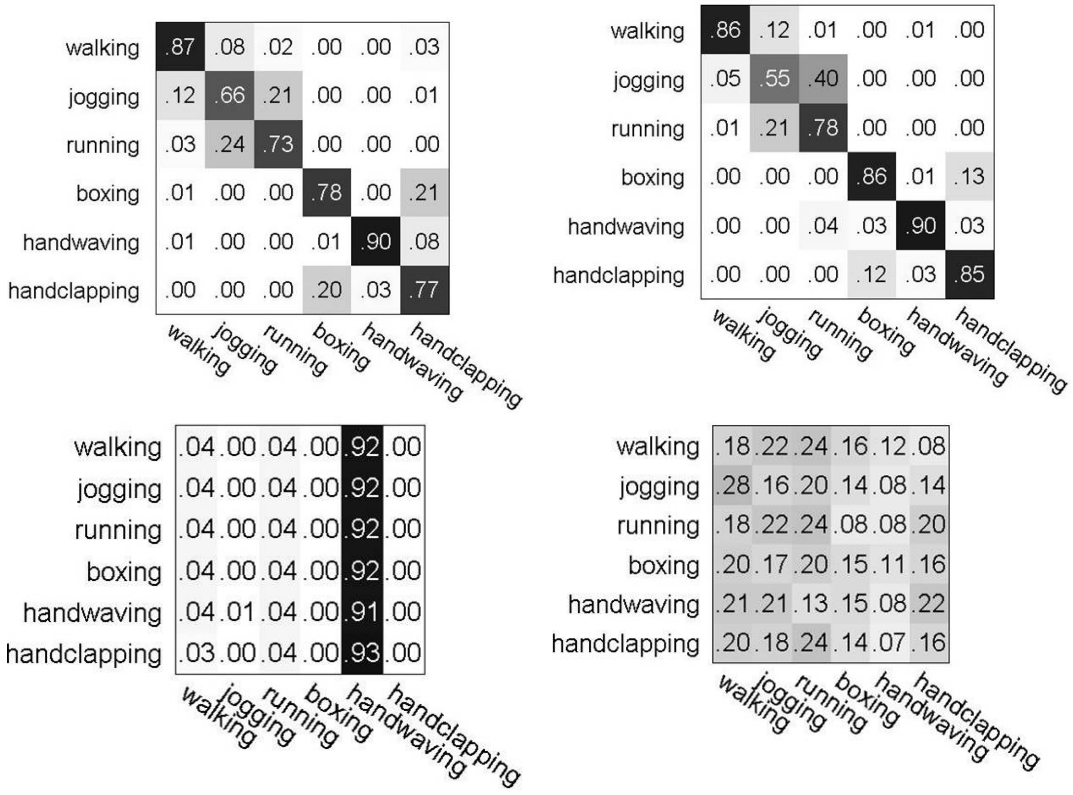

图 21: KTH dataset by LSSVM. 左上：linear, 分割精度 0.7844；左下：rbf,0.1656; 右上:Polynomial,0.7996; 右下：Chi-squared,0.1624

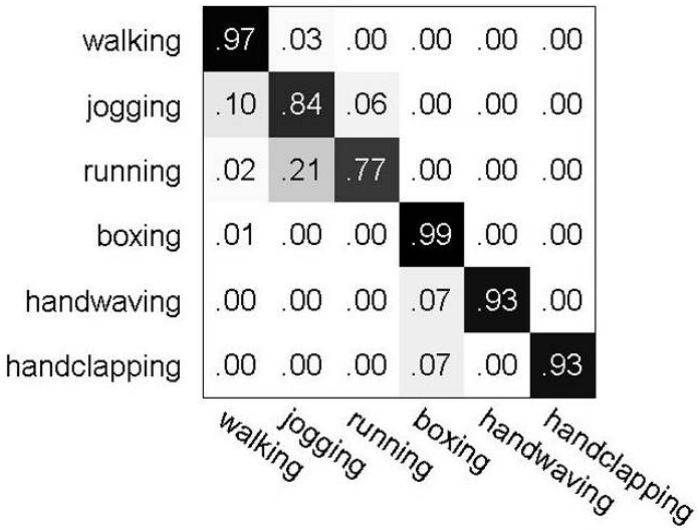

图 22: KTH dataset by Voting, 总分割精度为 0.9047

图 23: 自制的多动作视频截图

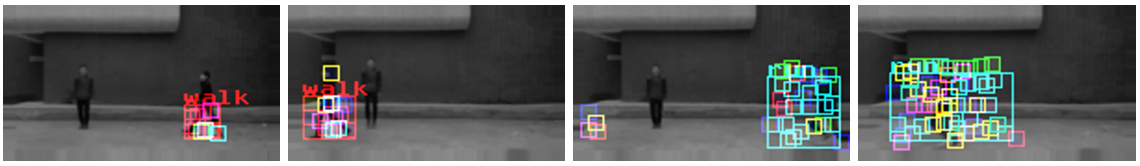

图 25: 一个视频在 weizmann 数据集上测试结果截图

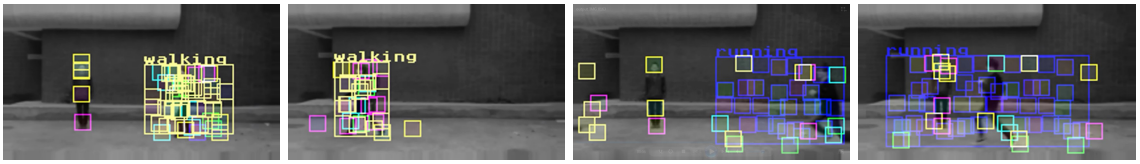

图 26: 一个视频在 KTH 数据集上测试结果截图

5.3.4 实验四：多动作数据测试

最后，我们测试了多动作视频的检测与识别，由于没有合适的数据集提供测试，我们自己拍了几段视频，如图 (23). 用 weizmann 训练的识别结果部分截图见图 (25)，用 KTH 训练的识别结果部分截图见图 (26)。采用不同的训练集需要采用其相对于的参数配置，比如如果用 weizmann 数据集则在提取测试数据的兴趣点时，设置 $\sigma = 1.2, \tau = 1.2, kpca = 100$ 而用 KTH 数据集才设置 $\sigma = 1.5, \tau = 1.5, kpca = 200$。否则其 cuboid 描述子无法对齐。也正是因为这个原因，使得同一个视频数据提取出的兴趣点有差异，显然由于拍摄的原因，具有稍大的尺度 (σ, τ) 更好的提取视频中原地站立且离摄像机较远的人的特征点。由于视频中人物的动作是根据 weizmann 数据集定义的动作制作的，因此在用 KTH 训练分类的结果中的部分分类错误也在情理之中。更多信息详见输出视频。

6 总结

所有的测试输出结果实际为一段视频，文档不可能全部展现，详见相关文件夹的输出视频。

作为非计算机视觉为研究对象的研究生，通过对” 动作检测与识别 “这个专题内容的学习, 研究与程序实现，不仅很好的窥测到该领域的惊鸿掠影，而且在程序实现的级别上对图像与视频处理中的各个细节有了一个基本认识和基本技能，而不是停留在对一个方程或一个算法流程的宏观理解上。实验代码将会放在我的 github 主页 (**https://github.com/huajh**) 上，如有疑问或发现一些错误，可电子邮箱联系我。

References

[1] Catherine Achard, Xingtai Qu, Arash Mokhber, and Maurice Milgram. A novel approach for recognition of human actions with semi-global features. *Machine Vision and Applications*, 19(1):27–34, 2008.

[2] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[3] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE, 2005.

[4] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001.

[5] Matteo Bregonzio, Shaogang Gong, and Tao Xiang. Recognising action as clouds of space-time interest points. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1948–1955. IEEE, 2009.

[6] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[7] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[8] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064, 1999.

[9] Ross Cutler and Matthew Turk. View-based interpretation of real-time optical flow for gesture recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 416–416. IEEE Computer Society, 1998.

[10] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

[11] Alexei A Efros, Alexander C Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 726–733. IEEE, 2003.

[12] Amit Kale, AN Rajagopalan, Naresh Cuntoor, and Volker Kruger. Gait-based recognition of humans using continuous hmms. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 336–341. IEEE, 2002.

[13] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506. IEEE, 2004.

[14] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Human activity recognition using a dynamic texture based method. In *BMVC*, pages 1–10, 2008.

[15] Toby HW Lam, King Hong Cheung, and James NK Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern recognition*, 44(4):973–987, 2011.

[16] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[17] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[18] Wei-Lwun Lu and James J Little. Simultaneous tracking and action recognition using the pca-hog descriptor. In *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, pages 6–6. IEEE, 2006.

[19] Ying Luo, Tzong-Der Wu, and Jenq-Neng Hwang. Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks. *Computer Vision and Image Understanding*, 92(2):196–216, 2003.

[20] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, 79(3):299–318, 2008.

[21] Sebastian Nowozin, Gökhan Bakir, and Koji Tsuda. Discriminative subsequence mining for action classification. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[22] John Platt et al. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

[23] Ramprasad Polana and Randal Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, pages 77–82. IEEE, 1994.

[24] Cen Rao and Mubarak Shah. View-invariance in action recognition. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–316. IEEE, 2001.

[25] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[26] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, pages 357–360. ACM, 2007.

[27] Yaser Sheikh, Mumtaz Sheikh, and Mubarak Shah. Exploring the space of a human action. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 144–149. IEEE, 2005.

[28] Johan AK Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, Joos Vandewalle, JAK Suykens, and T Van Gestel. *Least squares support vector machines*, volume 4. World Scientific, 2002.

[29] Ashok Veeraraghavan, Amit K Roy-Chowdhury, and Rama Chellappa. Matching shape sequences in video with applications in human movement analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12):1896–1909, 2005.

[30] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.

[31] 何卫华. 人体行为识别关键技术研究. PhD thesis, 重庆大学, 2012.

[32] 孔祥斌. 视频中的人体行为识别算法研究. Master's thesis, 电子科技大学, 2012.

[33] 杜友田, 陈峰, 徐文立, and 李永彬. 基于视觉的人的运动识别综述. 电子学报, 35(1):84–90, 2007.

[34] 王生进谷军霞, 丁晓青. 行为分析算法综述. 中国图象图形学报, (3):377–387, 2009.

[35] 赵海勇. 基于视频流的运动人体行为识别研究. PhD thesis, 西安电子科技大学, 2011.