参赛队员姓名：刘知宜

中学：北京大学附属中学

省份：北京市

国家/地区：中国

指导教师姓名：肖然、尹建芹

论文题目：基于计算机视觉的
重复性动作计数研究

基于计算机视觉的重复性动作计数研究

刘知宜

摘要

生活中对各种重复性动作计数的需求广泛，而目前较多采用各种传感器进行信息获取和分析，涉及到的传感器种类和数量多，设备复杂且通用性不足。本文提出一种基于计算机视觉进行视频中重复性动作计数的方法，并尝试应用于中考体育测试项目的计数。该方法首先基于 RGB 图和光流图使用深度卷积神经元网络（Convolutional Neural Networks，CNN）进行视频特征提取；然后使用主成分分析（Principal Component Analysis，PCA）对特征序列进行分析，提取第一维主成分，发现其很好地体现了重复动作的运动规律；进一步使用分段阈值滤波的傅里叶变换和傅里叶反变换去除噪声信息，获取明显的运动特性波形；最终通过峰值检测统计重复运动次数。在两个国际公开的测试数据集上测试，经与国际上已有的三种最优方法比较，此方法在两个数据集上均接近最优，且标准误差均达到了最好的效果，证明了该方法的有效性。另外，也将此方法应用于中考体育测试项目（仰卧起坐、引体向上、排球垫球）中重复性动作的计数，结果表明该方法具有较强的应用可行性。

关键词：计算机视觉，重复动作计数，深度卷积神经元网络，主成分分析，滤波去噪

Abstract

In daily life, there is a wide demand for various repetitive motion counts. Currently, various sensors are mostly used for information acquisition and analysis, with too many types and quantities of sensors, too complex equipment and insufficient universality. This paper presents a method of counting repetitive motion in video based on computer vision, and tries to apply it to the counting of sports test items of high school entrance examination. In this method, video features are extracted by using convolutional neural network (CNN) based on RGB and optical flow diagram. Then, principal component analysis (PCA) is used to analyze the feature sequence and to extract the principal component of the first dimension that reflects the motion rules of repeated actions. The Fourier transform and inverse Fourier transform with piecewise threshold filtering are further used to remove the noise information, and then the obvious motion characteristic waveform is obtained. Finally, the repeated motion times are counted by peak detection. The validity of this method was verified by comparing with three existing state-of-the-art methods on two internationally disclosed test data sets, which proved that the presented method was close to the state-of-the-art method in both data sets and its standard error was the smallest. This method was applied to the counting of repetitive movements in the sports test items (sit-up, pull-up, volleyball mat) in the high school entrance examination, which showed that this method has strong feasibility of application.

Key words: computer vision, repetitive motion counting, deep convolutional neural network, principal component analysis, filtering and de-noising

目录

一、引言

视觉重复在我们周围的世界中无处不在，划船时反复划桨、鸟儿振动翅膀、举哑铃健身，在自然和城市环境中的交通模式、闪烁的灯光和风中树叶摇摆。这种节奏和重复性被用来做速度估计、运动估计和做触发信号[1]。在计算机视觉中，理解和分析视频中的重复性非常重要，因为它可以服务于动作分类[2][3]、动作定位[4][5]、人体运动分析[6][7]、三维重建[8]和摄像机标定[9]。

在中学的学习生活中，同样存在对重复性动作的检测和计数的需求。比如北京市学生参加中考体育测试中的仰卧起坐、排球垫球、引体向上等考试科目，都是需要对这些重复性动作进行计数。平时体育课上都是同学或老师进行人工计数，较易存在人为误差，而中考体育测试中采用了一系列的专用测试设备。对于仰卧起坐和引体向上等项目普遍采用放置在特定位置的红外传感器检测考生身体某个部位达到要求位置，从而触发计数；对于排球垫球的考试则是通过特定位置放置的传感器检测排球到达要求位置而触发计数的方法。每个考试项目需要的传感器种类不同、数量不同，摆放位置不同，见图 1.1。而且，仰卧起坐考试还要根据考试学生的身高调整传感器位置才能满足不同学生的测试需求。计数检测仪器的测量传感器数量多、测量原理多样[10][11]、调整操作麻烦，而且各考试项目使用的仪器不通用，带来了很大的资源浪费。而且平时的体育课或课外锻炼中也没有条件使用这些专用计数检测仪器，应用普及性不足。

（1）仰卧起坐测试仪；　（2）排球垫球测试仪；　（3）引体向上测试仪

图 1.1 北京市中考体育测试用检测仪器

近年来，随着手机应用的普及及其视频功能的不断提升，每个手机用户都可以轻而易举地录制各种视频信息，为重复性动作计数提供了丰富的视频数据资源；同时，计算机视觉技术及机器学习技术的飞速发展，为重复性动作计数奠定了良好的技术基础；使得基于计算机视觉技术进行生活中各种重复性动作计数具备了可能。相比传统的光电传感器检测方法，用手机录制视频，并用计算机视觉与机器学习的方法来进行各种重复性动作的计数，将更易于操作、具有更好的应用普及性。

对现实生活中的重复性动作进行计算机视觉识别并计数的难点主要体现在 5 方面。一是被检测的做重复动作的对象无限制。可以是人，如做仰卧起坐、引体向上的人；也可以是物体，如排球垫球中上下往复运动的排球。二是各种重复性动作种类不同。仰卧起坐是人体上半身躺下再坐起的一个角度范围内的摆动，引体向上和排球垫球是垂直地面的上下运动，如图 1.2。三是各种重复性动作的运动幅度不同、重复周期时长（即频率）可变。仰卧起坐运动幅度是 0°~120°上半身的摆动，一分钟 49 次满分；引体向上是全身垂直方向 30-50cm 的起伏，13 次满分，两次动作之间不超过 10 秒；排球垫球是垫球高度离地面超过 2.15 米（女生）/2.35 米（男生），一分钟 40 次满分[12]。四是各种重复运动的总次数较低，不到 50 次。五是由于摄像过程中的抖动，导致视频图像背景呈现不同幅度的运动，即在运动的背景下进行重复性动作的识别。可见，视觉重复计数的研究具有非常大的挑战性[13]。

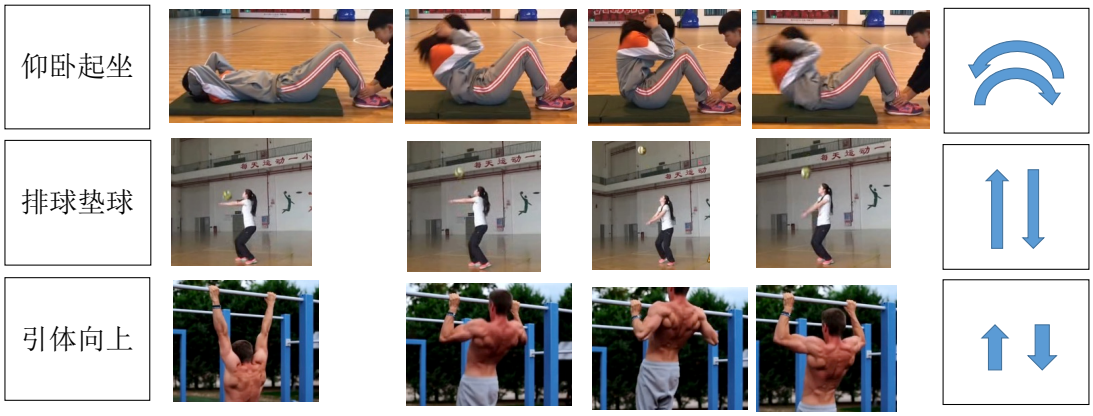


图 1.2 几种不同重复动作的运动模式区别

本文的研究目标是针对现实生活中的重复性动作进行基于计算机视觉的识别并计数，不限于被测对象种类（人体或物体），不限于动作的种类，不限于动作的幅度，不限于动作的频率及频率变化，在重复动作数量不多，且视频的摄像头以及视频的背景可能也在不同幅度地运动的情况下，有效地对视频中的重复动作进行识别和计数。

二、视觉重复动作计数的国内外相关研究的现状

基于计算机视觉的重复动作检测是在周期运动检测研究的基础上进一步深化扩展的技术。早期是针对固定周期的重复动作，力求识别动作重复的准确周期时长。而实际生活中很多重复性动作的循环周期不是常数，计数实际是对不同频率的重复动作进行数量统计。

2.1 运动周期检测方法

运动周期检测方法基本上可以分为两类：基于转换的方法和基于波形的方法。

基于转换的方法首先将视频信号转换到某个变换域，然后进行相应的周期检测。例如，一些基于频域的方法将视频信号转换到频域，提取除零频率外的最大峰值作为估计周期长度。Briassouli 和 Ahuja[14]将每个视频帧的像素值投影到 x 轴和 y 轴上，得到两个随时间变化的信号，然后对这两个信号进行时频分析，进行周期估计。然而，由于基于转换的方法是基于变换域的分析，它们只能处理周期恒定的运动，而不能处理周期长度变化的运动。此外，基于转换的分析通常需要依据大量重复周期的数据来估计准确的周期长度。因此，当应用于包含少量周期的视频序列时，其精度可能较低。

基于波形的方法首先从视频序列中提取反映运动随时间的周期变化的一维波形。然后通过分析波形提取周期。由于基于波形的方法更加灵活，对周期数的要求较低，因此在周期检测中得到了更广泛的应用。基于波形的算法已有很多种。例如，Cutler 和 Davis[15]根据帧与帧之间的绝对差异计算自相似波形，然后创建一个二维晶格结构来匹配自相似波形来找到周期。然而，由于该方法对变长周期运动的处理能力较差，因此仍存在一定的局限性。Wang 等人利用[16]提取物体下半身的宽度作为波形提取步态周期。虽然这种方法可以有效地检测步态周期，但它是基于非常具体的假设，不能扩展到检测其他运动。

此外，在实践中，许多周期性的运动只是由物体的局部动作反映出来的（例如，挥手运动中的手）。当包含整个对象时，其他无关部分的噪声运动可能会对最终的结果产生很大的影响。对于只有轻微运动的周期运动，这些噪声效应将变得非常明显。Gaojian Li 提出了一种基于凸包的（Convex-hull-based，CHB）程序来自动确定运动的感兴趣区域。这样，在周期检测时只考虑感兴趣区域，有效地排除了其他无关部件的噪声运动[17]。Y. Ren 提出了基于运动历史图像的算法实现运动模式定位，以检测重复运动的目标区域[18]。

2.2 重复性动作的计数方法

现有的视频重复估计方法通常将视频表示为一维信号，维持运动的重复结构。然后通过傅里叶分析[19]、峰值检测[20]或奇异值分解[21]提取频率信息。Pogalin 等人[22]通过跟踪一个对象，在跟踪区域上执行主成分分析，并使用基于傅里叶周期图来估计视频中的运动频率。Pogalin 等人在二维视场中识别了四种视觉周期性运动类型(平移、旋转、变形和强度变化)，并补充了三种运动连续性(振荡、恒定和间歇)。

Briassouli 和 Ahuja[14]采用短时傅里叶变换用时频分析方法处理多周期运动。Burghouts 提出了一种用于估计视频重复的时空滤波器组[23]。他们的过滤器在线工作，并且在正确调优后是有效的。为了提高的分辨率，采用连续小波变换代替短时傅里叶变换[24]。

Levy&Wolf [25]提出了一种通过卷积神经网络实时对视频中重复动作进行计数，该方法使用合成数据模拟在四种运动类型下的周期运动，进行网络的训练和预测，在测试时，对测试数据通过运动阈值法计算出感兴趣的区域，并通过分类网络对运动周期进行分类，完成重复计数任务，该方法在 YT_segments 数据集上显示了良好的性能。

依赖于周期运动的傅里叶分析方法不能处理非平稳运动。Runia[13]采用小波变换来更好地处理非静态和非平稳的视频动态。从流场和流场的微分出发，推导出三维空间中具有内在周期性的三种基本运动类型和三种运动连续性。

2.3 研究现状的分析

根据上述已有研究现状的总结，发现早期的视频重复性动作分析都是针对整幅图像进行分析，与重复动作不相关的部分会带来噪声和干扰，为了降低噪声与干扰对重复动作分析的影响，通过选取视频中某一特殊感兴趣区域进行分析，可以降低其他部分的干扰，但不同视频中重复动作的区域不确定，使视觉分析方法的通用性受到影响。其次，以往算法对视频中重复动作的运动形式也有所局限，不能推广到广泛的任意动作中，使得应用范围受限。另外，对不平稳的运动分析较弱，也约束了视觉重复计数的应用准确性。

基于已有的视觉重复动作分析方法，针对存在问题和不足，本论文的主要工作首先是解决视频重复动作的特征提取问题，使被检测的重复动作不受到视频中其他区域的干扰，不受动作在视频中所在区域的限制，不受动作类型的限制；而后把视频特征的重复性通过降维表示为保持运动重复结构的一维波形；再后，通过去噪使得一维波形平滑；最后通过峰值个数统计实现重复性动作的计数。

三、视觉重复动作计数的研究方案

本文针对视频数据源的重复性动作识别和计数的需求，建立了由视频数据进行特征提取、重复动作特征降维转化为重复波形、波形去噪、波峰计数的解决思路，具体分为四个步骤，图 3.1 为本文重复动作计数思路的框架图。

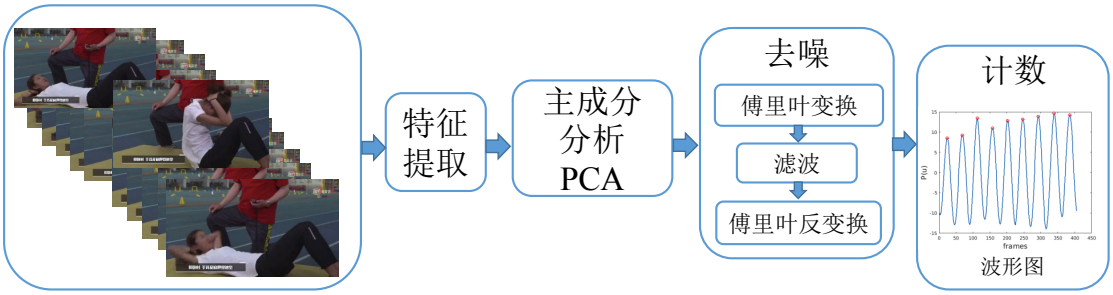


图 3.1 基于计算机视觉的重复动作计数思路图

第一步实现视频中重复性动作的特征提取。先对 RGB 图和光流图通过双流卷积神经网络进行特征的提取，为的是将视频转换为特征数据，学习到具有代表性和分辨性的有效高语义特征，去除某些不必要的底层语义特征。

第二步对高维特征数据进行降维。主成分分析可以有效检测时序信号中的共变部分。通过主成分分析 PCA 将高维数据降至低维，保留与重复性动作频率最相关的一些维度。

第三步对降维后的波形进行去噪。经过降维后的数据大体周期变化规律明确，但仍然存在很多细节上的波动，影响计数统计。使用快速傅里叶变换将时域信息转为频域信息，结合分段阈值滤波，再做逆运算转回时域信息，达到平滑波形的作用。

第四步统计图像波峰个数即可得重复动作次数。

四、视觉重复动作识别和计数的算法原理

视频数据源的重复性动作识别和计数分为四个步骤，下面对其中的特征提取、数据降维、滤波去噪的算法原理进行逐一分析。

4.1 重复性动作视频的特征提取

本文分段处理视频，并使用双流卷积神经网络提取视频特征。

（1）双流卷积神经网络

只有将视频转换为数据，才便于进行运算。所谓“特征提取”就是进行这一步的转换。一种常见的方法是使用卷积神经网络（CNN）[26]。它最普通的形式是以一张图片作为输入，经过多层的运算，得到一组数据作为输出。在此基础上稍加改造，就可以应用于本文中所需的对于视频的特征提取。

视频是由一帧一帧的画面构成的，想要提取视频的特征，要先提取构成它的画面的特征，也就是把一帧一帧的画面抽离出来，分别提取这些画面的特征。除此之外，视频所独有的是画面和画面之间的联系——光流。光流是空间中运动的

物体投影在平面上的像素运动的瞬时速度，是利用图像序列中像素在时域上的变化以及视频相邻帧之间的相关性来找到上一帧跟当前帧之间存在的对应关系，从而计算出相邻帧之间物体的运动信息的一种方法[27]。光流图的特点就是主要反映运动主体的运动，表示动态的过程。

由上面的分析，想到可以把卷积神经网络的输入拓展，拓展为单帧的 RGB 图像和视频的光流图，其中单帧的 RGB 图像表示静态信息，光流图表示动态信息。这样的卷积神经网络称为“双流卷积神经网络”——既有空间流又有时间流的卷积神经网络。空间流便是 RGB 图像，时间流便是光流图。空间流卷积神经网络和时间流卷积神经网络分别计算，然后再融合，最后得到整体的结果。

（2）视频的分段处理

由于在真实的体育考试中，视频可能时间较长，比如在仰卧起坐的计数中，一般持续时间至少为一分钟，引体向上测试也能达到半分钟之久，所以如果把整个视频的每一帧画面和光流图直接作为输入的话运算量太大了。在长视频处理中，如果遇到这类问题，一般有两个解决思路：只处理一段或几段视频，或者稀疏采样[28]。前者容易理解，就是截取视频中的一部分或者几部分代表整个视频。后者是指将视频分成等长的几段，在每段中随机抽取一些画面和光流图。但是以上这两种方法都不适用于重复动作计数：因为为了统计视频中重复动作的次数，必须对完整的视频进行计算，不能只计算其中的一个或若干个片段，也不能稀疏采样。本文采用的方法是将视频分成等长的几段，分别对每一段计算，最后再连接起来。这样可以在保证质量的情况下稍降低一些计算量。

（3）单帧图像特征提取

卷积神经网络最核心的部分有两个，分别是它的网络结构和它的参数，它们很大程度上直接决定了提取出来的特征的有效性。本文使用的是基于文献[28]的双流卷积神经网络。它使用 BN-Inception 网络结构，在效率和准确性之间达到较好的平衡[29]。该模型的参数是在大型公共数据集 Kinetics 上训练得来的。文献[28]研究的问题是对视频中的行为进行分类，并且使用了稀疏采样策略处理长视频。所以本文在应用该模型的时候，去掉了所有有关分类的部分，并且改稀疏采样为密集采样。对实验数据集，分别以单帧 RGB 图和光流图作为输入，进行特征提取，如图 4.1 所示。

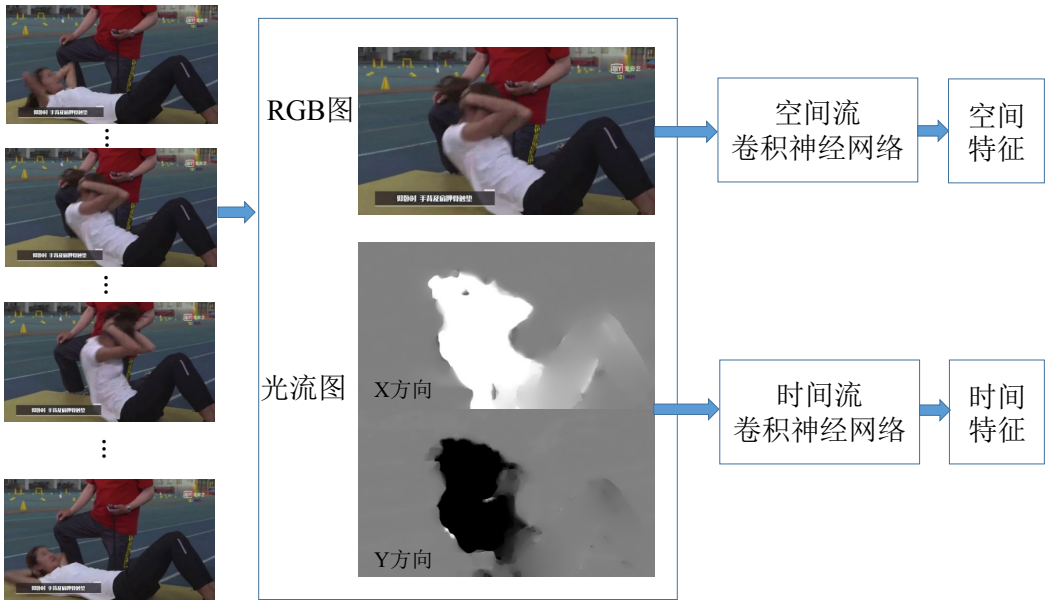


图 4.1 视频特征提取的框图

这一环节结束以后，每个视频都得到两个大小为 $N\times1024$ 的矩阵，一个是空间特征，即以 RGB 图像作为输入的结果，一个是时间特征，即以光流图作为输入的结果，其中 N 是该视频的帧数。矩阵的每一行是一个样本，每一列是一个维度。

4.2 RGB 图和光流图的数据降维

（1）降维分析

一个时长为 1 分钟的标准的仰卧起坐测试视频，帧率为 30fps，那么整个视频一共有 1800 帧，则上一步得到的矩阵大小为 1800×1024。这对于数据的存储和后续计算要求太高，所以应当尝试将特征数据降维。

应当如何降维？首先考虑去掉那些在不同维度之间，变化很小的维度，因为它们总是一成不变，对计数并没有带来什么贡献。并且过多的维度并不利于信息的识别，因为动作每次重复不可能一模一样，会有不同程度的出入。所以在数据的降维过程中，把这些干扰信息去除。因此，降维的两个主要任务是去掉在不同样本之间变化甚微的维度和干扰信息。PCA 算法可以将数据降至任意纬度 k，保留最主要的成分[30][31]。

（2）降维方法

需要降维的对象除了上一步得到的两个大小为 $N\times1024$ 的矩阵以外，还有一个大小为 $N\times2048$ 的矩阵，为上一步得到的两个矩阵合并在一起的结果。它融合了动态和静态信息。

降维最终的目标是在不丢失重要信息的前提下，将三个矩阵降为 $N\times k$ 的大

小。本文中取 k=10，问题的关键在于如何选择要保留的 10 个维度。

设对一个矩阵 $A_{N\times D}$, D=1024 或 2048 操作， PCA 算法的流程如下：

1）计算协方差矩阵[32]；

这里指的是要计算不同维度之间的协方差，所以形成的协方差矩阵为 $C_{D\times D}=(c_{ij})=(\mathrm{cov}(d_i,d_j))$，其中 d_i 表示矩阵 A 的第 i 列数据，d_j 表示矩阵 A 的第 j 列数据。按定义计算协方差矩阵时间复杂度太高，另一种计算协方差矩阵的方法是先将矩阵 A 中心化，即每个元素都减去其所在列的所有元素的算术平均值，然后用这个新矩阵右乘上它的转置，最后乘上系数 1/(N-1)即得协方差矩阵 C：

$$E_{N\times D}=(e_{ij})=(a_{ij}-\frac{1}{N}\sum_{k=1}^{N}a_{kj})$$

$$F_{D\times N}=E^T=(f_{ij})=(e_{ji})=(a_{ji}-\frac{1}{N}\sum_{k=1}^{N}a_{ki})$$

$$FE=(\sum_{l=1}^{D}f_{il}e_{lj})=(\sum_{l=1}^{D}(a_{li}-\frac{1}{N}\sum_{k=1}^{N}a_{ki})(a_{lj}-\frac{1}{N}\sum_{k=1}^{N}a_{kj}))=(N-1)C$$

2）求协方差矩阵的所有特征值及对应特征向量，将特征值从大到小排序，以前 k 个所对应的特征向量为基，构建 k 维线性空间；

协方差矩阵 C 的对角线上是每个维度内部的方差，表示每个维度在不同的样本之间的差异有多大，对角线之外的部分就是不同维度之间的协方差，表示这些维度之间的关联有多强。我们的目标是让不同维度之间的干扰尽量少，并且让每个维度内部的差异尽量大，也就是让矩阵 C 中非对角线上的元素尽量小，让对角线上的元素尽量大。于是想到可以将矩阵 C 对角化，之后它的对角线上就是 C 的特征值，其余位置都为 0，达到了我们最小化协方差的目的。在此之后，只需要保留对角线上元素前 k 大的维度，舍弃其余维度即可，因为这 k 个维度的方差最大，是最主要的成分。

所以，我们需要求出协方差矩阵 C 的所有的特征值以及对应的特征向量，并按特征值从大到小排序，用前 k 个特征值所对应的特征向量从左到右排列，形成投影矩阵 P。

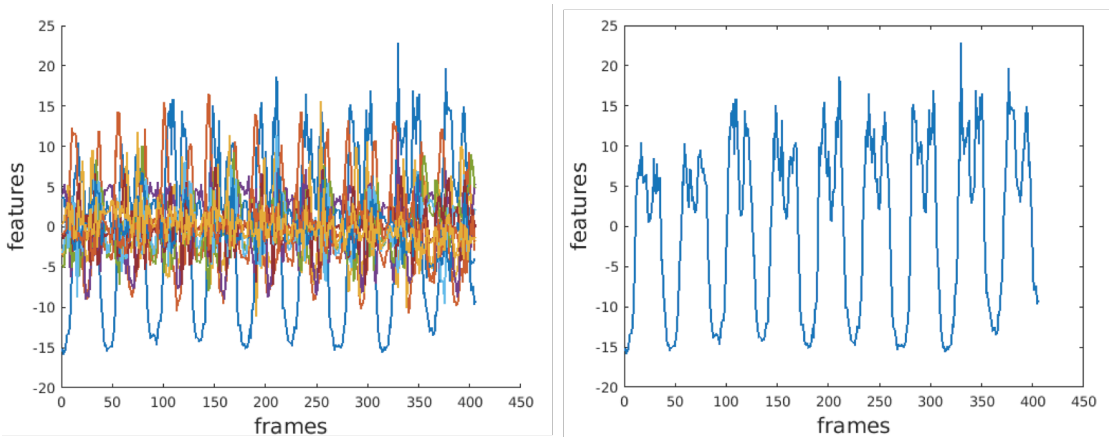
3）将原矩阵 A 投影到新的空间中，即得到降维后的矩阵。

得到投影矩阵 P 之后，计算 $A'=PA$，即可得到降维后的矩阵。

通过这三步，得到了经过 PCA 降到 10 维后的数据。

4.3 波形图的分段阈值滤波

接下来要做的是尝试从这大小为 $N\times 10$ 的矩阵中找出周期变化的规律，并计数。在观察数据的时候，应当纵向观察，因为矩阵每一行是一个样本，由上至下代表着时间的推进。把数据用图像呈现出来更有利于观察周期变化。于是可以把数据画在一个平面直角坐标系中，横轴代表时间，纵轴是数据，分别把每一维度的数据画在同一张图像上。具体地，把矩阵的 10 列数据分开，对于每一列，把该列的数据作为纵坐标，配以均匀递增的横坐标，描在坐标平面内，再用平滑的曲线连接（这里用软件作图，是用的直线段连接），如图 4.2a。发现第一维数据具有很明显的周期变化，且变化幅度较大，比较清晰，如图 4.2b。



a. 10 维主成分　　b. 第一维主成分

图 4.2　PCA 降维后的数据

第一维主成分已经可以很好地表现整个周期运动，所以接下来只用第一维的数据运算，效果等同于在 PCA 步骤中取 k=1。

单看第一维数据的图像，它类似正弦波，有着上下的往复变化。与正弦波不同的是，它每次往复的幅度以及频率是不同的，因为每次动作的幅度和频率是不同的。想要统计动作重复的次数，就要统计图像中往复变化的次数。因为往复的频率是变化的，所以简单地用横轴坐标的总变化量除以频率是不可取的，只能一次一次地数。可以将数往复变化的次数归结为数“正弦波波峰”的个数。理想状态下，有多少个波峰就代表动作重复了几次。但是从上图中可以看到，直接统计有多少个中间比两边高的数据点是得不到正确的“波峰”数的，因为曲线不够平滑。所以首先要进行降噪处理，让图像上的线条尽量平滑，便于统计。

在数字信号处理中，通过傅里叶变换将时域的信息转换到频域中，设定某一阈值然后滤波从而达到去掉干扰信息是一种常用的方法。时域是以时间为横坐标的域，是客观存在的域，也是最直观的域；频域则是将时域中的曲线拆分成若干

个正弦波的线性组合后从侧边的投影图。经过一次快速傅里叶变换，时域中的信息转换到了频域中，也就是说，时域中的曲线被拆分成了若干个正弦波的线性组合。这个线性组合是唯一的，它们在频域中的图像如图 4.3 所示。

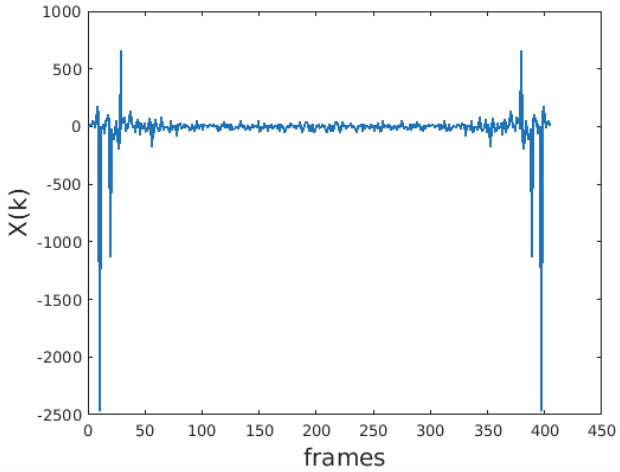
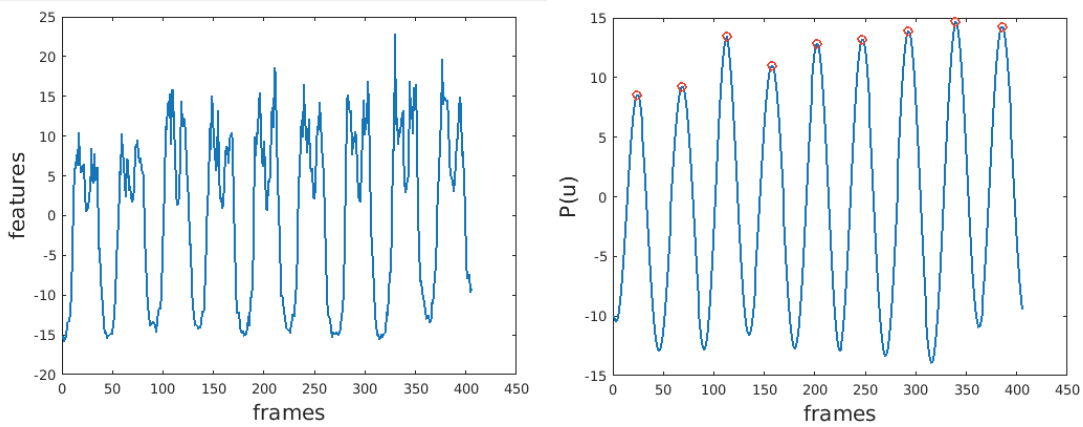


图 4.3 图 4.2b 中第一维主成分的频域图像

这些正弦波并不是所有都重要，应该将其中带来干扰的部分去掉。从图 4.3 中可以看出，频域图像是具有对称性的，这是因为傅里叶变换具有对称性：两侧对应的正弦波的振幅较大，中间对应的正弦波的振幅较小。为了去除干扰部分的影响，可以将中间振幅较小的部分过滤掉：选取一个合适的阈值 *threshold*，记分解出来的正弦波一共有 *L* 个，把下标在[*threshold*, *L-threshold*]范围内的波都去掉，也就是将这些波在线性组合中的系数改为 0，然后做一次快速傅里叶逆变换，将滤波以后的频域信息转换回到时域中。滤波后的信息在时域中的图像如图 4.4b。可以看到，曲线变得平滑了很多，圆圈处是按照“判断是否两侧数值都小于中间”识别出来的波峰。现在，波峰的统计变得准确了很多。



a. 滤波前　　b. 滤波后

图 4.4 分段阈值滤波前后的第一维主成分

以上流程中，阈值 *threshold* 的选择非常重要：如果太小，那么无法过滤掉足

够多的干扰信息，导致还是无法统计波峰；如果太大，过滤掉了有用的信息，也达不到正确计数的目的。本文中，*threshold* 并不是一个固定的常数，而是在 YT-segments 数据集上实验得来，见 5.3 节。

五、视觉重复动作计数方法的实验验证

本文实验数据源于多样化和具有挑战性的真实生活场景，具有不同的重复长度，且包含相机和背景的运动。本文对比了两个来自 youtube 的数据集：YT segments[25]和 QUVA[13]。本文特征提取工作使用基于文献[29]的训练模型，该模型使用大型主流的数据集 Kinetics, 该数据集包含 30 万个来自真实场景的剪辑动作视频，共有 400 个动作类别。该方法的显著性在 2017 Activity 挑战赛中得到了很好的证明。另外，本文的重复动作计数任务没有训练过程，直接对实验数据提取的特征进行重复动作计数分析。

5.1 视频实验数据集

YT_segments 数据集：包含具有重复内容的 100 个视频数据集，这个测试数据集很好的显示了各领域的组合，包括锻炼、烹饪、建筑、生物等，视频只包含重复的动作，每个视频的重复次数被预先标记，其中最短重复和最长重复次数分别为 4 和 50，视频平均时长为 14.96s。其中包含 30 个视频具有不同程度的摄像机运动。

QUVA 数据集：由 100 个视频组成，展示各种重复的视频动态，包括游泳、搅拌、切割、梳理和音乐制作。与数据集 YT_segments 相比，该视频数据在周期长度、运动外观、摄像机运动和背景复杂度方面有更多的变化。通过增加场景复杂性和时间动态的难度，使得该数据集作为一个更现实和更具挑战性的基准估计重复视频。

5.2 评估准则

本文使用的评估准则是把真值 G 和预测值 R 之间计数的绝对差的百分比作为评估结果:$\frac{|G-R|}{G}\times 100$。对于 N 个视频，计算平均绝对误差(MAE)$\pm$标准偏差(σ)[25]，其中

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|G_i-R_i|}{G_i},$$

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}(G_i - R_i)^2。$$

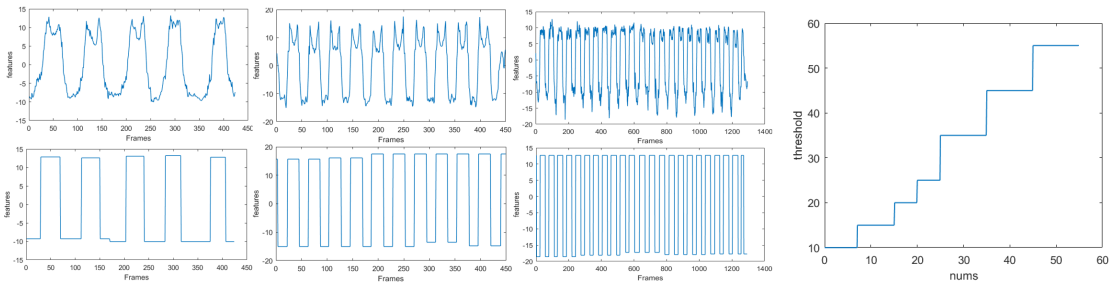
5.3 滤波阈值分析

在 4.3 节中，我们基于频谱图滤除噪声频带，来修正波形图。为了验证不同滤波阈值对实验结果的影响，基于 RGB 特征进行分析，首先根据经验值设置固定 *threshold* (α={10, 15, 20, 25, 30, 35, …})，然后在 YT_segments 数据集上对重复动作计数进行实验，结果如表 5.1 所示。

表 5.1 不同阈值的对比分析

α	MAE
10	23.3±63.4
15	19.9±42.8
20	30.9±38.14
25	44.6±42.8
30	56.0±62.7
35	67.2±91.6
multi-stage α	8.7±3.9

实验结果表明滤波阈值 *threshold* 过大和过小都会改变原本的波形信息，从而失去消除噪声的目的，反而会使检测效果更差，因为固定的阈值并不适合多样化的频率运动，而分段阈值傅里叶变换表现了较好的性能。



a. 不同运动频率可视化效果　　b. 分段阈值选择

图 5.1 滤波阈值选择分析

本文根据不同频率变量的运动特性进行频谱阈值选择。基于一维主成分的高频信息计算高通频带数，如图 5.1a 所示。然后优化频谱阈值，得到图 5.1b 所示的分段阈值选择基准，来修正运动波形图，使重复运动具有明显的时序特性，更利于重复动作计数。

5.4 视觉重复性动作计数的实验结果分析

分别对基于 RGB 和光流图提取的空间和时序特征进行分析，然后对比 RGB + 光流特征的融合结果，同时还分析了傅里叶变换模块对实验效果的影响，结果

表明在 YT_segments 数据集上加入傅里叶变换后准确率得到了较大提升，另外，也表明基于 RGB 流的特征达到了最好的效果，如表 5.2 所示。

表 5.2 TY_segments 数据集的实验结果对比分析

数据集	未滤波 MAE			滤波 MAE		
	RGB	光流	RGB +光流	RGB	光流	RGB + 光流
YT_segments [25]	13.7 ±6.2	29.2 ±21.3	18.2 ±9.44	8.7 ±3.9	21.9 ±12.7	15.8 ±6.6

本文的方法与现存较显著的三种方法进行对比，实验结果如表 5.3 所示。在 YT_segments 数据集上文献[25]表现最好，MAE 为 6.5。其中文献[13]的方法 MAE 为 10.3，优于基于文献[22]的方法。本文的方法 MAE 为 8.7，优于文献[22]和[13]的方法，但标准误差与以上方法相比达到了最好的性能。在更具挑战性的 QUVA 数据集上，本文的实验结果 MAE 为 25.1，也达到了良好的性能。方法[25]表现最差，MAE 为 48.2，这是因为他们的网络在训练时只考虑四种运动类型。方法[22]的 MAE 评分为 38.5，文献[13]中 MAE 为 23.2。在两个公共数据集上，本文的方法得到的标准误差均达到了最好的效果，证明了该方法的有效性。

表 5.3 不同方法的实验结果对比

数据集	YT_segments	QUVA
	MAE	MAE
Pogalin *et al.*[22]	21.9±30.1	38.5±37.6
Levy & Wolf [25]	6.5±9.2	48.2±61.5
Runia &Snoek[13]	10.3±19.8	23.2±34.4
我们的实验	**8.7±3.9**	**25.1±25.2**

六、视觉重复动作计数的应用

6.1 中考体育项目计数的实验

实验视频来自网络，只包含仰卧起坐、引体向上和排球垫球的视频。这些视频中的重复动作次数从 3 次到 48 次不等，视频时长从 5 秒到 65 秒，镜头和背景有移动的，也有固定的。按特征提取、数据降维、滤波去噪、统计计数四步骤分别对三种体育项目的视频进行计数，每个项目各举一个例子如下，计数视频已按附件上传。

（1）仰卧起坐计数

图 6.1 为仰卧起坐视频的计数过程。采用 CNN 进行特征提取后，分别采用 RGB 图和 XY 两方向光流图进行 PCA 降维，并进行波形去噪和计数。仰卧起坐计数真实值：9 个；RGB 特征提取计数：9 个，光流图特征提取计数：9 个。说明两种计数均很准确。

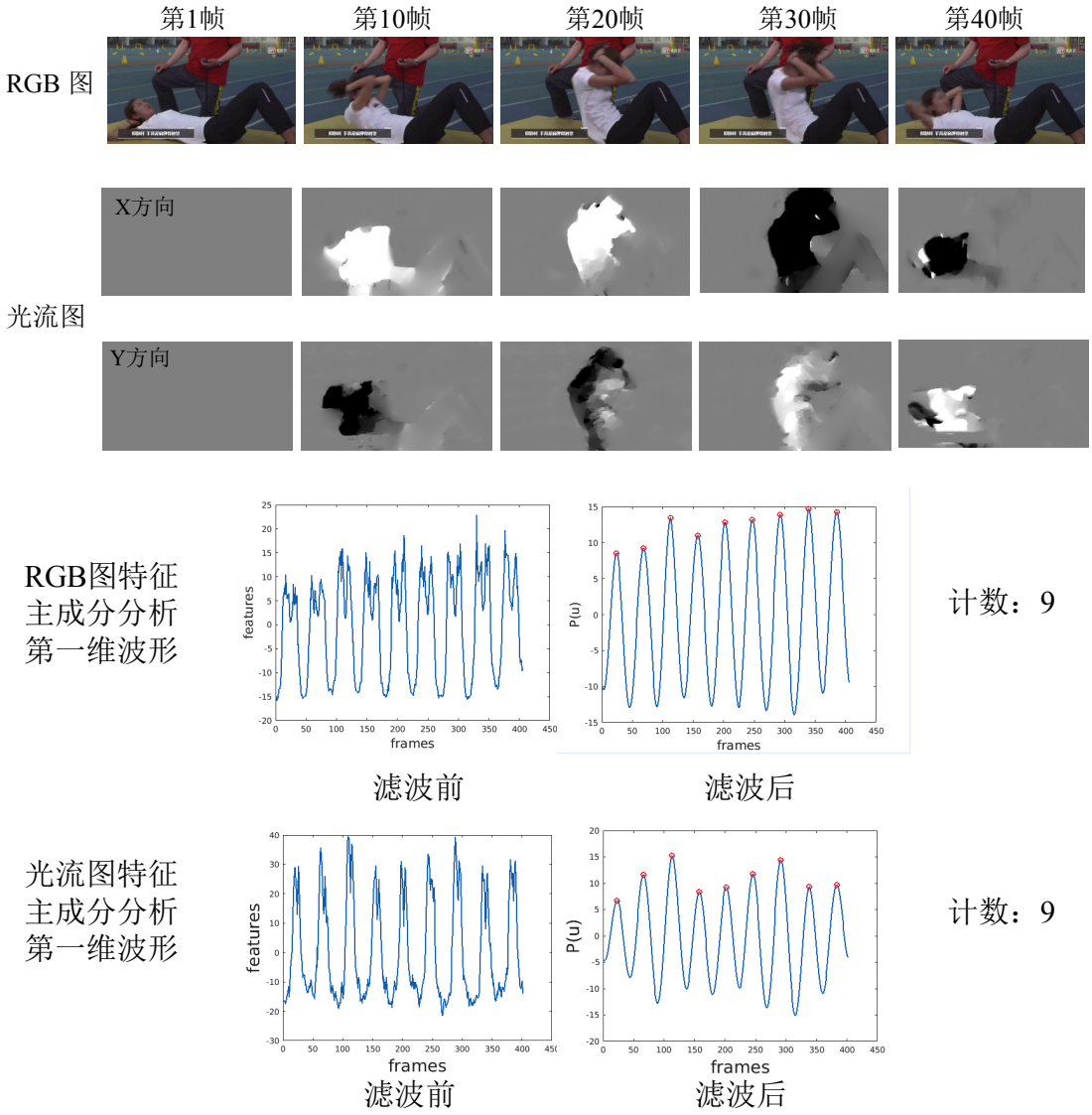


图 6.1 仰卧起坐重复动作视觉计数过程

（2）引体向上计数

图 6.2 为引体向上视频的计数过程。采用 CNN 进行特征提取后，分别采用 RGB 图和 XY 两方向光流图进行 PCA 降维，并进行波形去噪和计数。引体向上计数真实值：19 个；RGB 特征提取计数：19 个，光流图特征提取计数：22 个。说明 RGB 特征计数准确，而光流图特征计数有偏差。这个引体向上视频中背景

有很多围观的学生，且由于手机是竖直放置拍摄，所以双侧有较大视野是没有信息的，但通过重复性动作的特征提取，这些背景信息没有对计数造成太大影响，说明本文的算法具有较好的抗干扰能力。

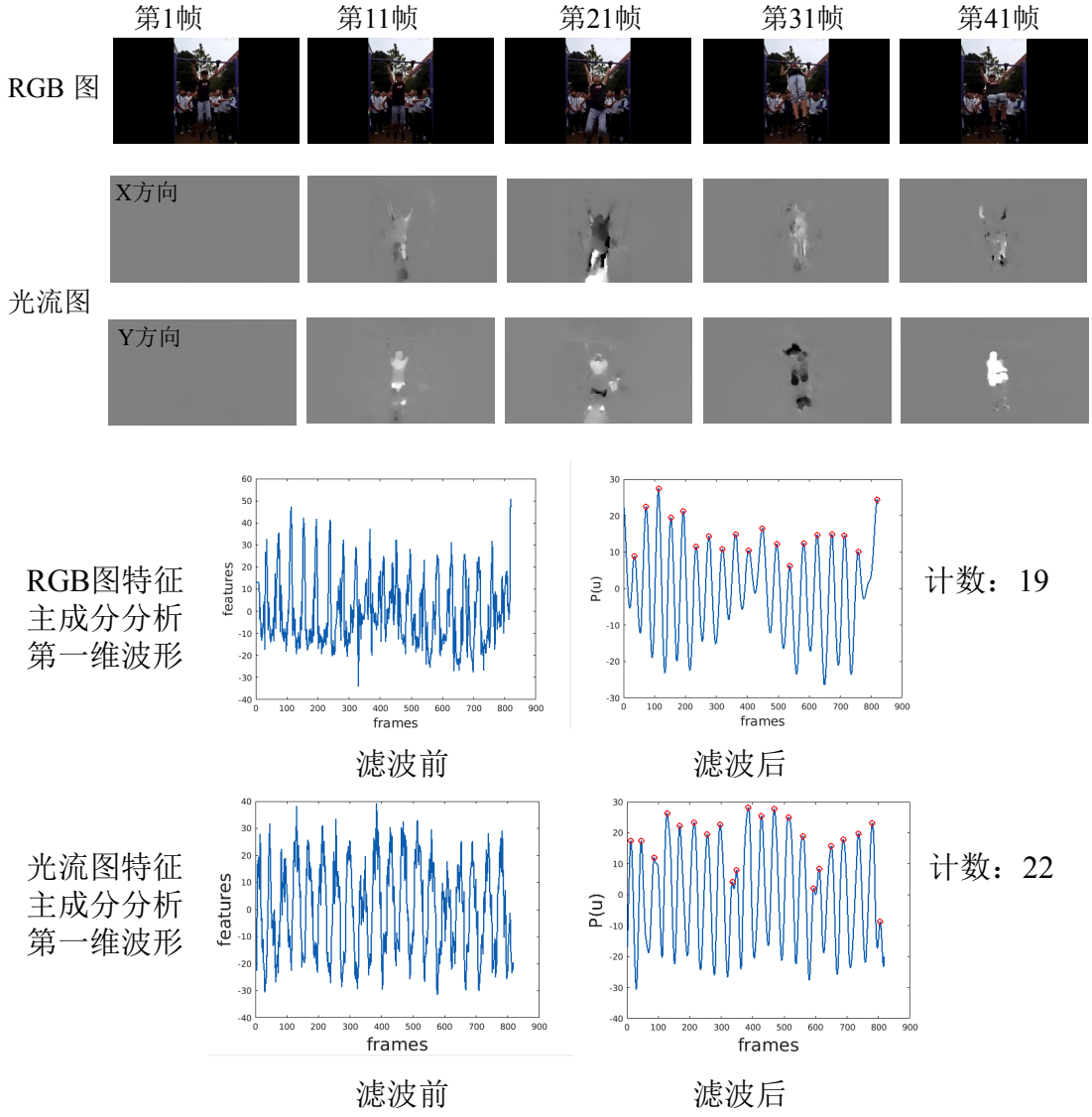


图 6.2 引体向上重复动作视觉计数过程

（3）排球垫球计数

图 6.3 为排球垫球视频的计数过程。采用 CNN 进行特征提取后，分别采用 RGB 图和 XY 两方向光流图进行 PCA 降维，并进行波形去噪和计数。排球垫球计数真实值：20 个；RGB 特征提取计数：11 个，光流图特征提取计数：18 个。排球垫球计数的准确性不如仰卧起坐和引体向上，主要原因是垫球过程中，测试人员在一定区域内往复移动、旋转，使得每次动作的相似性降低。且人的手臂在做上下的小幅度重复性动作，而排球在做大幅度上下运动，因此视频中同时出现

了两个做重复性动作的特征，形成了交互的干扰，导致计数准确率下降。

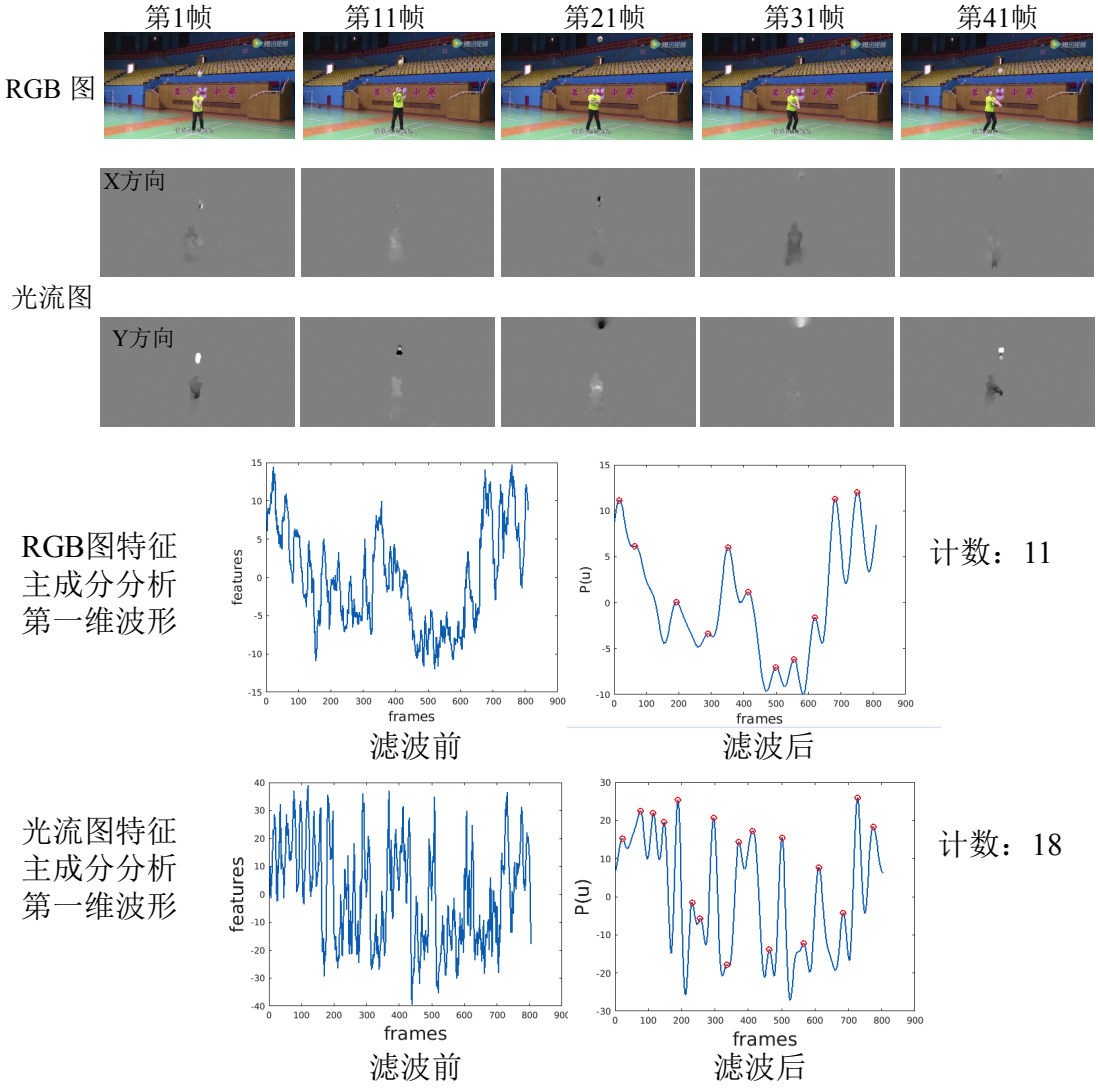


图 6.3 排球垫球重复动作视觉计数过程

6.2 视觉重复动作计数的讨论

（1）视觉计数方法准确性

对比基于 RGB 图和光流图的特征提取和计数，基于 RGB 的准确性更高，尤其是对仰卧起坐、引体向上这种单一重复动作的计数。但在排球垫球考核中，不够熟练的考生会在一个 3 米×3 米的正方形区域内移动，存在平移、旋转等情况，一些视频中镜头也会随着考生移动，每次动作之间的相似性减少，这样会很大地影响识别的准确度。另外，人的手臂在做上下的小幅度重复性动作，而排球在做大幅度上下运动，因此视频中同时出现了纵向两个耦合重复性动作，此时基于光流图特征提取的计数准确性就高于基于 RGB 特征提取的计数，说明了不同的特

征提取方法是要考虑被检测对象的耦合干扰问题。

（2）视觉计数方法的影响因素

影响识别准确度的主要因素有镜头或动作主体的移动、多种重复动作的干扰、动作重复的次数、和动作重复的频率。

在排球垫球考核中，由于垫球人会存在平移、旋转等情况，镜头也会随着人移动。这样会很大地影响识别的准确度，因为每次动作之间的相似性减少。另外，垫球过程中，人的手臂的小幅度重复性动作，与排球大幅度上下运动耦合，形成了手臂动作对排球重复性动作分析的干扰，导致计数准确率下降。

动作重复的次数如果太少，会影响阈值的选择，很难正确计数。应用本文的算法，在动作重复频率合理的情况下，一般动作的重复次数在 6 次以上的情况识别准确率较高。

动作重复的频率也会影响识别准确度。如果频率过高，相邻两帧之间的差异太大，可能会忽略掉一些次数。如果频率过低，会将过多噪声信息记录了进来，也会影响计数准确性。

七、结论与展望

7.1 结论

本文针对现实生活中的重复性动作进行基于计算机视觉的识别并计数，在不限制被测对象种类、动作种类、动作幅度、动作频率及频率变化，且重复动作数量不多，视频背景运动的情况下，能够实现有效地对视频中的重复动作进行识别和计数。

本文的核心思想是通过对具有重复性动作的视频数据进行特征提取和降维，转化为与重复动作频率一致的重复性波形，把重复性动作计数转变为波峰数统计。

本文的重复性动作计数方法分为四步：视频重复动作特征提取；视频重复性特征的降维和特征波形的选取；特征波形的去噪；特征波形峰值个数统计。该方法首先基于 RGB 和光流图应用深度卷积神经元网络（CNN）进行视频特征提取。然后使用主成分分析（PCA）对特征时间序列进行分析，提取体现重复动作随时间变化运动规律的主成分，发现第一维主成分波形与重复性动作频率直接相关。进一步使用分段阈值滤波的傅里叶变换和反变换去除主成分波形的噪声信息，获取明显的运动特性波形。最终通过峰值检测统计重复运动次数。

通过本文方法对比了两个来自 youtube 的数据集 YT segments 和 QUVA 的各 100 个重复动作视频。以所有视频的平均绝对误差和标准偏差作为评价标准。根

据不同频率变量的运动特性进行频谱分段阈值选择，来修正运动波形图，使重复运动具有明显的时序特性，更利于重复动作计数。经基于 RGB、光流与 RGB + 光流特征融合提取的空间和时序特征进行分析，表明在 YT_segments 数据集上基于 RGB 流的特征达到了最好的效果，且加入傅里叶变换和反变换后准确率得到了较大提升。最后经与其他三种较优的已有视觉重复计数方法比较，无论在数据集 YT segments 还是 QUVA 上，本文的方法均接近最优，且标准误差均达到了最好的效果，证明了该方法的有效性。

将本文视觉重复计数方法应用于中考体育测试项目（仰卧起坐、引体向上、排球垫球）中重复性动作的计数，证明了该方法对于单一重复动作计数的准确性很高，具有很好的未来应用可行性。本文的计数方法不需要训练或者限定动作类型，在应用上比使用大型的仪器计数更加简便、易于普及。

7.2 工作展望

现在影响计数准确率的因素为镜头或动作主体的移动、多个重复性动作的干扰、动作的重复次数，和动作的重复频率。以后的工作将当着重于对计数干扰因素的研究，不断提高计数准确性。

参考文献

[1] G. Johansson. Visual perception of biological motion and a model for its analysis. Perception & Psychophysics, pp. 201-211, 1973.

[2] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. Behavior classification by eigendecomposition of periodic motions. Pattern Recognition, 38(7):1033-1043, 2005.

[3] C. Lu and N. J. Ferrier. Repetitive motion analysis: Segmentation and event classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(2): 258-263, 2004.

[4] I. Laptev, S. J. Belongie, P. Perez, and J. Wills. Periodic motion detection and segmentation via approximate sequence alignment. IEEE International Conference on Computer Vision, 2005.

[5] B. Sarel and M. Irani. Separating transparent layers of repetitive dynamic behaviors. IEEE International Conference on Computer Vision, 2005.

[6] A. B. Albu, R. Bergevin, and S. Quirion. Generic temporal segmentation of cyclic human motion. Pattern Recognition, 41(1):6–21, 2008.

[7] Y. Ran, I. Weiss, Q. Zheng, and L. S. Davis. Pedestrian detection via periodic motion analysis. International Journal of Computer Vision, 71(2):143–160, 2007.

[8] S. Belongie and J. Wills. Structure from periodic motion. In Spatial Coherence for Visual Motion Analysis, pp. 16-24. Springer Berlin Heidelberg, 2006.

[9] S. Huang, X. Ying, J. Rong, Z. Shang, and H. Zha. Camera calibration from periodic motion of a pedestrian. IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[10] 沈永根，一种体育用坐位体前屈和仰卧起坐测试仪及使用方法，CN105935488B，2018 年 9 月 4 日。

[11] 任佳，袁祖斌， 体育教学计数装置，CN204159010U，2015 年 2 月 18 日。

[12] “2018 北京中考体育分数设置与具体评分标准”,北京市中考在线，http://www.zgkao.com/zk/201708/23805.html

[13] Tom F. H. Runia, Cees G. M. Snoek, Arnold W. M. Smeulders, Real-world repetition estimation by Div, Grad and Curl. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 9009-9017.

[14] A. Briassouli and N. Ahuja, Extraction and analysis of multiple periodic

motions in video sequences, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 1244-1261, 2007.

[15] R. Cutler and L. Davis, Robust real-time periodic motion detection, analysis and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 781-796, Aug. 2000.

[16] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, Chrono-gait image: A novel temporal template for gait recognition, European Conference on Computer Vision, 2010.

[17] Gaojian Li , Xintong Han , Weiyao Lin , Hui Wei, Periodic motion detection with ROI-based similarity measure and extrema-based reference-frame selection, IEEE Transactions on Consumer Electronics, Vol58(3) , August 2012, pp. 947-954.

[18] Y. Ren, B. Fan, W. Lin, X. Yang, H. Li, W. Li, and D. Liu. An efficient framework for analyzing periodical activities in sports videos. In Image and Signal Processing, 2011.

[19] O. Azy and N. Ahuja. Segmentation of periodically moving objects. International Conference on Pattern Recognition, 2008.

[20] A. Thangali and S. Sclaroff. Periodic motion detection and estimation via space-time sampling. IEEE Winter Conf. on Applications of Computer Vision, 2005.

[21] D. Chetverikov and S. Fazekas. On motion periodicity of dynamic textures. In The British Machine Vision Conference, 2006.

[22] E. Pogalin, A. W. M. Smeulders, and A. H. Thean. Visual quasi-periodicity. In IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[23] G. J. Burghouts and J.-M. Geusebroek. Quasi-periodic spatiotemporal filtering. IEEE Transactions on Image Processing, 15(6):1572-1582, 2006.

[24] O. Rioul and M. Vetterli. Wavelets and signal processing. Signal Processing Magazine, 8(4):14–38, 1991.

[25] O. Levy and L. Wolf. Live repetition counting. IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[26] 卢宏涛，张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述，《数据采集与处理》，2016 年第 1 期，1-17。

[27] Barron, John L., David J. Fleet, Steven S. Beauchemin． Performance of optical

flow techniques：International Journal of Computer Vision，1994, pp. 43-77.

[28] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition, European Conference on Computer Vision. Springer, Cham, 2016: 20-36.

[29] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning, PMLR, 37, 448-456.

[30] Svante Wold, Kim Esbensen and Paul Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems, 2 (1987) 37-52, Elsevier Science Publishers B.V., Amsterdam - Printed in The Netherlands.

[31] 周志华，《机器学习》，第 10 章降维与度量学习，10.3 主成分分析，清华大学出版社，2016 年 1 月，ISBN 978-7-302-42328-7。

[32] 刘吉佑、莫骄，《线性代数与几何》，第 2 章矩阵，北京邮电大学出版社，2012 年 8 月， ISBN 978-7-5635-3141-7。

致谢

论文的选题来源于中考体育测试中很多项目（仰卧起坐、排球垫球、引体向上）属于重复性动作计数，而相应的测试仪器复杂、传感器数量多、传感器位置还需因人调整、不同项目之间测试仪器不通用，带来了很大的操作不便和资源浪费问题，且在日常课上课外的体育活动中难以普及应用。结合目前手机摄像功能和手机使用普及率的不断提高、及计算机视觉和机器学习技术的飞速发展，进行基于计算机视觉的视频重复动作计数方法的研究具有很强的挑战性和广泛的应用价值。

作者自主选题，查找科研文献，在北京邮电大学自动化学院智能认知与信息处理实验室尹建芹老师指导下，系统学习了与课题相关的机器学习、计算机视觉理论知识。学习已有的深度卷积神经元网络和主成分分析对视频数据进行特征提取的原理和算法。在其科研团队的研究基础上，通过大量实验与讨论，发现视频的深度特征提取结合主成分分析可以提取较好的计数规律。作者主要贡献在于针对包含任意运动模式的任意视频，提出基于主成分分析反应计数规律的特征时序信号提取方法的思想，解决了本课题中的一个重要问题，并应用于实际场景下的体育视频中的重复动作计数，取得了良好的识别效果。

作者在北大附中肖然老师指导下学习了matlab、python程序设计，独立进行视频重复动作计数实验的验证。

两位老师均为无偿指导。

作者独立地完成了论文的撰写工作。

学术诚信声明

本参赛团队声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛队员：刘知宜　　指导老师：

2019 年 8 月 31 日