Journal of System Simulation

Volume 27 | Issue 10 Article 26

8-7-2020

# Human Action Recognition Method Based on Key Frames

Xiangbin Shi
*1. Department of Computer, Shenyang Aerospace University, Shenyang 110136, China;;2. Liaoning General Aviation Key Laboratory, Shenyang Aerospace University, Shenyang 110136, China;;3. College of Information, Liaoning University, Shenyang 110036, China;*

Shuanpeng Liu
*1. Department of Computer, Shenyang Aerospace University, Shenyang 110136, China;;*

Deyuan Zhang
*1. Department of Computer, Shenyang Aerospace University, Shenyang 110136, China;;*

# Human Action Recognition Method Based on Key Frames


## Abstract

Abstract: More and more researchers have begun to study the human action recognition based on depth information and skeleton information since the Kinect has been released. *A method of human action recognition based on the skeleton feature of key frames is proposed in order to improve the accuracy and timeliness of the human action recognition, and reduce the computational complexity.* The clustered data was obtained by using K-means clustering algorithm, and then the key frames were extracted by using the clustered data. Two features for human action recognition were extracted, one is the feature of the position of human joint, another is the feature of the skeleton angle between rigid body and rigid body. The sequence of action video was classified through the SVM classifier. This method leads to a more accurate recognition rate, and the real-time capability has been improved at the same time according to the result showed on the data set MSR-DailyActivity3D.



## Keywords

Kinect, human action recognition, key frames, k-means

## Recommended Citation

Shi Xiangbin, Liu Shuanpeng, Zhang Deyuan. Human Action Recognition Method Based on Key Frames[J]. Journal of System Simulation, 2015, 27(10): 2401-2408.

第 27 卷第 10 期 系统仿真学报© Vol. 27 No. 10
2015 年 10 月 Journal of System Simulation Oct., 2015

# 基于关键帧的人体动作识别方法

石祥滨[1,2,3]，刘拴朋[1]，张德园[1]
(1. 沈阳航空航天大学计算机学院，沈阳 110136；2. 沈阳航空航天大学辽宁通用航空重点实验室，沈阳 110136；
3. 辽宁大学信息学院，沈阳 110036)

**摘要：**Kinect 问世以来，越来越多的研究者开始研究基于深度信息和骨架信息的人体动作识别。*为了提高动作识别的准确率和实时性，并且降低计算过程中的计算复杂度，提出了一个基于关键帧的骨架特征的人体动作识别方法。*采用 K-均值聚类算法对人体动作视频序列做聚类，通过聚类出的数据提取人体动作视频序列中的关键帧。提取关键帧中的关节点位置和人体刚体部分之间的骨架角度两种特征，利用 SVM 分类器对动作序列进行分类。在 MSR-DailyActivity3D 数据集上的实验结果表明，该方法具有较高的识别率，并且提高了实时性。

**关键词：**Kinect；人体动作识别；关键帧；k-means
中图分类号：TP391.4 文献标识码：A 文章编号：1004-731X (2015) 10-2401-08

## Human Action Recognition Method Based on Key Frames

*Shi Xiangbin*[1,2,3], *Liu Shuanpeng*[1], *Zhang Deyuan*[1]
(1. Department of Computer, Shenyang Aerospace University, Shenyang 110136, China;
2. Liaoning General Aviation Key Laboratory, Shenyang Aerospace University, Shenyang 110136, China;
3. College of Information, Liaoning University, Shenyang 110036, China)

**Abstract:** More and more researchers have begun to study the human action recognition based on depth information and skeleton information since the Kinect has been released. *A method of human action recognition based on the skeleton feature of key frames is proposed in order to improve the accuracy and timeliness of the human action recognition, and reduce the computational complexity.* The clustered data was obtained by using K-means clustering algorithm, and then the key frames were extracted by using the clustered data. Two features for human action recognition were extracted, one is the feature of the position of human joint, another is the feature of the skeleton angle between rigid body and rigid body. The sequence of action video was classified through the SVM classifier. This method leads to a more accurate recognition rate, and the real-time capability has been improved at the same time according to the result showed on the data set MSR-DailyActivity3D.

**Keywords:** Kinect; human action recognition; key frames; k-means

## 引言

人体动作识别是机器视觉中很重要的一个研究方向，有很广泛的应用，如：视频监控、人机交互、虚拟现实等。人体动作识别的方法有基于模板、基于概率匹配、基于语义的等[1]。徐光祐[2]从行为的定义、运动特征的提取和动作表示以及行为理解几个方面对目前的工作做了分析和比较，并指出了这些工作面临的挑战和将来的研究方向。在以前的研究中，很多关于人体动作识别的研究方法都是基

收稿日期：2015-06-14 修回日期：2015-07-30；
基金项目：国家自然科学基金(61170185)；航空科学基金(2013ZC54011)；辽宁省博士启动基金(20121034)；辽宁省教育厅资助项目(L2014070)；
作者简介：石祥滨(1963-)，男，辽宁，博士，教授，研究方向为虚拟现实，图像处理，网络游戏。



于单目 RGB 视频[3]。但是，单目 RGB 视频数据的处理有很多的难点。比如：不具有视角不变性，对光照和背景的变化敏感，对噪声不鲁棒等。虽然近几年通过研究者的努力，人体动作识别方面取得了一些很有意义的成果，但是人体动作识别的研究仍然非常具有挑战性。比如：使用 RGB 视频数据获取 3D 空间数据比较困难。使用动作捕获系统虽然可以获取人体身上的 3D 坐标，但是这一系统的成本非常高，而且需要使用者穿戴一些会阻碍自然运动的动作捕捉器。随着低成本的深度传感器的出现，如微软的 Kinect，由于它可以获取图像中的深度信息和骨架信息，通过 Shotton[4]提出的方法，可以通过一张深度图像快速并且准确的估计人体骨架关节点的 3D 位置。所以近几年以骨架为基础的人体动作识别得到了关注。目前已存在的以骨架为基础的人体动作识别主要有两大类[5]，基于关节的方法和基于身体部分的方法。以关节点为基础的方法是将人体骨架用一系列的点来表示[6-7]。另一方面，以身体部分为基础的方法是将人体骨架表示为一系列刚体片段[5]。在动作执行过程中，根据环境的影响或每个动作执行者的习惯，动作的执行会各不相同，会存在动作执行速率不一致的问题。会造成同一个动作视频序列的长短不一致，对识别结果产生影响。Meinard Muller[6]利用动态时间归整(DTW)处理速率不一致的问题。不同于该方法，J. Wang[7]提出的采用傅里叶时间金字塔(FTP)方法表示时间模型。FTP 是描述性模型。它不涉及生成模型(HMM，CRF 和动态系统)中复杂的学习，并且它相比于 DTW，对于噪声和时间不对准更具有鲁棒性。

在人体动作视频序列中，提取的关键帧要能够反映视频序列中要表示的人体动作，因为视频是渐变的，所以帧与帧之间可能存在着冗余，这样会对人体动作识别的识别率产生不良影响。而基于关键帧的特征提取方法，能够去除冗余数据产生的影响，而且还能够减少用于人体动作识别的特征数据，故能够降低计算复杂度。提取的视频序列中关键帧的特征，能够更加准确、全面的对人体动作进行表示，提高识别的准确性和实时性。本文使用 K-均值聚类算法对人体骨架中关节点的 3D 坐标数据进行聚类，得到每一个关节点的聚类中心，提取关键帧，然后提取关键帧中的人体骨架特征。最后使用 SVM 分类器对其进行分类。

# 1 关键帧提取

关键帧即特征帧，即在一个动作视频序列中，能够反映该动作的有代表性的视频帧，可以利用从关键帧中提取的特征识别人体动作。考虑每一个动作序列的执行动作速率不一致问题，利用 K-均值聚类算法进行聚类，提取出相似的数据的聚类中心，然后进行关键帧的提取。

## 1.1 k-means 聚类算法

k-means 聚类算法将相似对象归入同一簇，将不相同的对象归到不同簇。利用该算法将一个视频序列中的每一帧中的骨架关节点的 3D 坐标进行聚类，每一帧中的 3D 坐标组成一个 60 维的向量，通过聚类可以得到 k 簇 60 维的向量。

假如样例为 $\{x^{(1)},x^{(2)}...x^{(N)}\}$，$x^{(i)}\in R^N$，$N$ 为一个视频中的总帧数，$i$ 表示动作视频序列长度的帧数为 $N$ 中的第 $i$ 帧，$x^{(i)}$ 为序列中第 $i$ 帧的 20 个关节点 3D 坐标位置的向量，$x^{(i)}$ 为 d 维的向量，d 在此为 60，$R^N$ 为一个动作视频序列中的每一帧中关节点的 3D 坐标数据组成的向量的集合。将样例数据(一个动作视频序列所有帧)以帧为单位，将帧中的关节点数据聚类成 K(K<=N)个簇(cluster)：

1、随机选取 K 个聚类质心(cluster centroids)为 $u_1,u_2...u_K\in R^n$。

2、重复下面过程直到收敛

{

对于每一个样本 $x^{(i)}$，计算计算样本到每一个质心的距离最小值，将其归入应属于的类

$$D=\arg\min\sum_{i=1}^{N}\sum_{j=1}^{K}\left\|x^{(i)}-u_j\right\|^2$$

对于每一个类 j，重新计算该类的质心

$$u_j = \frac{\sum_{i=1}^{N} r_{ij} x^{(i)}}{\sum_{i=1}^{N} r_{ij}}$$

}

K 为要聚类出的簇的个数，$D$ 表示聚类和样本中心的距离最小值，$D$ 最小时，表示归入 j 类。$r_{ij}$ 表示数据向量 $x^{(i)}$被归类为 j 时，为 1，否则为 0。$u_j$ 表示计算出的样本中心，为 60 维的向量。

## 1.2 关键帧提取

通过上面 k-means 聚类算法，提取每一个视频序列所有帧中的 k 个样本中心。提取出的 k 个样本中心，其中每一个样本中心为 60 维的向量，由视频帧中 20 个关节点的 3D 坐标聚类得到，其中每一个关节点的聚类中心，和相应视频序列中的视频帧中的 20 个 3D 关节点中对应的关节点坐标计算欧氏距离，寻找关键帧，如：关节点 1 聚类出的样本中心和每一帧中关节点 1 的 3D 坐标计算欧氏距离，依次类推，欧式距离最小的关节点记为 1，其他记为 0，比如视频中第一帧得到结果为{1,1,1,1,1,0,0,0,0,1,0,1,0,1,0,0,1,1,0,0}，最后，计算视频序列中得到 1 最多的帧，即为关键帧。

提取关键帧步骤：

选取数据集中一个动作的视频序列，该视频序列帧数 N，k 值为动作视频序列中要提取的关键帧的个数，首先利用 k-means 算法聚类，利用聚类出来的样本中心寻找关键帧。得到的 k 簇聚类出来的数据表示为：

$$M_i = \{(x_{i1}, y_{i1}, z_{i1}), (x_{i2}, y_{i2}, z_{i2}) \ldots (x_{ij}, y_{ij}, z_{ij})\},$$
$$i \in (1,2,3 \ldots k),\ j = 20$$

i 为聚类出来的第 i 簇。j 为每帧中关节点的个数。此处用$(x_{i1},y_{i1},z_{i1})$表示所有帧的第 1 关节点聚类出的样本中心。一个视频序列中的一帧中的骨架关节点的 3D 坐标表示如下：

$$F_m = \{(x_{m1}, y_{m1}, z_{m1}), (x_{m2}, y_{m2}, z_{m2}) \ldots$$
$$(x_{mj}, y_{mj}, z_{mj})\},\ m \in (1,2,3 \ldots N),\ j = 20$$

m 为一个视频序列中第 m 帧，N 为视频序列中的总帧数，$(X_{m1}, Y_{m1}, Z_{m1})$ 表示视频序列中第 m 帧的第 1 个关节点位置，以此类推… $M_i$ 和 $F_m$ 均为 60 维向量，将聚类出的每个关节点样本中心和视频序列帧中的相应关节点的 3D 坐标数据之间计算欧氏距离，表示如下：

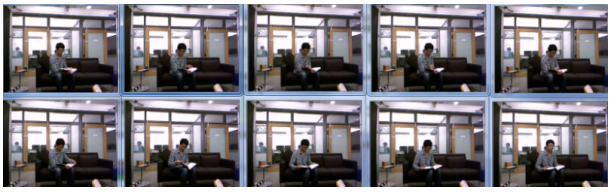
$$C = \mathrm{Min}(M_i, F_m) =$$
$$\{\mathrm{dis}|(x_{i1}, y_{i1}, z_{i1}), (x_{m1}, y_{m1}, z_{m1})|,$$
$$\mathrm{dis}|(x_{i2}, y_{i2}, z_{i2}), (x_{m2}, y_{m2}, z_{m2})| \ldots$$
$$\mathrm{dis}|(x_{ij}, y_{ij}, z_{ij}), (x_{mj}, y_{mj}, z_{mj})|\},$$
$$m \in (1,2,3 \ldots N),\ j = 20$$

$\mathrm{dis}|(x_{ij}, y_{ij}, z_{ij}), (x_{mj}, y_{mj}, z_{mj})|$ 表示利用 k-means 算法聚类出的关节点 j 的聚类中心和第 m 帧中的相应的第 j 关节点坐标的欧氏距离，对所有帧计算距离后，$\mathrm{dis}|(x_{ij}, y_{ij}, z_{ij}), (x_{mj}, y_{mj}, z_{mj})|$最小的记为 1，否则记为 0。为了保证提取的关键帧和视频序列中帧的先后顺序一致，计算过程为，使用每个视频序列中聚类出的关于 3D 关节点位置样本中心的 k 个向量，每个向量中关节点的聚类中心分别和视频序列中每一帧中的相应的 3D 关节点坐标计算欧氏距离。Min 表示样本中心和关节点的距离最小，每帧的 20 个关节点中，距离最小的关节点的个数即为 C，计算出 C 值最大的视频帧，并保存得到该 C 值的帧的索引，最后按索引值排序，即得到该视频序列关键帧。使用上面的方法提取的关键帧，可以保证提取出来的关键帧的顺序和视频中帧的顺序一致。而且对于所有帧中的骨架数据都相似的视频动作序列，如图 1 中的看书动作，所有帧的结果几乎相同，可能会存在提取出的关键帧的顺序和视频中的顺序不一致的情况出现。但是由于所有帧之间的相似度比较大，所以对识别效果不会产生太大的影响。

利用上述方法对日常人体动作数据集 DailyActivity3D dataset 数据集中的动作：喝水、站起来和看书，提取关键帧，设置 k=10 时提取关键帧如图 1 所示。

(a) 喝水

(b) 站起来



(c) 看书

图 1 视频序列中关键帧提取

聚类的作用就是将相似对象归入同一簇，这就避免了将很多相似对象作为特征进行分类，即选择的特征都是能够显著表示该视频动作序列的显著性特征，避免了数据冗余的发生，并提高了运算速度，而且提出了一种自动计算 k 值的方法，通过实验表明，一个动作视频序列的帧数和提取的关键帧数之间的关系，和随着帧数的增长，识别时间间的关系，如图 2 所示。

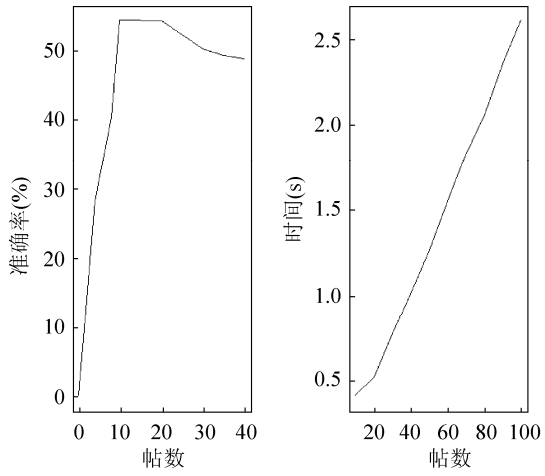



图 2 时间和准确率随帧数增加变化图

从图 2 可以看出，利用每一个视频序列中所有帧数的十分之一到二十分之一作为关键帧，取得的识别效果最好，且识别速度很快，这样可以根据要处理视频的帧数决定要提取关键帧的个数，避免了每次都重新设置 k 值。对视频序列的长短具有鲁棒性。

## 2 特征提取

提取每一个动作视频序列的关键帧后，要从关键帧中的 3D 关节点坐标数据中提取特征，利用所提取的特征进行人体动作识别。将关节点坐标数据和人体刚体部分之间的骨架角度作为特征用于人体动作识别，骨架表示及关节点索引，如图 3 所示。两种特征提取方法如下：

1) 关节点位置

通过 Kinect 获得的人体骨架的信息，骨架关节点坐标可以反映关节点位置，所以可以用骨架的关节点坐标作为特征。对于每一个关节点，都可以得到三个信息：一是这关节点的索引。每一个关节点有一个单一的索引值。二是每一个关节点的位置。这三个坐标用米表示。x，y 和 z 轴是在深度传感器的主体轴。这是一个右手坐标系，传感器所在位置为原点。z 轴方向是传感器阵列点的方向。y 轴的方向是向上，x 轴的方向是向左(相对于传感器阵列)，如图 4 所示。三是关节点的状态，如果 Kinect 能够跟踪到关节点，状态就设置为跟踪。如果关节点不能够被跟踪，算法就会根据其他关节点的位置试着推断该关节点的位置。如果可能，关节点的状态被推断出来，否则，这个关节点的状态为未跟踪。关节点可以提高如此丰富的信息，所以使用关节点位置作为人体动作识别的特征。每一个关节点为(x，y，z)三维坐标组成，每一帧图像提取 20 个关节点，所以一帧图像就可以得到一个 60 维的特征向量，如关节点 i 的 3D 坐标为 $(x_i, y_i, z_i)$，$i \in \{A, B...T\}$。所以每一帧图像得到的特征向量表示如下：

$$joint\_position = \{(x_A, y_A, z_A), (x_B, y_B, z_B)...(x_T, y_T, z_T)\}.$$

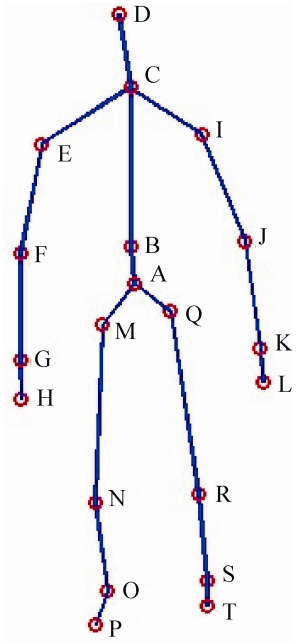



图 3 骨架表示

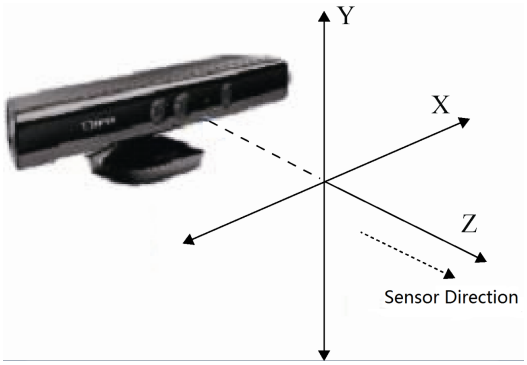



图 4 关节点位置坐标系

2) 骨架角度

利用提取的关节点 3D 坐标计算人体刚体部分之间的角度作为人体动作识别的特征，从一帧图像的关节位置中计算出的 20 个角度组成的特征向量如下表示：

$$\text{joint\_angle}=\{(\overrightarrow{CD},\overrightarrow{CI}),(\overrightarrow{CD},\overrightarrow{CE}),(\overrightarrow{CE},\overrightarrow{CB}),(\overrightarrow{CB},\overrightarrow{CI}),(\overrightarrow{EC},\overrightarrow{EF}),(\overrightarrow{FE},\overrightarrow{FG}),(\overrightarrow{GF},\overrightarrow{GH}),(\overrightarrow{AB},\overrightarrow{AQ}),(\overrightarrow{AB},\overrightarrow{AM}),(\overrightarrow{AM},\overrightarrow{AQ}),(\overrightarrow{QA},\overrightarrow{QK}),(\overrightarrow{MA},\overrightarrow{MN}),(\overrightarrow{NM},\overrightarrow{NO}),(\overrightarrow{ON},\overrightarrow{OP}),(\overrightarrow{KQ},\overrightarrow{KS}),(\overrightarrow{SK},\overrightarrow{ST}),(\overrightarrow{IC},\overrightarrow{IJ}),(\overrightarrow{JI},\overrightarrow{JK}),(\overrightarrow{KJ},\overrightarrow{KL}),(\overrightarrow{BC},\overrightarrow{BA})\}\ .$$

角度计算公式如下：

假如肩膀、肘和手的关节点坐标分别为 $E(x_E,y_E,z_E)$、$\mathrm{F}(x_F,y_F,z_\mathrm{F})$ 和 $\mathrm{G}(x_G,y_\mathrm{G},z_\mathrm{G})$ 表示。则前臂和后臂两个刚体由向量 $\overrightarrow{FE},\overrightarrow{\mathrm{FG}}$ 表示为：

$$\overrightarrow{FE}=(x_E-x_F,y_E-y_F,z_E-z_F),$$

$$\overrightarrow{\mathrm{FG}}=(x_E-x_G,y_E-y_G,z_E-z_G)$$

假设计算向量 $\vec{a}(\mathrm{a}_1,\ \mathrm{a}_2,\ \mathrm{a}_3)$ 和向量 $\vec{\mathrm{b}}(b_1,\ b_2,\ b_3)$ 量的夹角，根据两向量间的角度计算公式，如下所示：

$$\cos\left\langle\vec{a},\vec{b}\right\rangle=\frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\cdot|\vec{b}|}=\frac{a_1b_1+a_2b_2+a_3b_3}{\sqrt{a_1^2+a_2^2+a_3^2}\cdot\sqrt{b_1^2+b_2^2+b_3^2}}$$

利用上述的角度计算公式，得出向量 $\overrightarrow{FE}$和$\overrightarrow{\mathrm{FG}}$ 组成的角度可以表示为 $\cos\left\langle\overrightarrow{FE},\overrightarrow{\mathrm{FG}}\right\rangle$。

## 3 人体动作识别

使用 SVM 分类器对人体动作进行训练和识别，支持向量机 SVM (Support Vector Machine)是一种有监督学习算法，SVM 分类器可以解决高维度数据的分类。SVM 由于其很高的准确性和对于高维数据的分类能力，能够生成非线性的高维分类器，所以被选用作分类。

假定训练样本($x^{(i)}$，$y^{(i)}$)，$x^{(i)}$是特征，$y$ 是类别标签，$i$ 表示第 $i$ 个样本，SVM 使用支持向量分类(C-SVC)算法找到 SVM 的最优超平面：

$$\mathrm{f}(x)=w^Tx^{(i)}+b$$

$x^{(i)}$为向量形式，即多特征形式，为了将训练数据分开，在非线性情况下软间隔定义如下：

$$\min_{\gamma,w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t. } y^{(i)}(w^Tx^{(i)}+b)\geqslant 1-\xi_i,i=1,2,\ldots m\ \xi_i\geqslant 0$$

$\xi$ 是松弛向量，为非负参数，目标函数后面加上 $C\sum_{i=1}^{m}\xi_i$ 就表示离群点越多，目标函数值越大，而我们要求的是尽可能小的目标函数值。这里的 C

是离群点的权重，C 越大表明离群点对目标函数影响越大，目标函数控制了离群点的数目和程度，使大部分样本点仍然遵守限制条件。拉格朗日公式定义为：

$$L(w,b,\xi,\alpha,\gamma)\frac{1}{2}w^T w+$$

$$C\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}\alpha_i[y^{(i)}(x_i^T w+b)-1+\xi_i]-\sum_{i=1}^{m}\gamma_i\xi_i$$

此处的 $\alpha_i$ 和 $\gamma_i$ 为拉格朗日乘子，经过求导、KKT 条件约束等一系列操作得出：$w=\sum_{i}^{n}a_i y_i x_i$ n 为支持向量的个数，将上式子代入拉格朗日公式，可得极大值：

$$L=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j\theta(x^{(i)})^T\theta(x^{(j)})$$

$$s.t.\ 0\le\alpha_i\le C, i=1\ldots m, j=1\ldots m, \sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

为了提高计算效率，将 $\theta(x^{(i)})^T\theta(x^{(j)}$ 映射到核函数 $k(x_{(i)},x^{(j)})$

在此实验中用的核函数为径向基函数(简称 RBF)核函数，又称为高斯核函数：

$$K(x,z)=\exp(-\frac{\|x-z\|^2}{2\sigma^2})$$

最后得到的超平面为：

$$f(x)=\sum_{i=1}^{m}\alpha_i y^{(i)}\left\langle x^{(i)},x^{(j)}\right\rangle+b$$

## 4 实验结果与分析

实验中，使用 k-means 聚类算法聚类出要提取的关键帧的个数 k，然后通过 2.2 中关键帧提取办法提取出视频序列中能够表示人体动作的关键帧。之后，使用 3 中的特征用来表示视频序列中的人体动作，最后使用 SVM 分类器进行人体动作的分类和识别。设置 libsvm[1]中 svm 类型为 C_SVC，核类型是 rbf(radial basis function)，同时，对数据进行了缩放，缩放范围是[-1,1]。使用交叉验证的方式对所用方法的正确率进行了评估。实验中利用两种特征(关节位置、刚体之间角度)，最后结果表明使用该关键帧提取方法，可以提高人体动作识别率。在 MSR-DailyActivity3D[9]数据集上进行了验证。

MSR-DailyActivity3D 数据集是一个由 Kinect 设备获得的日常动作数据集。该数据集有 16 个动作类型：drink, eat, read book, call cellphone, write on a paper, use lap-top, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down.有十个人，每个人执行每个动作两次。有 320 个动作样本，一共 3×320 个文件，分别记录了 RGB、深度、和骨架。一些动作示例如图 5 所示。

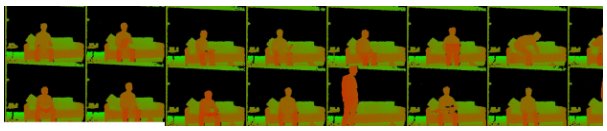


图 5 MSR-DailyActivity3D 数据集动作

实验结果如下：

表 1 展示了使用关节点位置坐标和人体刚体部分之间的骨架角度作为特征时，两种方法对识别率的影响。

表 1 识别率比较

| Method | Acc. (%) |
|---|---|
| Only Joint Position features | 0.49 |
| Joint angle features | 0.55 |
| k-means key frame Joint Position | 0.54 |
| k-means key frame Joint angle | 0.62 |

表 2 显示其他一些动作识别方法的在 MSR-DailyActivity3D 数据集上的识别率，和使用关节帧的方法进行比较。可以看出，使用关键帧进行的人体动作识别的方法的识别率要比使用其他一些方法的识别率要好。

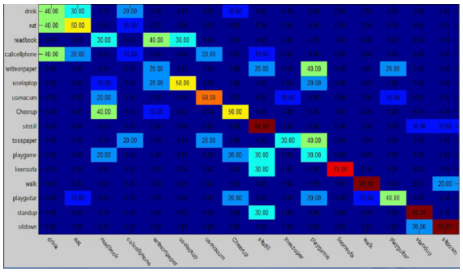
表 2 MSR-DailyActivity3D 数据集中的识别率比较

| Method | Acc. (%) |
|---|---|
| Dynamic Temporal Warping [9] | 0.54 |
| Only LOP features[5] | 0.43 |
| k-means key frame Joint Position | 0.54 |
| k-means key frame Joint angle | 0.62 |

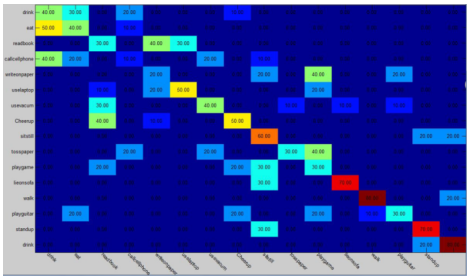


由以上实验结果得出，观察表 1 可以发现，使用骨架刚体之间的角度作为特征，要比使用关节点位置特征得到的准确率要高出 6%。因为关节点位置会因为人的不同，提取的坐标位置各不相同，因为每个人的骨架刚体是不同的，如前臂和大腿等骨架刚体的长度会因为人的不同而不同，所以提取的关节点位置不同，而使用骨架刚体角度会避免这一问题产生的影响。同时，在表 1 中，对于同一种特征，使用关键帧提取方法，使用关节点位置和人体刚体部分之间的骨架角度作为特征时，识别率比不使用该方法分别提高了 6%和 8%。观察表 2 中的数据，在 MSR-DailyActivity3D dataset 数据集中，相比于其他论文中的方法，使用关键帧的方法，要比 J. Wang 中[5]使用的方法，人体刚体部分之间的骨架角度作为特征时，关键帧的方法要高出 19%。关节点位置作为特征，关键帧的方法要比 J. Wang 中[7]高出 11%，而人体刚体部分之间的骨架角度作为特征，关键帧的方法要比 M. Muller [11]中高出 8%。使用文章中提出的关键帧的方法，使用人体骨架刚体部分之间的角度特征，得到的 MSR- DailyActivity3D 数据集的混淆矩阵如图 6(a)所示，不使用文中方法得到的混淆矩阵如图 6(b)所示。



(a) 关键帧方法产生的混淆矩阵



(b) 非关键帧方法产生的混淆矩阵

图 6 MSR-DailyActivity3D 数据集动作识别混淆矩阵

表 3 比较了 4 种人体动作识别方法的识别时间比较。

表 3 动作识别时间比较

| Method | Acc. (s) |
|---|---|
| Only Joint Position features | 2.418 |
| Joint angle features | 0.887 |
| k-means key frame Joint Position | 0.316 |
| k-means key frame Joint angle | 0.158 |

由表中实验结果表明，使用关键帧的方法，提取骨架角度作为特征进行人体动作识别耗费的时间最短。相比于 Vemulapalli[5]的方法，由于计算复杂，运行一次需要大约 5.5 小时，识别速度要大大提高，但是识别率相比还有待改进。

## 5 结论

提出了一种基于视频关键帧的人体动作识别方法。通过使用 k-means 聚类算法，对人体动作视频序列中的骨架关节点坐标数据进行聚类，之后利用聚类出的数据，提取出能够表示视频序列中人体动作的关键帧，利用该方法提取出的关键帧，从关键帧的人体骨架中的关节点 3D 坐标数据中提取特征，最后，使用 SVM 分类器对其进行分类和识别。该方法减少了不同帧之间坐标数据产生的冗余，并且减少了用于动作识别的特征的数据量，可以提高计算速度和识别精度，所以对识别率会产生好的影响。

## 参考文献：

[1] 胡琼, 秦磊, 黄庆明. 基于视觉的人体动作识别综述 [J]. 计算机学报, 2013, 36(12): 2512-2524.

[2] 徐光祐, 曹媛媛. 动作识别与行为理解综述 [J]. 中国图象图形学报, 2009, 14(2): 189-195.

[3] J K Aggarwal, M S Ryoo. Human Activity Analysis: A Review [J]. ACM Computing Surveys(CSUR), 2011, 43(3): 16:1-16:43.

[4] Shotton J, Sharp T, Kipman A, et al. Real-time human pose recognition in parts from single depth images[J]. Communications of the ACM, 2013, 56(1): 116-124.

[5] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3d skeletons as points in a lie group [C]// Computer Vision and Pattern Recognition

(CVPR), 2014 IEEE Conference on. USA: IEEE, 2014: 588-595.

[6] Meinard Müller. Information Retrieval for Music and Motion [M]. USA: Springer-Verlag New York, Inc., 2007.

[7] Wang J, Liu Z, Wu Y, et al. Mining actionlet ensemble for action recognition with depth cameras[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 1290-1297.

[8] Hussein M E, Torki M, Gowayyed M A, et al. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations[C] //Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013: 2466-2472.

[9] Thi-Lan Le, Minh-Quoc Nguyen. Human posture recognition using human skeleton provided by Kinect [C]// Computing, Management and Telecommunications (ComManTel), 2013 International Conference on. USA: IEEE, 2013: 340-345.

[10] Liu T, Song Y, Gu Y, *et al*. Human Action Recognition Based on Depth Images from Microsoft Kinect [C]// Intelligent Systems (GCIS), 2013 Fourth Global Congress on. USA: IEEE, 2013: 200-204.

[11] Müller M, Röder T. Motion templates for automatic classification and retrieval of motion capture data[C]//Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation. Eurographics Association, 2006: 137-146.

[12] Chih-Chung C, L Chih-Jen. LIBSVM: A library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 1-27.

[13] Le T L, Nguyen M Q, Nguyen T T M. Human posture recognition using human skeleton provided by Kinect [C]// Computing, Management and Telecommunications (ComManTel), 2013 International Conference on. USA: IEEE, 2013: 340-345.

(上接第 2400 页)

[11] D Zhao, X Wang, T Fang, C Wang. Research on occupant evacuation in a high-rise dormitory building based on building EXODUS [J]. Journal of Applied Fire Science (S1044-4300), 2014, 23(2): 249-268.

[12] V S Kalogeiton, D P Papadopoulos, I P Georgilas, *et al*. Cellular automaton model of crowd evacuation inspired by slime mould [J]. International Journal of General Systems (S0308-1079), 2015, 44(3): 354-391.

[13] W Zeng, H Nakamura, P Chen. A modified social force model for pedestrian behavior simulation at signalized crosswalks [C]// Proceeding of the 9th International Conference on Traffic and Transportation Studies. Beijing, China: Elsevier, 2014, 138: 521-530.

[14] X Yang, H Dong, Q Wang, *et al*. Guided crowd dynamics via modified social force model [J]. Physica A: Statistical Mechanics and its Applications (S0378-4371), 2014, 411(10): 63-73.

[15] F Martinez-Gil, M Lozano, F Fernández. MARL-Ped: A multi-agent reinforcement learning based framework to simulate pedestrian groups [J]. Simulation Modelling Practice and Theory (S1569-190X), 2014, 47: 259-275.

[16] 李本先, 迟妍. 突发事件过程中暴动人群演化模型与仿真 [J]. 重庆理工大学学报(自然科学版), 2014, 28(6): 81-87. (Li Benxian, CHI Yan. Emergency Incidents during the Riots Burst Evolution Model and Simulation of Crowd [J]. Journal of Chongqing University of Technology(Nature Science) (S1674-8425), 2014, 28(6): 81-87.)

[17] Helbing D, Farkas I, Vicsek T. Simulating dynamical features of escape panic [J]. Nature (S0028-0836), 2000, 407(6803): 487-490.

