人体运动视频关键帧优化及行为识别

赵 洪， 宣士斌

(广西民族大学信息科学与工程学院，广西 南宁 530006)

摘 要：在行为识别过程中，提取视频关键帧可以有效减少视频索引的数据量，从而提高动作识别的准确性和实时性。为提高关键帧的代表性，提出一种关键帧序列优化方法，并在此基础上进行行为识别。首先根据 3D 人体骨架特征利用 K-均值聚类算法提取人体运动视频序列中的关键帧，然后根据关键帧所在序列中的位置进行二次优化以提取最优关键帧，解决了传统方法中关键帧序列冗余等问题。最后根据最优关键帧利用卷积神经网络(CNN)分类器对行为视频进行识别。在 Florence3D-Action 数据库上的实验结果表明，该方法具有较高的识别率，并且与传统方法相比大幅度缩短了识别时间。

关 键 词：行为识别；关键帧；K-均值；卷积神经网络
中图分类号： TP 399 **DOI：**10.11996/JG.j.2095-302X.2018030463
文献标识码： A **文 章 编 号：**2095-302X(2018)03-0463-07

Optimization and Behavior Identification of Keyframes in Human Action Video

ZHAO Hong, XUAN Shibin

(School of Information Science and Engineering, Guangxi University for Nationalities, Nanning Guangxi 530006, China)

Abstract: In the course of behavior identification, extracting keyframes from the video can effectively reduce the amount of video index data, so as to improve the accuracy and real-time performance of behavior identification. A method for optimizing the keyframe sequence is proposed to improve the representativeness of keyframes, on which the behavior identification is based. Firstly, the K-means clustering algorithm is employed to extract keyframes in the human action video sequence according to 3D human skeleton features. Then, the quadratic optimization is performed in the light of the location of keyframes to extract the optimal keyframe, and it can reduce the redundancy of keyframe sequence, compared with traditional ways. Finally, the behavior video is identified by convolutional neural network (CNN) classifiers in accordance with the optimal keyframe. The experiment results on the Florence 3D Action dataset indicate that the method has a high identification rate, and drastically shortens the identification time, compared with the traditional method.

Keywords: behavior identification; keyframes; K-means; convolutional neural network

人体行为识别是近年来计算机视觉领域的一个研究热点，广泛应用于人机智能交互、视屏监控、虚拟现实等领域[1]。随着多媒体技术和网络信息的飞速发展，视频数据大量充斥在我们周边，如何在规定的时间内从大量视频数据中检索出有效的、关键的信息进行应用是当前一个急需解决的关键问题。关键帧则是反映镜头主要内容的一帧或者若干帧图像，不仅可以简单、概括的描述视频主要视觉

收稿日期：2017-07-18；**定稿日期：**2017-09-01
基金项目：广西自然科学基金项目(2015GXNSFAA139311)
第一作者：赵 洪(1991–)，女，山东济南人，硕士研究生。主要研究方向为视频图像处理及行为识别。E-mail：15777169369@163.com
通信作者：宣士斌(1964–)，男，广西南宁人，教授，博士。主要研究方向为图像处理、模式识别。E-mail：xuanshibin@mail.gxun.cn

内容，而且相比于原始视频中所含图像帧的数目，关键帧的使用可以大幅度减少视频索引的数据量，为后期的应用提供了很好的数据预处理作用。目前，关键帧提取技术主要包括以下 4 类：①基于镜头边界法[2]。该方法通常提取镜头固定位置上的帧作为关键帧，例如首帧、中间帧或尾帧。此类方法简单易行，但提取的关键帧有时因为视频数据的类型不能很好地反映镜头内容。②基于视觉内容分析法[3-4]。该方法将视频内容变化程度作为选择关键帧的标准，但当有镜头运动时，此类方法容易选取过多的关键帧，造成数据冗余并且所提关键帧不一定具有代表性。③基于运动分析法[5-6]。该方法通过计算镜头中的运动量，在运动量达到局部最小值处选取关键帧，该方法能很好地表达视频内的全局性运动，但计算量较大，耗时较长。④基于聚类的方法[7-9]。该方法在预先设定好聚类数目的前提下提取的关键帧能够很好地表达视频主要内容，提取关键帧的数量也可以根据视频内容和种类来动态确定，此类方法已经成为目前主流的关键帧提取方法。但这些方法提取的关键帧往往存在大量冗余，为此本文在由 K-均值聚类的方法提取的初始视频关键帧的基础上，提取距离每个聚类中心最近的帧作为关键帧,构造初始关键帧序列然后根据关键帧帧间位置对初始关键帧序列进行二次优化，提高关键帧质量，消减冗余信息构建最优关键帧序列，最后利用 CNN 在 Florence3D-Action 数据库上进行识别实验。

1 运动特征表示

在人体行为识别中，利用 Kinect 获取 3D 骨架信息，可以有效避免物体遮挡或者重叠问题，并很好地适应环境的变化，具有很好的鲁棒性。而在实际运动中，人体主要部位的骨骼运动对动作识别结果起到决定性的作用，细节骨骼运动对人体的整体运动起到的影响有限，因此，采用文献[10]中的 15 个主要关节点的骨架模型，骨架表示及关节点索引如图 1 所示。选取髋关节(O 点)为根节点即局部坐标系原点，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。

1.1 关节点位置

本文使用 15 个主要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 15 个关节点，所以一帧图像就可以得到一个 45 维的特征向量，如关节点 A 的 3D 坐标为(x_A,y_A,z_A),每一帧图像得到的 45 维特征向量表示为

$$\text{joint\_position}=\left\{\left(x_A,y_A,z_A\right),\left(x_B,y_B,z_B\right)\cdots\left(x_O,y_O,z_O\right)\right\}$$

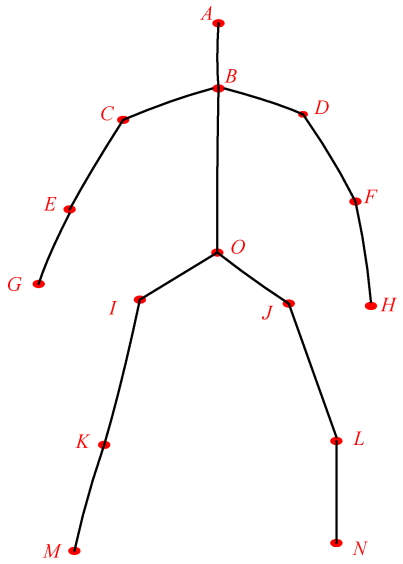

图 1 骨架表示及关节索引

1.2 角度信息

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 15 个角度组成的特征向量[11]为

$$\text{Joint\_angle}=\begin{Bmatrix}(\overrightarrow{BA},\overrightarrow{BD}),(\overrightarrow{BA},\overrightarrow{BC}),(\overrightarrow{BC},\overrightarrow{BO}),\\(\overrightarrow{BO},\overrightarrow{BD}),(\overrightarrow{CB},\overrightarrow{CE}),(\overrightarrow{EC},\overrightarrow{EG}),\\(\overrightarrow{OB},\overrightarrow{OJ}),(\overrightarrow{OI},\overrightarrow{OB}),(\overrightarrow{OI},\overrightarrow{OJ}),\\(\overrightarrow{IO},\overrightarrow{OK}),(\overrightarrow{KI},\overrightarrow{KM}),(\overrightarrow{JO},\overrightarrow{JL}),\\(\overrightarrow{LJ},\overrightarrow{LN}),(\overrightarrow{DB},\overrightarrow{DF}),(\overrightarrow{FD},\overrightarrow{FH})\end{Bmatrix}$$

角度计算公式中，肩膀、手肘、手部的关节点坐标分别用$C\left(x_C,y_C,z_C\right)$、$E\left(x_E,y_E,z_E\right)$、$G\left(x_G,y_G,z_G\right)$表示；上臂和下臂两个刚体则由向量$\overrightarrow{EC}$和$\overrightarrow{EG}$表示，即

$$\overrightarrow{EC}=\left(x_C-x_E,y_C-y_E,z_C-z_E\right)=(a_1,a_2,a_3)$$

$$\overrightarrow{EG}=\left(x_G-x_E,y_G-y_E,z_G-z_E\right)=(b_1,b_2,b_3)$$

根据向量夹角公式可得向量$\overrightarrow{EC}$和$\overrightarrow{EG}$之间的夹角$\cos\langle\overrightarrow{EC},\overrightarrow{EG}\rangle$为

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

2　关键帧提取

关键帧即特征帧，是在一个动作视频序列中可以概括反映该动作的视频帧，需要体现动作视频中具有代表意义的关键姿态。有效的关键帧序列意味着可以代表性的表示该行为，最大限度的使该行为区别于其他类型的行为，同时减少数据存储空间的使用。在动作识别过程中可以利用从关键帧中提取的特征识别人体动作，考虑每一个动作执行动作速率不一致问题，本文利用 K-均值聚类算法进行聚类，提取出相似数据的聚类中心，然后进行关键帧的提取。

2.1　K-means 聚类算法

K-means 算法[12]是最典型、最常用的基于划分的聚类算法，其基本思想是将相似对象归入同一簇，将不相同的对象归到不同簇。本文利用该算法将一个人体运动视频序列中的每一帧中的骨架关节点 3D 坐标进行聚类，每一帧中的 3D 坐标组成一个 45 维的向量，通过聚类可以得到 K 簇 45 维的向量。假如视频序列为 $\left\{x^{(1)},x^{(2)},\cdots,x^{(N)}\right\}$，$\boldsymbol{x}^{(i)}\in R^N$，其中，$N$ 为视频序列总帧数；i 为视频中的第 i 帧；$\boldsymbol{x}^{(i)}$ 为序列中第 i 帧的 15 个关节点 3D 坐标位置的向量；$\boldsymbol{x}^{(i)}$ 为 m 维向量，m=45。本文将视频中的关节点聚类成 $K(K\leqslant N)$个簇。算法具体过程如下：

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

(2) 计算每一个样本 $\boldsymbol{x}^{(i)}$ 到每一个聚类质心的距离最小值，D 取最小值时，表示将其归入 j 类，即

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

(3) 对于每一个 j 类，重新计算该类的质心 $u_j=\dfrac{\sum_{i=1}^{N}r_{ij}\boldsymbol{x}^{(i)}}{\sum_{i=1}^{N}r_{ij}}$。

(4) 重复步骤(2)、(3)直到函数收敛。

在进行聚类前，K-means 需要指定聚类个数，且初始聚类中心选取具有随机性，所以实验中提取 K=8、K=10、K=12 时的关键帧。以 Florence3D-Action 数据集中的动作：wave、drink、sit down 为例，K=10 时提取关键帧如图 2 所示，其中图 2(a) “挥手”序列关键帧从左至右依次为：1 帧、5 帧、9 帧、15 帧、19 帧、22 帧、24 帧、26 帧、29 帧、30 帧；图 2(b) “坐下”序列关键帧从左至右依次为：1 帧、3 帧、9 帧、11 帧、14 帧、15 帧、18 帧、22 帧、27 帧、29 帧。

(a) 挥手

(b) 坐下

图 2　视频序列关键帧提取

2.2 二次优化关键帧

从图 2 发现初次提取的关键帧有大量的重复，对比这些重复的关键帧，可以发现有些是因为动作运动过快，有些则是由于动作过于缓慢，最终导致相似的两帧相似度变小，误判为关键帧，例如在挥手关键帧序列中 22 帧与 24 帧；坐下关键帧序列中 14 帧与 15 帧。另外还可以看出重复的关键帧在视频镜头中的位置序列比较近，因此本文提出基于视频帧间隔的二次提取关键帧的方法，对初次聚类得到的关键帧进行二次提取，优化关键帧序列，具体方法如下：

(1) 假设第一次提取的关键帧数量为 M，定义 f 为视频的帧率；μ 为帧率倍数系数，则帧间隔阈值 $d=\mu f$ 。依次记录初次提取的每一关键帧在视频中的位置序列号，记为 p，并得到视频关键帧位置序号数组 $P=\{p_i|i=1,2,\cdots,M\}$。

(2) 依次计算相邻两关键帧的位置序列号差值，得到含有 $M-1$ 个数值的数组 $A=\{a_i|i=1,2,\cdots,M-1\}$。

(3) 将数组 A 中的值依次与间隔阈值 d 比较，若 $a_i<d$，则将位置序号为 p_{i+1} (相邻两冗余关键帧的后一帧)的关键帧去除，否则都保留。

对聚类中心为 10 时初次提取的视频关键帧进行二次提取优化，其中，帧率倍数系数 μ 设定为 1.5，以 Florence3D-Action 数据集中 wave、sit down 视频序列为例，视频帧率 f 为 20 帧/秒，则帧间隔阈值 $d=3$。

①记录初次提取的关键帧在视频中的位置可得序列号数组 P，即

$$P_{\text{wave}}=\{p_1=1、p_2=5、p_3=9、p_4=15、p_5=19、p_6=22、p_7=24、p_8=26、p_9=29、p_{10}=30\}$$

$$P_{\text{sit down}}=\{p_1=1、p_2=3、p_3=9、p_4=11、p_5=14、p_6=15、p_7=18、p_8=22、p_9=27、p_{10}=29\}$$

②依次计算相邻两关键帧的位置序列号差值，得到数组 A，即

$$A_{\text{wave}}=\{a_1=4、a_2=4、a_3=6、a_4=4、a_5=3、a_6=2、a_7=2、a_8=3、a_9=1\}$$

$$A_{\text{sit down}}=\{a_1=2、a_2=6、a_3=2、a_4=3、a_5=1、a_6=3、a_7=4、a_8=5、a_9=2\}$$

③将数组 A_{wave} 与 $A_{\text{sit down}}$ 中的值依次与帧间隔阈值 $d=3$ 比较，例如在 A_{wave} 数组中 $a_6<3$，则将位置序号为 p_7 的关键帧去除，即把初始关键帧序列中第 24 帧去除。

④最终得到的最优后的关键帧序列(图 3)，其中图 3(a) “挥手”序列关键帧从左至右依次为：1 帧、5 帧、9 帧、15 帧、19 帧、22 帧、29 帧，共 7 帧；图 3(b) “坐下”序列关键帧从左至右依次为：1 帧、9 帧、14 帧、18 帧、22 帧、27 帧，共 6 帧。

(a) 挥手

(b) 坐下

图 3 二次优化后的关键帧

利用上述方法对视频初次提取的关键帧进行二次优化，从图 3 中可以看出，优化后的关键帧并没有丢失运动序列的重要信息，所提取的关键帧也很好地反映了视频的内容，并且有效地去除了首次提取关键帧的冗余，达到优化目的。在方法中间隔阈值采用 $d=\mu f$，对于不同视频帧率的视频镜头具有一定的自适应性。

3 行为识别

卷积神经网络(convolutional neural network, CNN)[13]最先应用到手写识别，后来广泛应于模式识别各领域，共有 3 种类型的层：卷积层、下采样

层和全连接层。全连接层的连接方式与以往的神经网络连接方式相同，即一个神经元连接上一层所有的输出。卷积层的输出是通过一些核来卷积上一层的输入得到的，卷积操作公式为

$$y_{mn}=f\left(\sum_{j=0}^{J-1}\sum_{i=0}^{I-1}\boldsymbol{x}_{m+i,n+j}w_{ij}+b\right)\left(0\leqslant m<M,0\leqslant n<N\right)$$

其中，$\boldsymbol{x}$ 为二维输入向量；w 为尺寸 $J\times I$ 卷积核；b 为偏置；y 为尺寸 $M\times N$ 的输出；函数 f 为激活函数。一个卷积层有多个核去卷积上一层的输出，可得到多个新的二维输出特征图(feature map)。

下采样层就是对上一层得到的每一个特征图进行采样操作，使得特征图的尺寸减小。尺度为 $S_1\times S_2$ 的采样操作公式为

$$y_{mn}=\frac{1}{S_1S_2}\sum_{j=0}^{S_2-1}\sum_{i=0}^{S_2-1}\boldsymbol{x}_{m\times S_1+i,n\times S_2+j}$$

其中，$\boldsymbol{x}$ 为二维输入向量；y 为采样后得到的输出。

本文整个识别流程包括对数据的预处理、构造 3D 卷积神经网络和 softmax 分类训练过程，最终输出测试结果。算法流程如图 4 所示。在预处理阶段，将优化后的关键帧图像大小统一到120×160 像素中，得到统一大小尺度下的图像；然后分类标记图像信息，将统一大小后的图像均分为 5 份，1~4 份作为训练样本集使用，第 5 份作为测试样本集，得到标记后的图像；进行卷积和下采样操作，构造 3D 卷积神经网络；最后连接 softmax 分类器，实现多分类，Florence3D-Action 数据库中共含有 9 种动作类别，最终的输出神经元为 9，得到训练文件。将测试样本集在训练文件中进行测试，输出测试结果。

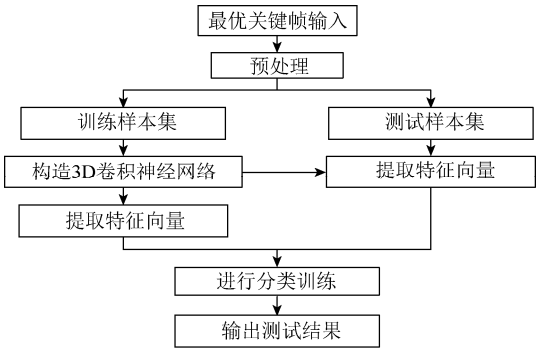

图 4　算法流程图

4　实验结果与分析

在实验中，使用 K-means 聚类算法聚类出原始关键帧序列，然后对初始关键帧序列进行二次优化，得到最优关键帧序列。最后使用 CNN 分类器进行人体动作的分类和识别。实验结果表明，对行为视频进行关键帧提取后，通过分析关键帧进行行为识别不但没有降低识别的效果，而且在识别时间上与直接对原始视频进行识别有大幅度的缩减。在 Florence3D-Action 数据集上进行了验证。

Florence3D-Action 数据集由一个固定的 Kinect 传感器获得，含有 10 个人执行的 9 个基本动作，即:挥手(wave)、喝水(drink)、接电话(answer phone)拍手(clap)、系鞋带(tight lace)、坐下(sit down)和站起来(stand up)、看手表(read watch)、弯腰(bow)共 215 个行为序列。

实验中选取数据集中 9 种动作视频序列，记录 K-means 聚类算法提取 K=8、K=10、K=12 时的关键帧、经过本文二次优化算法得到的关键帧以及消除冗余关键帧。实验结果见表 1。

表 1　关键帧提取及优化实验结果

视频序列	视频总帧数	K-means 算法选取关键帧数	经二次优化后关键帧数	消除冗余帧数
wave	30	8	4	4
drink	32	8	5	3
answer phone	26	8	4	4
clap	23	10	6	4
tight lace	33	10	7	3
sit down	34	10	6	4
stand up	33	12	7	5
read watch	33	12	8	4
bow	38	12	7	5

从表 1 可以看出，本文算法提取出的关键帧准确率高，冗余度小。经 K-means 聚类算法提取出的关键帧存在一定的冗余，但是通过对初始关键帧序列进行二次优化处理后，基本上消除了冗余帧，达到了预期优化目的。由于数据集中的视频序列的总帧数有限，初次提取的关键帧和二次优化的关键帧数目也有限。但随着视频总帧数的增加，消除冗余关键帧的效果会越来越明显。

表 2 展示了文献[6]、文献[14]、文献[15]、文献[16]以及本文算法在 Florence3D-Action 数据集上的实验结果。文献[6]和文献[14]对原始视频序列进行识别，平均识别率 88.0%和 94.5%，用本文算法提取关键帧后利用二次优化后的关键帧序列进行识别的平均识别率为 93.1%，在保证识别精度的前提下大幅度缩短了识别时间，提高了识别效率。相比文献[15]、文献[16]同样对关键帧序列进行识别，本文采用的二次优化后的关键帧序列识别的精度分别提高了 2.7%和 0.8%。实验结果表明，使用基于关键帧运动序列识别的方法，提取人体骨架角度特征进行分类识别所需的时间最短。相比于传统方法中直接对原始视频序列进行识别大大缩减了识别时间，在保证识别精度的前提下提高了识别效率。基于视频关键帧的识别在视频监控、网络视频数据库等大数据中有更突出的表现，可以大幅度的减少识别时间，减少人力物力的消耗。

表 2　各方法在 Florence3D-Action 数据集上的实验结果

算法	输入	ARA (%)	识别时间(s)
文献[6]	视频序列	88.0	3.723
文献[14]	视频序列	94.5	2.418
文献[15]	关键帧序列	90.4	0.276
文献[16]	关键帧序列	92.3	0.316
本文	关键帧序列(优化)	93.1	0.023

使用本文提出的基于关键帧序列的行为识别的方法，采用人体骨架刚体部分之间的角度特征，得到的 Florence3D-Action 数据集的混淆矩阵如图 5 所示，其中 drink 和 answer phone 这两个动作由于都是头面上的运动，并且手臂对头部也有一定的遮挡作用，使得识别过程中容易混淆。而 tight lace、sit down、stand up 和 bow 这些近似全身运动的动作具有很高的识别率，分别为 98%、98%、100%、99%。所以在今后的研究和改进中对混淆动作或者只调动局部肢体部分动作的识别是一项挑战性任务。

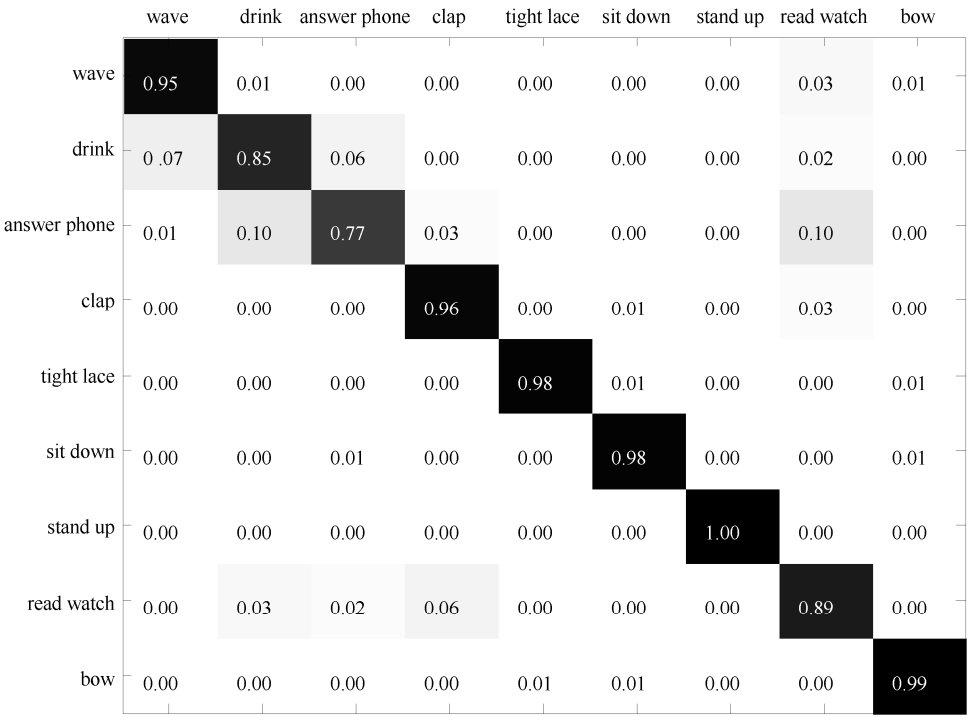

图 5　Florence3D-Action 数据集混淆矩阵

5 结 束 语

本文提出了一种基于视频关键帧序列的人体行为识别方法，主要思想是对原始视频运动序列聚类获取关键帧序列，再对初始关键帧序列进行二次优化，提高关键帧质量，获得最优关键帧序列。实验表明使用该方法提取的关键帧能较好地反映视频镜头的内容，利用卷积神经网络在Florence3D-Action数据库上的识别实验结果表明对视频关键帧序列进行识别在保证识别精度的前提下与传统方法相比提高了识别效率。

尽管实验结果达到了预期效果，但在以下方面还可以进行改进：①实验中只使用了人体骨架关节角度作为关键帧的特征，在下一步的工作中，将会添加更多特征，如：形状，纹理等，以期得到更好的效果。②对数据集中的混淆动作的识别结果还有待于提高，在今后的研究中对局部动作或者极易混淆动作识别会更加努力。此外，基于视频关键帧的识别可以应用于日常视频监控调看、互联网视频数据筛选等领域。

参 考 文 献

[1] 朱煜, 赵江坤, 王逸宁, 等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6): 848-857.

[2] PRIYA G G, DOMNIC S. Shot based keyframe extraction for ecological video indexing and retrieval [J]. Ecological Informatics, 2014, 23 (9): 107-117.

[3] SUN Z H, JIA K B, CHEN H X. Video key frames extraction based on spatial-temporal color distribution [C]//International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Los Alamitos: IEEE Computer Society Press, 2008: 196-199.

[4] HANNANE R, ELBOUSHAKI A, AFDEL K, et al. An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram [J]. International Journal of Multunedia Information Retrieval, 2016, 5(2): 89-104.

[5] 潘志庚, 吕培, 徐明亮, 等. 低维人体运动数据驱动的角色动画生成方法综述[J]. 计算机辅助设计与图形学学报, 2013, 25(12): 1775-1785.

[6] DEVANNE M, WANNOUS H, BERRETTI S, et al. 3-D human action recognition by shape analysis of motion trajectories on riemannian manifold [J]. IEEE Transactions on Cybernetics, 2014, 45(7): 1340-1352.

[7] LIU F, ZHUANG Y T, WU F, et al. 3D motion retrieval with motion index tree [J]. Computer Vision and Image Understanding, 2003, 92(2/3): 265-284.

[8] 王方石, 须德, 吴伟鑫. 基于自适应阈值的自动提取关键帧的聚类算法[J]. 计算机研究与发展, 2005, 42(10): 1752-1757.

[9] SONG X M, FAN G L. Joint key-frame extraction and object segmentation for content-based video analysis [J] IEEE Transactions on Circuits and Systems for Video Technology, 2006 16(7): 904-914.

[10] 田国会, 尹建芹, 韩旭, 等. 一种基于关节点信息的人体行为识别新方法[J]. 机器人, 2014, 36(3): 285-292.

[11] 石祥滨, 刘拴朋, 张德园. 基于关键帧的人体动作识别方法[J]. 系统仿真学报, 2015, 27(10): 2401-2408.

[12] 孙淑敏, 张建明, 孙春梅. 基于改进 K-means 算法的关键帧提取[J]. 计算机工程, 2012, 38(23): 169-172.

[13] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos. [J]. Computational Linguistics, 2014, 1(4): 568-576.

[14] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3D skeletons as points in a lie group [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society Press, 2014: 588-595.

[15] ZHANG Q, YU S P. An Efficient method of keyframe extraction based on a cluster algorithm [J]. Journal of Human Kinetics, 2013, 39(1): 5-14.

[16] WANG C Y, WANG Y Z, YUILLE A L. Mining 3D key-pose-motifs for action recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society Press, 2016: 289-293.