

基于深度学习和动态时间规整的人体运动检索

楚超勤,肖秦琨,高 嵩

(西安工业大学 电子信息工程学院,陕西 西安 710032)

摘要:随着计算机动画在各种应用中的日益普及,市场上出现了很多人体运动捕获设备,人们使用这些设备制作了大量的人体运动数据库。为了节约成本和高效地利用已有数据资源,提出了一种基于深度学习和动态时间规整相结合的人体运动检索方法。该方法包括两个主要阶段,在学习阶段,针对运动数据库中的运动序列,首先利用模糊聚类获取运动代表性帧,进而建立关键帧图像集合,然后应用深度神经网络学习关键帧图像集合,得到自动编码器,再应用自动编码器提取各个关键帧运动姿态的特征,建立运动特征数据库。在运动检索阶段,针对待查询运动序列,根据阶段1获取的自动编码器对每一关键帧图片提取特征,进而使用基于曼哈顿距离的动态规划方法计算待查询运动与数据库中运动的相似度,并根据相似度量值对检索结果进行排序。最后通过实验验证了该方法的有效性。

关键词:运动检索;模糊聚类;自动编码器;曼哈顿距离;动态规划

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2018)06-0059-05

doi:10.3969/j.issn.1673-629X.2018.06.013

Human Motion Retrieval Based on Deep Learning and Dynamic Time Warping

CHU Chao-qin, XIAO Qin-kun, GAO Song

(School of Electronic Information Engineering, Xi'an University of Technology, Xi'an 710032, China)

Abstract: With the popularity of computer animation in various applications, human motion capture equipment is produced on the market, which can be used to produce lots of human motion databases. To reduce cost and utilize the existing data resources better, we propose a method of human motion retrieval based on depth learning and dynamic time warping. It consists of two main phases. In the learning, first of all, we obtain motion representative frame by the fuzzy clustering in view of the sequence in the motion database and set up a collection of key frame images. Then we use the deep neural network to learn the collection of key frame images for the automatic encoder and following the automatic encoder to extract the feature of each key motion frame for establishment of the motion feature database. In the motion retrieval, the automatic encoder attained by former phase extracts the feature of each key frame motion image. We use the dynamic programming method based on Manhattan distance to calculate the similarity between the motion sequences queried and the motions in the database, and sort the search results according to similarity measures. Finally the experiment proves the effectiveness of the proposed method.

Key words: motion retrieval; fuzzy clustering; automatic encoder; Manhattan distance; dynamic programming

0 引言

近年来,随着创新科技的发展,计算机动画在各种应用中日益普及^[1-8]。人体运动编辑对计算机动画制作尤为重要,在动画制作领域,很多企业公司对制作非常逼真的人类动画产生了极大的需求。现在市场上已经有很多方法来产生人体运动数据。运动捕获(Mo-Cap)是一种众所周知的获取运动数据的方法,因此运

动捕获设备的应用价值也越来越突出,推动了大规模人体和物体运动数据库的发展^[8-9]。然而,随着各种运动数据的增长,检索能满足特定要求的动画运动是一件困难的事情。因此,运动检索技术成为近年来在运动捕捉动画领域的研究重点。

目前已经提出了一些运动检索方法,其中许多创新方法是在现有的音频检索方法上加以拓展应用,如

收稿日期:2017-07-12

修回日期:2017-11-23

网络出版时间:2018-02-24

基金项目:国家自然科学基金(61671362,61271362);陕西省自然科学基金(2017JM6041)

作者简介:楚超勤(1991-),男,硕士生,研究方向为智能信息处理;肖秦琨,博士,教授,研究方向为图模型理论及智能信息处理;高 嵩,博士,教授,研究方向为智能信息处理。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20180224.1521.080.html>

所熟知的动态时间规整 (dynamic time warping, DTW)^[10]。然而,因为这种类型的数据的属性和参数数据量很大,所以单一使用 DTW 方法对运动捕获数据的检索效率低。为了支持索引和提高 DTW 的检索性能,提出一种基于均匀缩放 (uniform scaling, US) 的算法^[11]。然而,基于均匀缩放的方法通常具有较高的计算成本。基于 DTW 方法和典型相关分析 (canonical correlation analysis, CCA) 扩展方法,被称为广义的典型时间规整 (generalized canonical time warping, GCTW),这种方法被用于调整多模态序列^[12]。除了基于 DTW 的方法,其他方法是寻求逻辑上类似的运动匹配。例如,用于呈现运动的模板技术,以及使用模板匹配的运动搜索^[13]。此外,提出使用几何特征构建索引树,使用聚类和分割,然后根据峰值点进行动作匹配^[14]。然而,这些方法都不能很好地区分紧密匹配的

运动。

文中利用给定查询的运动序列,从运动数据库检索非常相似的运动。如上所述,基于 DTW 的检索方法的表现比统计匹配方法有更好的性能,但检索效率较低,因此提出将基于深度学习和动态时间规整的人体运动检索方法,以提高运动匹配的性能和效率。然后,基于优化的代表性识别特征通常比原始无序描述符具有更好的性能,使用模糊聚类将冗余姿态描述符转换成判别描述符^[15]。最后通过实验对该算法进行验证。

1 检索算法

1.1 算法概述

提出算法的图解说明如图 1 所示,其中算法分为两个阶段:系统学习和运动检索。

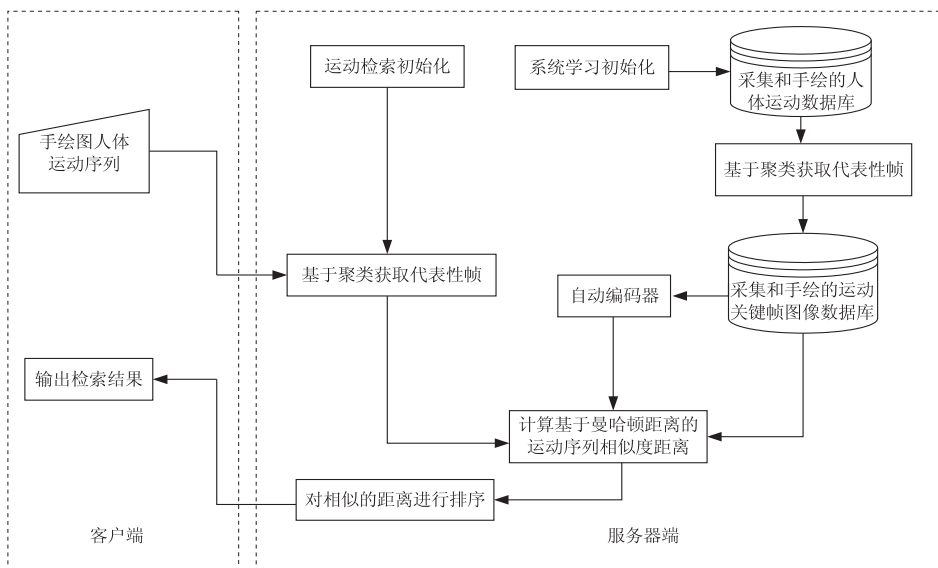


图 1 算法说明

在系统学习阶段,首先由相互间可分辨的运动集合构建运动数据库,其次通过聚类方法获取代表性帧图像,进而基于代表帧图像集合训练自动编码器模型,使用已经获取的编码器模型提取运动帧图像的特征;在运动检索阶段,基于上述步骤获取查询运动关键帧图像,进而应用自动编码器提取每一帧图像特征,应用曼哈顿动态规划算法计算待查询运动与运动数据库运动之间的相似度距离,排序输出检索结果。

1.2 系统学习

(1) 基于聚类获取代表性的帧。系统学习的关键一步是获取基于聚类的代表性框架。给定运动序列 $\{F_i\}_{i=1:n}$, 其中 n 是帧数,使用模糊 c -均值 (fuzzy c -means, FCM) 聚类方法生成代表性帧。为了计算两帧之间的距离,使用四元数^[15]来呈现身体姿态。设 F_i 为第 i 帧中的运动描述符, F_1 和 F_2 之间的距离计算公式如下: 万方数据

$$d(F_1, F_2) = (F_1 - F_2) (F_1 - F_2)^T \quad (1)$$

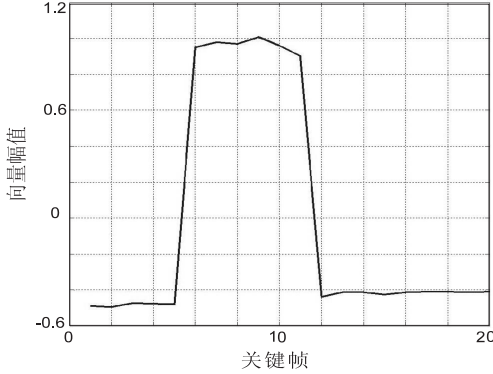
如果使用聚类方法来生成 c 个聚类中心,则选择距离聚类中心最短距离的一些帧作为代表帧,然后代表帧可以表示为 $RF = \{rf_k\}_{k=1:c}$, 其中 rf_k 对应于第 k 个聚类中心。因此可以使用 FCM 聚类对代表性帧进行提取。

图 2(a) 显示了代表性帧的第一主成分,对于图 2(b) 中的运动序列,在卡纳基梅隆大学 CMU 数据库中对应于“01-01.bvh”,从所有运动视频帧中找到 20 个聚类中心,不同的聚类数据用不同的颜色表示。为了便于展示,原始特征 (84 维四元数矢量) 使用主成分分析 (principal component analysis, PCA) 来缩小维数,只保留第一和第二主要分量。图 2(b) 显示了与聚类中心对应的代表帧。

1.3 使用自动编码器提取运动特征

自动编码器可以看作是神经网络。使用自动编码

器可以减小输入数据的维数,并将重构的信号作为输出。在深层网络中,自动编码器始终作为自动学习对象特征的良好模式,其在无监督的学习机制环境下训练,这一训练过程是必不可少的。自动编码器由编码器和解码器组成。



(a) 运动特征的第一主要部分



(b) 20 个代表性框架对应于 20 个中心

图2 使用 FCM 聚类的代表帧提取示例

假设自动编码器的输入为 x , 首先,该编码器将原始信号 x 映射到特征信号 $z^{[16]}$:

$$z^{(e)} = h^{(e)}(W^{(e)}x + b^{(e)}) \quad (2)$$

其中,“(e)”是指神经网络编码层; $h^{(e)}$ 是传递函数; $W^{(e)}$ 是加权矩阵; $b^{(e)}$ 是偏置向量。

接下来,解码器将特征信号 z 映射到估计值 \hat{x} 中^[16]:

$$\hat{x} = h^{(d)}(W^{(d)}z + b^{(d)}) \quad (3)$$

其中,“(d)”是指第 d 网络层; $h^{(d)}$ 是解码器的传递函数; $W^{(d)}$ 是权重矩阵; $b^{(d)}$ 是偏置向量。

一般来说,自动编码器的性能是由系统参数优化后决定的,并且代价函数始终是自动编码器参数训练的关键因素。根据深度学习理论,一个对象,诸如图像,通过输入到深层网络来提取特征并且进行特征重构,完成训练任务,输入(表示为 x)与特征重构后输出(表示为 \hat{x})之间的误差需要控制到最小的值。由此可以建立一个代价函数来描述这个误差^[16]:

$$E = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K (x_{kn} - \hat{x}_{kn})^2 + \lambda * \Omega_{\text{weights}} + \beta * \Omega_{\text{sparsity}} \quad (4)$$

代价函数 E 由 3 部分组成,第 1 部分是均方误差,第 2 部分是 L_2 正则化,第 3 部分是稀疏正则化, L_2 正则化系数为 λ , 稀疏正则化系数为 β 。

如果将 L_2 正则化:

$$\Omega_{\text{weights}} = \frac{1}{2} \sum_{j=1}^L \sum_{i=1}^n (w_{ji}^{(l)})^2 \quad (5)$$

其中, L, n, k 分别是训练数据中的隐层数、观测数和变量数。

通常添加一个正则化来激励稀疏项,如果将第 i 个神经元激活估量定义为^[17]:

$$\hat{\rho}_i = \frac{1}{n} \sum_{j=1}^n z_i^{(1)}(x_j) = \frac{1}{n} \sum_{j=1}^n h(w_i^{(1)T} x_j + b_i^{(1)}) \quad (6)$$

其中, n 是训练样本数; x_j 是第 j 个训练样本; $w_i^{(1)T}$ 和 $b_i^{(1)}$ 分别表示 $w_i^{(1)}$ 第 i 行的转置向量和偏移向量。

接下来,利用 Kullback Leibler 发散呈现稀疏正则化^[17-18]:

$$\Omega_{\text{sparsity}} = \sum_{i=1}^{D^{(1)}} \text{KL}(\rho \parallel \hat{\rho}_i) = \sum_{i=1}^{D^{(1)}} \rho \log\left(\frac{\rho}{\hat{\rho}_i}\right) + (1 - \rho) \log\left(\frac{1 - \rho}{1 - \hat{\rho}_i}\right) \quad (7)$$

当 ρ_i 和 $\hat{\rho}_i$ 相等, Kullback Leibler 发散是 0, 否则, 由于它们彼此偏离, 发散是较大的。

1.4 运动检索

为了实现运动之间的相似距离计算,使用曼哈顿距离动态规划 (Manhattan distance dynamic programming, MDDP)。MDDP 方法的目的是比较两个序列 $\mathbf{R}\mathbf{F}^X = (r_1^X, r_2^X, \dots, r_c^X)$ 和 $\mathbf{R}\mathbf{F}^Y = (r_1^Y, r_2^Y, \dots, r_c^Y)$ 之间的相似性,让整体匹配代价为 $C_p(\mathbf{R}\mathbf{F}^X, \mathbf{R}\mathbf{F}^Y)$:

$$C_p(\mathbf{R}\mathbf{F}^X, \mathbf{R}\mathbf{F}^Y) = [d_{\text{MDDP}}(r_i^X, r_j^Y)]_{c \times c} \quad (8)$$

$\mathbf{R}\mathbf{F}^X$ 和 $\mathbf{R}\mathbf{F}^Y$ 之间的一个规整路径 p 可以定义为本地代价度量,而 $d_{\text{Manh}}(r_i^X, r_j^Y)$ 表示 r_i^X 与 r_j^Y ($i, j = 1, 2, \dots, c$) 之间的曼哈顿距离。如果 \mathbf{r}_i^X 和 \mathbf{r}_j^Y 是 t 维向量,则 $\mathbf{r}_i^X = (x_1, x_2, \dots, x_t)$ 和 $\mathbf{r}_j^Y = (y_1, y_2, \dots, y_t)$, r_i^X 与 r_j^Y 之间的曼哈顿距离是:

$$d_{\text{MDDP}}(r_i^X, r_j^Y) = \sum_{i=1}^t |x_i - y_i| \quad (9)$$

此外, $\mathbf{R}\mathbf{F}^X$ 和 $\mathbf{R}\mathbf{F}^Y$ 之间的最佳规整路径是在所有可能的规整路径中具有最小总成本的规整路径 p^* 。然后定义 $\mathbf{R}\mathbf{F}^X$ 和 $\mathbf{R}\mathbf{F}^Y$ 之间的 MDDP 距离是 p^* 的总成本:

$$d_{\text{MDDP}}(\mathbf{R}\mathbf{F}^X, \mathbf{R}\mathbf{F}^Y) = C_{p^*}(\mathbf{R}\mathbf{F}^X, \mathbf{R}\mathbf{F}^Y) \quad (10)$$

为了确定最优路径 p^* , 使用动态规划,根据文献[6],有以下定理:

定理 1: 累积成本矩阵 D 满足:

$$\begin{cases} D(n, 1) = \sum_{k=1}^n d_{\text{MDDP}}(r_k^X, r_1^Y), n \in [1:c] \\ D(1, m) = \sum_{k=1}^m d_{\text{MDDP}}(r_1^X, r_k^Y), m \in [1:c] \\ D(n, 1) = \min\{D(n-1, m-1), D(n-1), m\}, \\ D(n, m-1) + d_{\text{MDDP}}(r_n^X, r_m^Y) \end{cases} \quad (11)$$

根据定理 1, 最终优化 MDDP 的距离是:

$$d_{\text{MDDP}}(\mathbf{R}\mathbf{F}^x, \mathbf{R}\mathbf{F}^y) = C_p(\mathbf{R}\mathbf{F}^x, \mathbf{R}\mathbf{F}^y) = D(n, m) \quad (12)$$

文中选择曼哈顿距离作为本地成本测量, 与使用欧几里德距离作为本地成本测量的传统 DTW 算法相比, 提出的检索方式^[6]具有更好的性能, 接下来的实验将会对此进行验证。基于两个关键步骤, 代表帧提取和相似性距离匹配, 可以根据相似距离顺序获得检索结果。

2 实验

实验选择使用 HDM5 运动数据库^[9], 从数据库中获得 3 000 个不同的动作片段, 将 3 000 个运动片段分类到 100 个运动集合中。得到 30 个随机选择的运动集合, 其中每个集合包括 10 个运动。实验的测试环境是在具有奔腾 6 GHz CPU 和 2 GB RAM 的电脑上进行评估。

由于每个原始动作通常包含不止一个活动, 为了获得准确的测试结果, 这些片段被分割成由单个活动组成的基本运动序列。为了与提出的方法进行比较, 也实施了 DTW 方法和 US 方法。测试目的是根据给定的查询运动序列从运动数据库中搜索最佳匹配的运动序列。

自动编码器深度学习神经网络模型^[19]由四个模块构成, 即输入端信号是 1 600 维的向量组, 对输入的数据进行编码的编码器模块, 对编码后的数据进行重构的解码器模块以及输出模块。每一个运动姿势的图像是 40 * 40 像素的尺寸大小, 构成 1 600 维的向量, 这一运动序列存储在向量组中, 经过深度学习网络预处理把原运动序列降至 100 维。

为计算 30 种运动集合的平均精度值, 同时采用了 Deep-DTW、Quat-DTW 和 Deep-US 方法。运动序列不同检索精度的对比如图 3 所示, 图 3 代表了数据库中的一个运动动作: clap5Resp。

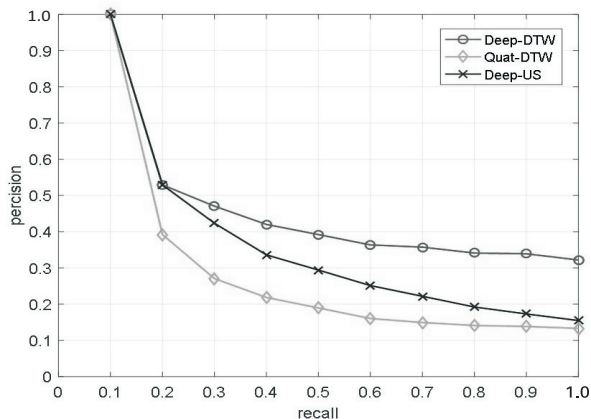


图 3 运动序列检索精度的对比

图 4 是运动序列检索精度仿真结果, 对应于图 3 一样的运动, 并将所有查询的运动序列的精度值进行平均, 得到运动类别的平均值, 检索精度使用 PR (precision-recall) 曲线图进行评估:

$$\text{precision} = \frac{\#\{\text{relevant} \cap \text{retrieved}\}}{\#\text{retrieved}} \quad (13)$$

$$\text{recall} = \frac{\#\{\text{relevant} \cap \text{retrieved}\}}{\#\text{relevant}} \quad (14)$$

其中, #retrieved 是检索到的运动序列的数量; #relevant 是相关运动序列的数量。

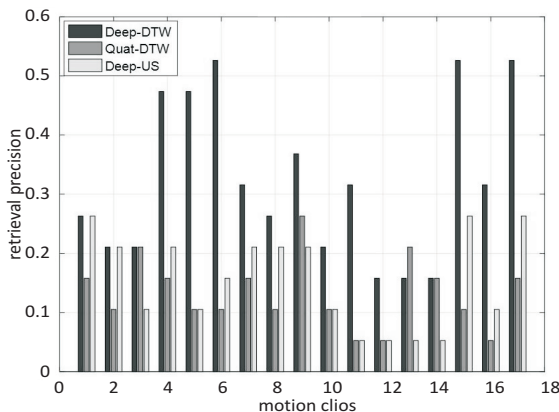


图 4 PR 曲线检索精度仿真结果

从图 4 可以看到, 使用 Deep-DTW 方法对序列进行检索^[20-25], 精度都高于其他两种方法。说明使用 Deep-DTW 方法对序列进行检索较其他检索方法性能好。

3 结束语

提出一种基于深度学习和动态时间规划相结合的运动检索算法。针对运动数据库中的运动序列, 首先利用模糊聚类获取运动代表性帧及其对应的权重值, 进而建立关键帧图像集合。基于深度学习, 通过对图像集合学习来训练自动编码器, 运用训练的自动编码器提取各个关键帧运动姿态特征, 建立运动特征数据库。为了计算相似度距离, 在运动检索方法中, 针对待查询运动序列, 使用训练获取的自动编码器对每一关键帧图片提取特征, 进而使用基于曼哈顿距离的动态规划方法计算待查询运动与数据库中运动的相似度, 并根据相似度量值对检索结果进行排序。实验结果证明, 该方法具有很好的精确性和有效性。

参考文献:

- [1] XIAO Jun, TANG Zhangpeng, FENG Yinfu, et al. Sketch-based human motion retrieval via selected 2D geometric posture descriptor[J]. Signal Processing, 2015, 113:1-8.
- [2] WANG Pengjie, LAU R W H, PAN Zhigeng, et al. An Eigen-based motion retrieval method for real-time animation[J].

- Computers & Graphics, 2014, 38: 255–267.
- [3] LI Meng, LEUNG H, LIU Zhiguang, et al. 3D human motion retrieval using graph kernels based on adaptive graph construction[J]. Computers & Graphics, 2016, 54: 104–112.
 - [4] MÜLLER M, RÖDER T, CLAUSEN M. Efficient content-based retrieval of motion capture data[J]. ACM Transactions on Graphics, 2005, 24(3): 677–685.
 - [5] MÜLLER M, BAAK A, SEIDEL H P. Efficient and robust annotation of motion capture data[C]//ACM SIGGRAPH/Eurographics symposium on computer animation. New Orleans; ACM, 2009: 17–26.
 - [6] KRUGER B, TAUTGES J, WEBER A, et al. Fast local and global similarity searches in large motion capture databases[C]//Proceedings of the 2010 ACM SIGGRAPH/Eurographics symposium on computer animation. Madrid, Spain; ACM, 2010: 1–10.
 - [7] VOGEL A, KRUGER B, KLEIN R. Efficient unsupervised temporal segmentation of human motion[C]//Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation. Copenhagen, Denmark; ACM, 2015: 167–176.
 - [8] MULLER M, RÖDER T, CLAUSEN M, et al. Documentation mocap database HDM05[R]. [s. l.]: [s. n.], 2007.
 - [9] Graphics Lab. Motion capture database[EB/OL]. 2012. <http://mocap.cs.cmu.edu/>.
 - [10] KEIGH E, RATANAMAHATANA C A. Exact indexing of dynamic time warping[J]. Knowledge & Information Systems, 2005, 7(3): 358–386.
 - [11] KAPADIA M, CHIANG I K, THOMAS T, et al. Efficient motion retrieval in large motion databases[C]//ACM SIGGRAPH symposium on interactive 3d graphics and games. Orlando, Florida; ACM, 2013: 19–28.
 - [12] ZHOU Feng. Generalized time warping for multimodal alignment of human motion[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2012: 1282–1289.
 - [13] BAAK A, MÜLLER M, SEIDEL H P. An efficient algorithm for keyframe-based motion retrieval in the presence of temporal deformations[C]//ACM international conference on multimedia information retrieval. Vancouver, British Columbia, Canada; ACM, 2008: 451–458.
 - [14] CHEN Cheng, ZHUANG Yueting, NIE Feiping, et al. Learning a 3D human pose distance metric from geometric pose descriptor[J]. IEEE Transactions on Visualization & Computer Graphics, 2011, 17(11): 1676–1689.
 - [15] ZHOU Feng, TORRE F D L, HODGINS J K. Hierarchical aligned cluster analysis for temporal clustering of human motion[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 35(3): 582–596.
 - [16] ZHAN Xiwu, HOUSER P R, WALKER J P, et al. A method for retrieving high-resolution surface soil moisture from hydros L-band radiometer and radar observations[J]. IEEE Transactions on Geoscience & Remote Sensing, 2006, 44(6): 1534–1544.
 - [17] KOVAR L, GLEICHER M. Automated extraction and parameterization of motions in large data sets[J]. ACM Transactions on Graphics, 2004, 23(3): 559–568.
 - [18] SALTON G, MCGILL M J. Introduction to modern information retrieval[M]. New York; McGrawHill, 1983.
 - [19] 尹征, 唐春晖, 张轩雄. 基于改进型稀疏自动编码器的图像识别[J]. 电子科技, 2016, 29(1): 124–127.
 - [20] 肖秦琨, 李俊芳, 肖秦汉. 基于四元数描述和 EMD 的人体运动捕获数据检索[J]. 计算机技术与发展, 2014, 24(3): 90–93.
 - [21] 吕刚, 郝平, 盛建荣. 一种改进的神经网络在小图像分类中的应用研究[J]. 计算机应用与软件, 2014, 31(4): 182–184.
 - [22] 杨涛. 运动捕获数据关键帧提取及检索研究[D]. 杭州: 浙江大学, 2006.
 - [23] 李婷. 基于运动捕获数据的人体运动编辑技术研究[D]. 武汉: 华中科技大学, 2008.
 - [24] 郑启财. 基于深度学习的图像检索技术的研究[D]. 福州: 福建师范大学, 2015.
 - [25] 连荷清. 人体运动捕获数据的检索方法研究[D]. 南京: 南京理工大学, 2013.