DOI:10.19322/j.cnki.issn.1006-4710.2017.02.008

基于时间维度局部特征的人体行为识别

张九龙[1]，张镇东[1]，杨　夙[1,2]，高　阳[1]，肖照林[1]

(1.西安理工大学 计算机科学与工程学院，陕西 西安 710048；
2.复旦大学 计算机科学技术学院，上海 201203)

摘要：视频人体行为识别算法中，局部特征三维模板卷积法难以避免背景中伪兴趣点且计算耗时。提出一种高效准确的融合时间维度和FAST角点特征的运动人体兴趣点检测方法，针对FAST角点不能表达时间维度信息的缺陷，将相邻三帧两两做差，然后在得到的前向和后向运动图像上进行FAST角点检测，取两个特征点集的交集作为当前帧运动人体局部兴趣点。该方法有效结合了时间维度信息和FAST算子的优点，具有耗时短、准确率高、运动相关性好的特点。最后应用词袋模型进行人体行为特征建模，分别应用SVM、KNN、决策树和LDA进行分类识别，在Weizmann、KTH数据库上进行测试，实验表明：SVM获得最好的分类性能，KNN获得最高的效率，因此KNN可以利用到实时的行为识别中。

关键词：行为识别；局部特征；运动信息；FAST角点；词袋模型

中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：1006-4710(2017)02-0169-06

Local feature extraction using time domain information for human action recognition

ZHANG Jiulong[1], ZHANG Zhendong[1], YANG Su[1,2], GAO Yang[1], XIAO Zhaolin[1]

(1. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China;
2. School of Computer Science, Fudan University, Shanghai 201203, China)

Abstract: In the human action recognition application, the traditional 3D template convolution method is time-consuming and difficult to avoid defects of pseudo interest points of background. To overcome this weak point, we propose a motion human local interest point detect method combining motion information and FAST feature. First, the difference is computed on every two adjacent frames, and the FAST(Features From Accelerated Segment) feature point detection is deployed on the two pieces of motion information by taking intersection of the two points set as final output with non-maximum suppression. With the low time-consuming FAST algorithm applied, this method should be an efficient motion intersection point detector with high accuracy and motion correlation. Finally, we use the BOW model to generate action feature vector. The classifier used is SVM (Support Vector Machine), KNN, Decision Tree and LDA. Performance is tested on deferent datasets, the simple KTH and Weizmann, SVM classifier, obtaining the best accuracy with KNN being more efficient.

Key words: action recognition; local feature; motion information; FAST corner; BOW(Bag of Words) model

人体行为识别技术已经成为计算机视觉领域中的一个热点研究方向，在现实生活中有着十分广泛的应用，是一个复杂而又富有挑战性的难题，同时也是一个兼具研究价值和应用意义的研究课题。人体行为识别的核心内容是特征提取，良好的特征应该能表征当前行为，同时扩大类间间距而缩小类内

收稿日期：2016-07-06

基金项目：国家自然科学基金资助项目(61402362)；陕西省自然科学基金资助项目(2015JQ6218，2016JQ6069)；陕西省教育厅专项科研计划资助项目(16JK1553)；西安市碑林区科技计划资助项目(GX1616)

作者简介：张九龙，男，副教授，博士，研究方向为图像处理与模式识别。E-mail:chinajiulong@hotmail.com

间距[1]。

人体行为特征可以分为整体特征和局部特征。整体特征是将视觉对象即人体作为目标进行处理得到的特征,通常是先应用一系列的前景提取算法得到运动前景,再对它进行描述。用这种方法提取的特征依赖于定位、前景提取以及跟踪的精度,而且对视角、遮挡和噪声比较敏感,因此,整体特征的表现往往取决于能否有效的处理以上这些因素。最早应用剪影信息的是 Bobick 和 Davis[2],他们将单一视角下二值化的剪影图叠加起来得到 MEI 图和 MHI 图,利用 Hu 矩来进行模板匹配。Ali 等[3]通过光流来获取人体运动学特征,赵晓健等人[4]利用稠密光流轨迹和稀疏编码识别人体行为,胡石等人[5]运用傅里叶变换与边缘小波矩描述子相结合,给出了多段定向距离轮廓描述矩阵,实现了轮廓特征的提取。Wang 等人[6]使用 DTW(Dynamic Time Warping)[7]匹配重心轨迹进行行为识别,整体特征还可以通过基于人体结构的方法,比如关节点模型[8]、3d 骨架[9]等进行描述。局部特征通常以独立的区块为单位进行处理,最后将区块特征串联起来得到完整的特征。最著名的时空特征点当属 Laptev 的 3D-Harris[10]特征点和 Dollár[11]特征点,以及 Bregonzio[12]提出的特征点检测方法。特征描述子最常见的有 HOG/HOF((Histograms of Oriented Gradients/Histograms of Optical Flow)[13-14],之后 Willems 等人[15]将 SURF 特征(Speed Up Robust Feature)也扩展到了三维,类似的还有 3D-SIFT[16]、3D-HOG[17]。

局部特征的方法可以保持一定程度的对噪声、遮挡的鲁棒性,而且也不一定需要背景减除和跟踪,因此得到了广泛的应用。但是传统的三维模板卷积求兴趣点的方法很难避免伪兴趣点的情况,有些兴趣点落在背景中部分目标的边缘上,或者人体轮廓内部纹理丰富的地方。Chakraborty[18]在原有方法上基于非极大值抑制的思想提出 selective-stip,但从效果来看,伪兴趣点依然存在。因此,寻求准确率更高、运动相关性更好的局部兴趣点检测方法仍然是行为识别领域的研究难点和重点。本文提出融合运动信息和 FAST 角点特征的兴趣点检测方法,针对 FAST 角点不能表达时间维度信息的缺陷,将相邻三帧两两做差,在得到的前向和后向运动图像上进行 FAST 角点检测,取两个特征点集的交集作为当前帧运动人体局部兴趣点。该方法有效结合了时间维度信息和 FAST 算子的优点,实验效果良好。

1 基于时间维度的局部行为特征提取

1.1 FAST 角点检测算法

角点也称为特征点,通常指的是两条边的交点,实际应用当中,角点的意思是拥有特定特征的点,比如局部最小或者满足一定的梯度特征、数学特征。角点检测被广泛的应用于运动跟踪、图像匹配、三维建模等领域。传统的几种角点检测算法结果如图 1 所示。

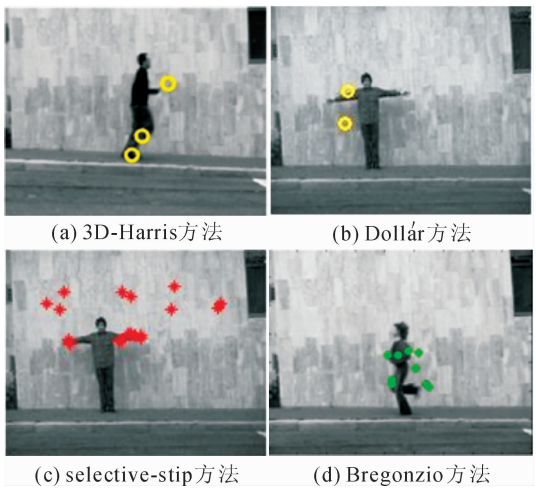


图 1 几种常见的局部特征点示意图

Fig. 1 Several commonly used local features

SIFT 特征点较早用于特征检测和图像匹配,拥有旋转、平移和尺度不变性,但是效率较低且耗时,故后来采用 SURF 特征点作为该方法的改进。对 Harris 角点而言,其缺点是特征点响应强度值只考虑离散的 8 个 45°角方向。

实际的应用当中,角点检测是动作识别中的基础部分,因此对其性能有较高的要求。在 FAST 的性能研究中发现,该角点与周围的像素点有足够的不同。具体来说,如果一点的灰度值比其周围领域内足够多的像素点的灰度值大或者小于某一阈值,该点就被定义为角点,其具体计算原理可以简单表述为:记该点像素值为 P,以该像素点为中心,3 像素为半径作圆,如果圆上连续 N 个像素的值都大于 $P+S$ 或者小于 $P-S$,那么该点就是角点,其中 S 为预先设定的阈值,一般取 10～20,N 一般取 9 或 12。

1.2 伪兴趣点消除预处理

运动前景提取的方法有帧差法、简单减背景法以及高斯建模法。其中帧差法简单高效,但是准确率较低,轮廓不完整,容易出现“空洞”和“重影”。实际应用中背景往往不固定,因此简单减背景法受到

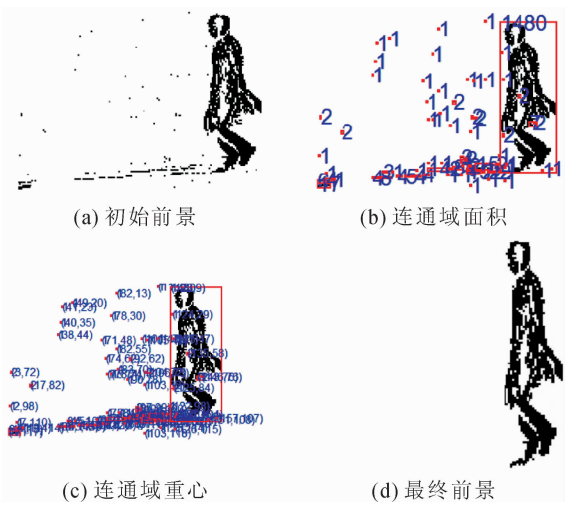
限制，而高斯建模法收敛较慢、耗时较长。本文在分析以上方法优劣的基础上，提出采用前后向结合的帧差法进行初始运动前景的提取：

$$D_i = | F_{i+1} - F_i | \cup | F_i - F_{i-1} | \quad (1)$$

式中，D_i 为第 i 帧对应的运动前景，F_i 为视频中第 i 帧图像。

图 2(a)是由前后向结合的帧差法所得的初始前景。可以看出，初始前景有较多的噪声点，这些点很容易被当成兴趣点(POI)，本文提出的伪兴趣点消除算法基于连通域的面积及连通域到图像重心的距离进行判别，其流程如下。

1) 如图 2(b)所示，用红色矩形框标出每个连通域的范围，用蓝色字体标示其大小。可以看出，大部分的噪声点连通域只包含 1～2 个像素。之后将面积小于阈值 T_1 的连通域抹去。依据实验结果，帧差法的噪声点面积较小，T_1 一般选 10～20，即可将大部分的潜在伪兴趣点去掉。



(a) 初始前景 (b) 连通域面积

(c) 连通域重心 (d) 最终前景

图 2 基于形态学的前景伪兴趣点消除

Fig. 2 Removal of pseudo POI by morphology

2) 计算每个连通域到图像整体重心的距离。图像整体重心计算公式为：

$$G(x) = \frac{\sum_{i=1}^{N} x_i}{N} \quad (2)$$

$$G(y) = \frac{\sum_{i=1}^{N} y_i}{N} \quad (3)$$

式中，x_i，y_i 分别代表第 i 个连通域中心的横纵坐标，$G(X)$、$G(Y)$ 分别为水平和垂直方向的重心坐标。

由式(2)和式(3)可求出每个连通域的重心坐标，如图 2(c)所示。将图像对角线长度的 1/4 作为阈值 T_2，如果某一连通域的重心与整幅图像的重心距离超过 T_2，就将该连通域抹去，这样可以将一些位于边缘的面积较大的潜在伪兴趣点去除，得到最终的前景如图 2(d)所示。

1.3 融合运动信息的局部兴趣点

本文提出的融合运动信息和 FAST 角点的兴趣点检测方法以帧间差信息为基础，提取其 FAST 角点，如图 3 所示，整个提取过程可以分为以下四步：

第 1 步 逐个提取视频帧图像，对当前帧进行前后帧求差，得到两幅帧差信息图，并根据式(1)进行时间域信息融合，接着进行伪兴趣点消除，去除方法见第 1.2 节；

第 2 步 在第 1 步的输出图像上进行 FAST 角点检测，检测结果如图 3(b)所示；

第 3 步 求第 2 步所得的两个 FAST 角点的交集，得到最终的对应于当前帧图像的角点，其效果如图 3(c)所示；

第 4 步 继续以上的第 1～3 步，直到视频结束。

由本方法得到的运动人体兴趣点结果如图 4 所示。

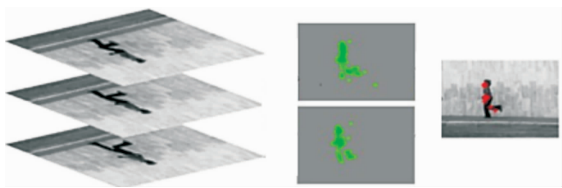


图 3 融合时间信息和 FAST 角点的兴趣点检测方法

Fig. 3 POI detection by combination of time domain information and FAST

(a) 前一帧 (b) 后一帧

图 4 基于时间信息的局部特征点检测结果

Fig. 4 Detection result using time domain information

2 基于 BOW 的行为建模

特征数量取决于视频中兴趣点的多少，使用 BOW(Bag of Words)模型可以解决特征数量不一致的问题。

BOW 模型的本质是统计某一个词汇(视觉特征)出现的频数，它最初被用在文本分类中，将文档

表示成特征矢量,以此进行训练识别。它的基本思想是:对于一串文本,可以将它看作一些相互独立的基本词汇的集合,这些基本词汇的集合就成为词袋,在此基础上统计该文本中词汇的频数,就可以得到其特征向量,从而进行分类。在行为识别中,词汇就是视觉特征,同样的,对这些特征进行统计,就可以得到长度一致的特征向量。本文选取 HOG\HOF 特征来生成局部描述子。

构建合理的视觉词典是 BOW 方法的核心,词典构建的合理与否,如词典的大小、词汇的抽象层次,都会影响最终的识别结果。初始的视觉特征往往比较杂乱,要得到视觉词汇,先要对这些特征进行聚类,将所得的聚类中心作为视觉词汇。下面介绍基于 K-Means 的视觉词汇聚类算法。

2.1 K-Means 聚类生成视觉词汇

K-Means 聚类的目的就是对于给定的数据集合$\{x_i\}_{i=1}^{N}$,其中 x_i 维数相同,在给定分类数目 k($k\leqslant N$)的条件下,将原始数据分成 k 类,$S=\{S_1,S_2,\cdots,S_k\}$,S_i 对应第 i 类的集合,K-Means 的总体目标即对式(4)求最小值,其中 μ_i 对应 S_i 的聚类中心。

$$\arg\min_{S}\sum_{i=1}^{k}\sum_{x_i\in s_i}||x_i-\mu_i||^2 \tag{4}$$

K-Means 的整个算法步骤可以表示如下:

第 1 步 从原始的数据中随机选取 k 个元素,作为 k 个簇的初始中心 μ_i。

第 2 步 分别计算剩下的元素到 k 个簇中心的距离,将其划归到距离最近的簇 S_i,判断依据可以表示为:

$$\arg\min_{i}||x_i-\mu_i||^2 \tag{5}$$

第 3 步 根据以上结果,重新计算 k 个簇各自的中心 μ_i,计算方法是取簇中所有元素各自维度的算术平均数,计算公式依据式(6),其中若 $x_i\in S_i$ 成立,ω 为 1,否则为 0。

$$\mu_i=\frac{\sum_{i=1}^{k}\omega x_i}{\sum_{i=1}^{k}\omega} \tag{6}$$

第 4 步 重复 2~3 步,将全部元素按照新的中心重新聚类。

第 5 步 重复第 4 步,直到聚类结果稳定,即每个簇的聚类中心不再变化为止,得到最终的聚类结果输出。

值得一提的是,对于人体行为识别来说,视觉词汇的数量一般大于 300,这也是文献给出的经验值。

2.2 AP 聚类生成视觉词汇

K-Means 方法是不稳定的聚类方法,其聚类的数目需人为设定,因此本文同时尝试 AP(Affinity Propagation)[19]聚类生成视觉词汇,比较二者的优缺点。AP 算法是稳定的聚类算法,自 2007 年在 Science 杂志上发表以来受到了广泛的关注,为聚类算法提供了一种新的思路。其核心思想在于对聚类中心的描述或者说是筛选上,该方法可以为其他的聚类算法提供一个“准则”。

该算法认为聚类中心应当具有以下特点:①其本身密度大,即它被密度不超过自身的点包围;②与其它密度更大的点距离“较远”。

对于给定的数据集 $A=\{x_i\}_{i=1}^{N}$,其中 $I_A=\{1,2,\cdots,N\}$为相应下标集,d_{ij} 表示 x_i 到 x_j 的距离,其中 $i,j\in\{1,2,\cdots,N\}$。以 ρ_i 和 δ_i 表示数据 x_i 对应的密度和到聚类中心的距离,其中 ρ_i 通过式(7)计算:

$$\rho_i=\sum_{j\in I_{A\{i\}}}X(d_{ij}-d_c) \tag{7}$$

$$X(x_i)=\begin{cases}1 & x_i<0\\ 0 & x_i\geqslant 0\end{cases} \tag{8}$$

d_c 为预先设定的一个值,文中称为截断距离,设$\{q_i\}_{i=1}^{N}$表示$\{\rho_i\}_{i=1}^{N}$的降序排列下标序,即满足$\rho_{q1}\geqslant\rho_{q2}\geqslant\cdots\geqslant\rho_{qN}$,其中 δ_i 通过式(9)定义:

$$\delta_{qi}=\begin{cases}\min\limits_{\substack{q_j\\ j<i}}\{d_{qiqj}\} & i\geqslant 2\\ \max\limits_{j\geqslant 2}\{\delta_{qj}\} & i=1\end{cases} \tag{9}$$

选取聚类中心时,将 $\rho_i\times\delta_i$ 做降序排列,选取前若干个作为聚类的中心,具体数目根据实验确定。

AP 聚类不需要指定聚类个数,它的聚类中心是原始数据中确切存在的一个数据点,多次执行 AP 聚类算法得到的结果是一样的,其算法复杂度为 $O(N\times N\times\log N)$。

3 实验结果

3.1 实验环境与数据集合

本文实验的硬件平台为 CPU Intel Core-i54 核 2.0 GHz、内存 8GB,操作系统是 Windows 764 位,编程平台为 Matlab R2012a。

本文实验数据集包括 Weizmann 数据集和 KTH 数据集。Weizmann 数据集共包括 90 段视频,由 9 个人执行 10 个不同的动作(弯曲、抬起、走步、侧移、前跳、猛跳、摇手 1、摇手 2、跑步、跳)。该数据库中,视角、视频背景以及摄像头都是静止的。KTH 数据集包括 6 类行为(跑步、击拳、挥手,拍手、慢跑、走),是由 25 个人在 4 个场景下执行的,一

共有 599 段视频。该数据集中，背景相对静止，除了镜头的拉近拉远，摄像机的运动比较轻微。与 Weizmann 数据集相比，KTH 数据集类间差异明显，视角和时长也有区别。

3.2 实验方案

本文利用 SVM、KNN、决策树、LDA 这四种方法作为分类器。SVM 的核函数分别取 linear、sigmoid、RBF、polynomial 核，对 BOW 模型建模得到的视频序列特征向量进行分类，在 Weizmann 和 KTH 数据集上进行测试，将样本分为 5 个子集进行交叉验证，将其中一个集合作为测试集，而将其余集合作为训练集，这样每一个子数据集都至少有一次作为测试集。每一回合中，几乎所有的样本皆用于训练模型，因此这样的训练集最接近原始样本的分布，同时能最大程度的消除随机因素对实验结果的影响，提高了结果的可靠性。

3.3 方法对比

本文方法在 KTH 数据集上得到的识别性能，如图 5 所示，图中包含 KNN、决策树、LDA 的识别性能。

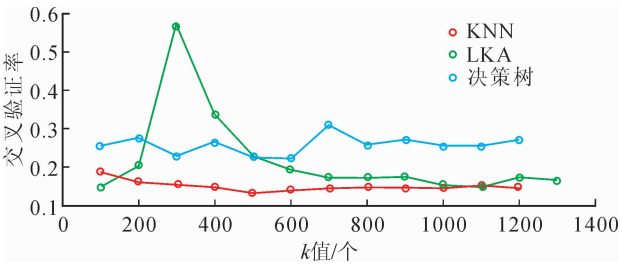


图 5 不同分类方法在 KTH 数据集上的分类性能

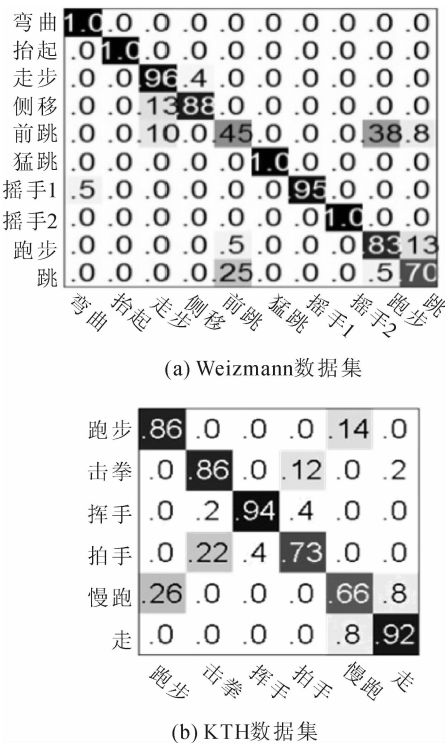
Fig. 5 Performance comparison on KTH

在 KTH 上，AP 聚类方法生成的词汇数量为 156，识别结果略低于 K-Means 方法。但实验中后者的词汇量和大多数文献中给出的参考值一样，都大于 300，增加了时间开销。

最终实验结果显示，SVM 分类器在取多项式核的情况下能获得最佳的分类效果；KNN 因其不需要进行训练，是一种高效的分类方法，当 k 值为 3 时，KNN 可获得最佳的分类效率。

本文应用了三种方法进行识别，分别为 KNN、LDA 和决策树方法。识别结果以混淆矩阵的形式出现在图 6～8 中。

另外，SVM 方法在 KTH 数据集上的识别率总体优于 KNN 方法，但计算耗时较长；KNN 在 Weizmann 数据集上的表现总体优于决策树方法，且时间代价小。



(a) Weizmann数据集

(b) KTH数据集

图 6 两个数据集上 KNN 方法的混淆矩阵

Fig. 6 Confusion matrix on two datasets with KNN classifier

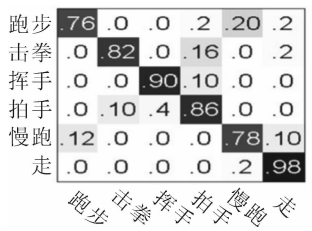


图 7 KTH 数据集上 SVM 方法的混淆矩阵

Fig. 7 Confusion matrix on KTH with SVM classifier

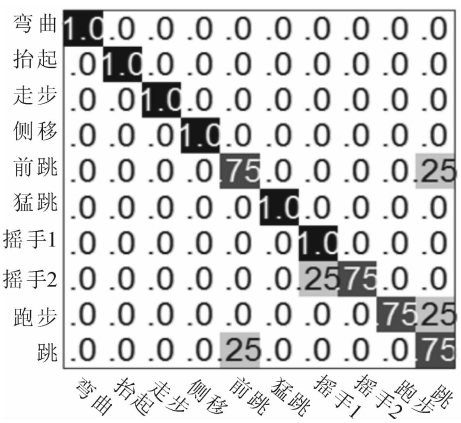


图 8 Weizmann 数据集上决策树方法的混淆矩阵

Fig. 8 Confusion matrix on Weizmann with Decision Tree classifier

4 结 语

本文算法充分结合了时间维度信息与 FAST 特征的高效，是一种快速的局部特征检测算法，这些特征点具有表征人体行为空域信息和时域信息的能力。另外，将 BOW 方法应用于行为识别解决了特征数量不一致的问题，但是 BOW 方法忽视了视觉词汇之间的高阶关联，因此，后续研究中可以探索高阶文法的效果。

参考文献：

[1] POPPE R. A survey on vision-based human action recognition[J]. Image and Vision Computing, 2010, 28(6): 976-990.

[2] BOBICK A F, DAVIS J W. The recognition of human movement using temporal templates[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2001, 23(3): 257-267.

[3] ALI S, SHAH M. Human action recognition in videos using kinematic features and multiple instance learning [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(2): 288-303.

[4] 赵晓健，曾晓勤. 基于稠密光流轨迹和稀疏编码算法的行为识别方法[J]. 计算机应用, 2016, 36(1): 181 - 187. ZHAO Xiaojian, ZENG Xiaoqin. Action recognition method based on dense optical flow trajectory and sparse coding algorithm[J]. Journal of Computer Applications, 2016,36(1):181-187.

[5] 胡石，梅雪. 人体行为动作的形状轮廓特征提取及识别[J]. 计算机工程, 2012, 38(2): 198-200.
HU Shi, MEI Xue. Shape contour feature extraction and recognition of human behavior motion[J]. Computer Engineering, 2012, 38(2): 198-200.

[6] WANG H, SCHMID C. Action recognition with Improved trajectories[C]//IEEE International Conference on Computer Vision(ICCV), December 3-6, 2013, Sydney, Australia. IEEE, 2013: 3551-3558.

[7] SEMPENA S, MAULIDEVI N U, ARYAN P R. Human action recognition using dynamic time warping [C]//Electrical Engineering and Informatics (ICEEI), 2011 International Conference on. July 17-19, 2011, Bandung, Indonesia. IEEE, 2011: 1-5.

[8] XIA L, CHEN C C, AGGARWAL J K. View invariant human action recognition using histograms of 3d joints [C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. June 16-21, 2012, Providence, Rhode Island, USA. IEEE, 2012: 20-27.

[9] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3d skeletons as points in a lie group[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2014, Columbus, Ohio, USA. 2014: 588-595.

[10] LAPTEV I. On space-time interest points[J]. International Journal of Computer Vision, 2005, 64(2-3): 107-123.

[11] DOLLÁR P, RABAUD V, COTTRELL G, et al. Behavior recognition via sparse spatio-temporal features [C]//Visual Surveillance and Performance Evaluation of Tracking and Surveillance(VS-PETS), 2005. 2nd Joint IEEE International Workshop on. October 15-16, 2005 Beijing, China. IEEE, 2005: 65-72.

[12] BREGONZIO M, GONG S, XIANG T. Recognising action as clouds of space-time interest points[C]// Computer Vision and Pattern Recognition (CVPR), 2009. IEEE Conference on. June 20-25 ,2009 Miami, Florida. IEEE, 2009: 1948-1955.

[13] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition(CVPR), 2005. IEEE Computer Society Conference on. June 25-29, 2005, Anchorage, Alaska, USA. IEEE, 2005, 1: 886-893.

[14] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies[C]// CVPR, 2008-IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society. June 23-28, 2008, Anchorage, Alaska. 2008: 1-8.

[15] WILLEMS G, TUYTELAARS T, GOOL L V. An efficient dense and scale-invariant spatio-temporal interest point detector[C]//European conference on computer vision. Springer Berlin Heidelberg, December 11-12, 2008, Dublin, Ireland. 2008: 650-663.

[16] SCOVANNER P, ALI S, SHAH M. A 3-dimensional sift descriptor and its application to action recognition [C] Proceedings of the 15th International Conference on Multimedia. September 9-12, 2009, Vienna, Austria. ACM. 2007: 357-360.

[17] KLASER A, MARSZALEK M, SCHMID C. A spatio-temporal descriptor based on 3d-gradients[C]// BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, May 14, 2008, London, UK. 2008: 1-10.

[18] CHAKRABORTY B, HOLTE M B, MOESLUND T B, et al. Selective spatio-temporal interest points[J]. Computer Vision and Image Understanding, 2012, 116(3): 396-410.

[19] FREY B J,DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(814): 972-976.

(责任编辑 周 蓓)