

基于混合时空特征描述子的人体动作识别

范晓杰, 宣士斌, 唐 凤

(广西民族大学 信息科学与工程学院 广西 南宁 530006)

摘要: 针对基于局部时空特征的行为识别中获取高效兴趣点、合理描述兴趣点及表征运动特征等关键问题,提出一种基于混合时空特征和SOM网络的新的行为识别框架。首先,从输入视频中提取出多尺度的Dollar时空兴趣点,并由时空兴趣点提取用于描述局部运动区域的视频块。然后,提出多向投影的光流直方图(DPHOF)构造方法,并与3D梯度方向直方图(HOG3D)结合描述视频块;利用SOM构造全局视频描述子。最后,用K最近邻(KNN)进行分类。对该方法在KTH和UCF-YT数据集上进行了验证,取得了很好的识别效果。实验结果表明,提出的DPHOF描述符能高效表示时空兴趣点,并优于HOG3D和HOF的描述性,且由SOM构造出的全局视频描述子可以高效地表示视频特征,该方法具有更好的识别结果。

关键词: 时空兴趣点; 3D有向直方图; 光流直方图; 自组织特征映射

中图分类号: TP31

文献标识码: A

文章编号: 1673-629X(2018)02-0098-04

doi: 10.3969/j.issn.1673-629X.2018.02.022

Realistic Human Action Recognition Based on Mixed Spatio-temporal Feature Descriptor

FAN Xiao-jie, XUAN Shi-bin, TANG Feng

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

Abstract: In view of key problems like efficient obtainment and reasonable description of interest points and characterization of movement in human action recognition based on local features of time and space, we present a new action recognition framework based on mixed space-time feature and SOM. Firstly, the multi-scale Dollar's spatio-temporal interest points are extracted from the input video, and then the video block of describing local motion region is extracted by means of spatio-temporal interest points. Furthermore, we propose a novel multidirectional projection optical flow histogram (DPHOF) descriptor to represent the video volume combined with the orientation histograms of 3D gradient orientations (3DHOG) and use SOM to generate the global video descriptor. Finally, the KNN is employed as classifier. This method is validated on the KTH and UCF-YT datasets with good recognition results. Experiment shows that the DPHOF descriptor proposed can efficiently represent the spatio-temporal interest points and is better than HOG3D and HOF. And the global video descriptor constructed by SOM can express the video features efficiently. The proposed method has better recognition effect.

Key words: spatio-temporal interest points (STIPs); orientation histograms of 3D gradient orientations (3DHOG); optical flow histogram (HOF); self-organizing feature map (SOM)

0 引言

近年来,人体行为识别已成为计算机视觉领域的重要研究方向,并在视频监控、人机交互等众多领域得到了广泛的应用^[1]。随着机器视觉得到越来越多的关注,作为其中的热点之一,人体行为识别成为一个重要的研究课题。

人体行为识别中一个至关重要的问题就是人体行为的描述。人体行为描述是从人体动作中提取部分特征信息来描述人体行为。根据当前的研究方法,人体

行为识别研究可以分为两类:基于整体运动信息的方法和基于局部特征的行为识别方法。

基于整体运动信息的方法通常采用光流和形状、边缘、轮廓形状等信息对检测出的整体感兴趣的人体区域进行描述。尽管整体运动信息对实际环境中的行为比较适合,但也面临许多问题,如对遮挡、噪声以及视角的变化比较敏感等。Wang等^[2]利用轨迹特征模拟连续帧间的时间关系;Zhen等^[3]对运动历史图像(MHI)和三个正交平面(TOPS)提取的时空立方体的

收稿日期: 2017-01-06

修回日期: 2017-04-21

网络出版时间: 2017-10-19

基金项目: 广西自然科学基金(2015GXNSFAA139311)

作者简介: 范晓杰(1993-),女,硕士研究生,研究方向为数字图像处理、模式识别;宣士斌,博士,教授,研究方向为数字图像识别、模式识别。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20171019.1624.024.html>

运动和结构信息进行编码, 并采用二维拉普拉斯金字塔编码描述符。

基于局部特征方法是通过在视频中定位一个局部视频块, 通过视频块描述人体运动信息。例如, Mota 等^[4]利用 3D-HOG 特征和光流特征来描述视频行为; Tang 等^[5]提取了视频序列中的 3D-SIFT 特征; LAPTEV 等^[6]结合 HOG 特征和 HOF 特征来描述视频序列中的时空立方体; 张飞燕等^[7]利用 HOF 特征来描述时空立方体, 取得了很好的识别效果。

Li Nijun 等^[8]结合使用 HOG3D 与 SOM 能够有效地进行行为识别, 但没有充分提取时空兴趣点运动信息。HOG3D 作为一种兴趣点描述方法, 能够对兴趣点周围的形态信息进行描述, 但该方法所包含的运动信息较少。为了能更全面高效地描述兴趣点信息, 文中提出一种新的多向投影光流特征直方图 (multidirectional projection optical flow histogram, DPHOF)。不仅能有效地表示光流的特征, 还能体现兴趣点及其邻域的运动情况, 并通过实验对该方法的有效性进行验证。

1 局部时空特征提取

文中算法的第一步就是提取时空兴趣点。为了获得较多的不同尺度的兴趣点, 采用比 Laptev^[9]的 STIP (space-time interest points) 更稠密的 Dollar^[10]的 STIP 作为局部特征。局部时空特征的计算是对视频的局部区域进行计算, 局部区域的选择在时空兴趣点的周围, 以时间和空间尺度为标准选取兴趣点的邻域块。兴趣点的表示是对其邻域块进行描述形成特征向量。最终的视频描述由一些不同位置、不同尺度特征点的特征向量来表示。

1.1 3D 有向梯度直方图 (HOG3D)

由于遵循了 HOG3D^[11]的提取流程, 因此有必要简单介绍一下 HOG3D 的基本思想。STIP 的邻域立方体块被划分为一系列的胞腔 (cell), 同样一个胞腔被划分成一系列的块 (block)。利用“积分视频 (integral video)”计算每个块中的 3D 平均梯度向量, 每个梯度方向的量化通过常规的多面体来进行, 得到每个块的直方图后, 叠加一个胞腔所有块的直方图得到胞腔直方图。最后, 级联 STIP 邻域立方体中所有胞腔的直方图得到 HOG3D 描述子。

假设 STIP 邻域立方体中 x 和 y 方向上有 M 个胞腔, t 方向上有 N 个胞腔, 每个胞腔的直方图维数是 d , 则级联所有胞腔直方图得到 M^2Nd 维的 HOG3D 描述子。实验取 $M=4$, $N=3$, 梯度方向量化到 Klaser^[11]推荐的正 20 面体的面法向量构成的 20×3 的投影矩阵 P 中, 即 $d=20$, 因此 HOG3D 描述子维数是 960。

1.2 多向投影光流直方图 (DPHOF)

传统的光流直方图方法是首先对图像块计算光流, 然后统计多个方向的光流分布情况。但传统的 HOF 描述方法仅能体现光流在兴趣点的特征, 不能体现出其邻域的运动情况。为了保证特征对行为的高描述性, 提出一种新的多向投影光流特征直方图 (DPHOF), 用金字塔 Lucas-Kanade^[12]光流算法来计算光流。光流特征计算完成后, 把对光流方向分布的统计转化为光流在多方向上投影分布的统计, 这样不仅能统计光流的方向分布情况, 也能按照投影的大小对速度分量进行加权。不同行为的光流特征在其速度分量上的分布是有很大的区别的, 用投影的方法对其进行加权更能准确高效地描述光流的特征。下面对 DPHOF 时空立方体描述符的构造进行详细描述。

在 DPHOF 描述方法中, 光流场的计算和 HOF 的计算方式一样, 选用金字塔 Lucas-Kanade 光流算法来计算光流特征。光流特征计算完成后, 开始计算时空兴趣点邻域立方体的描述符, 受 HOG3D 描述符生成方法的启示, 按照同样的流程生成多方向投影光流直方图。实验中, 在兴趣点的 x 和 y 方向上取 $M=4$ 个胞腔, 在 t 方向上取 $N=3$ 个胞腔, 每个胞腔由 $2 \times 2 \times 2$ 个块构成, 计算出每个块中的平均光流 $f_b = [v_{xmean}, v_{ymean}]$, 每个块中光流的量化是通过将其投影到 5×2 的投影矩阵 P 中, 生成光流直方图 h_b :

$$h_b = P \cdot f_b \quad (1)$$

$$P = (\cos \alpha, \sin \alpha)^T \quad (2)$$

其中, α 的取值范围为 $[0^\circ, 180^\circ]$ 并将其平分成 5 个扇形区域。统计每个块的平均光流在各个区域的投影, 得出每个块的投影光流直方图后, 叠加一个胞腔中所有的块直方图得到胞腔直方图。胞腔直方图的维数为 $d=5$ 。最后, 级联 STIP 邻域立方体中所有胞腔的直方图得到时空立方体的 DPHOF 描述子。因此 DPHOF 描述子的维数就是 240, 可以有效减轻“维数灾难”效应。

由上述计算过程可以看出, 利用 DPHOF 在构造光流特征的时空立方体描述子时更加紧凑高效, 通过投影量化使得在统计光流特征时, 不仅体现了光流方向的分布情况, 还更加精确地利用投影大小对光流速度分量加入权值, 保证了特征对立方体信息的高描述性。而且采用的 5 个方向的投影矩阵, 很大程度上减轻了“维数灾难”。

2 基于自组织特征映射 (SOM) 的全局描述子

2.1 SOM 网络

SOM 网络是由芬兰 Helsinki 大学的 Kohonen T 教

授提出的, 又称 Kohonen 网络。Kohonen 认为, 一个神经网络接受外界输入模式时, 将会分为不同的对应区域, 各区域对输入模式有不同的响应特征, 而这个过程是自动完成的。SOM 网络正是根据这一看法提出的, 其特点与人脑的自组织特性相类似。SOM 是一个两层的全连接网络(见图 1), 圆圈代表神经元, 线段标记直接相连的神经元。“竞争(competition)”、“合作(cooperation)”和“自适应(self-adaptation)”是 SOM 的 3 个核心过程。

2.2 全局视频描述子的构造

提取完时空特征后, 就要从所有动作类中随机选取 HOG3D 描述子和 DPHOF 描述子分别训练 SOM 网络。训练完成后, 把所有 HOG3D 描述子送入由 HOG3D 描述子训练的网络, 这样每个 HOG3D 描述子

就会激活一个神经元。最后统计测试结果就可以得到一个神经元击中率直方图, 将这个直方图称为该视频的 HOG3D 击中率直方图。对于 DPHOF 描述子, 以同样的方法送入由 DPHOF 描述子训练的网络进行测试, 同样会得到一个击中率直方图, 称为 DPHOF 击中率直方图。最后把 HOG3D 击中率直方图和 DPHOF 击中率直方图进行归一化处理, 并将两种描述方法的视频归一化直方图级联在一起作为该视频最终的全局描述符, 就由局部的时空特征得到了全局的视频描述子。

3 识别过程和算法

在测试过程中, 最终的判决结果由最终全局描述符的最邻近分类得到, 采用 χ^2 距离作为度量。动作识别流程如图 1 所示。

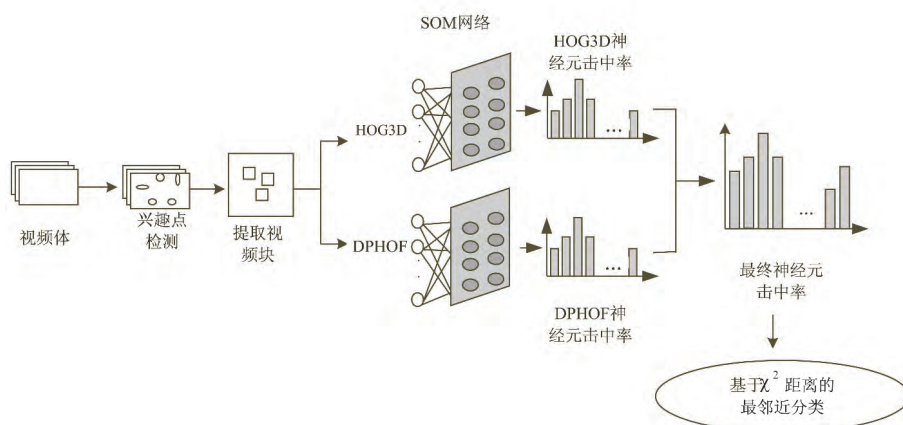


图 1 基于时空特征融合和 SOM 的动作识别流程

基于 HOG3D、DPHOF 和 SOM 的行为识别如下所述:

算法 1: 基于 HOG3D、DPHOF 和 SOM 的行为识别。

输入: 有标签的训练视频序列、测试视频序列;

输出: 测试视频的标签。

(1) 从所有的训练和测试视频中提取多尺度的 Dollar 的 STIP。

(2) 计算每个 STIP 的 HOG3D 描述子和 DPHOF 描述子。

(3) 分别用从训练集中随机选取的 HOG3D 描述子和 DPHOF 描述子训练 SOM 网络。

①初始化具有已知结构的 SOM 网络;

②利用在线学习机制将训练样本输入网络;

③找到对应于当前样本的获胜神经元;

④更新获胜神经元及其邻域神经元的权值;

⑤重复步骤 ②~④, 直至收敛或达到最大迭代次数。

(4) 分别用训练好的 SOM 网络计算所有训练和测试视频的神经元击中率归一化直方图。

(5) 将两种描述方法生成的视频归一化直方图进行级联作为视频的最终全局描述符。

(6) 用基于 χ^2 距离的 NN 分类器分类神经元击中率直方图得到识别结果。

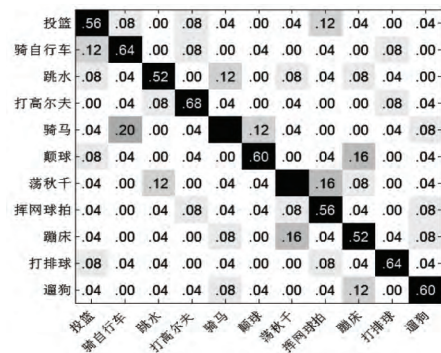
4 实验结果与分析

4.1 实验环境和数据库

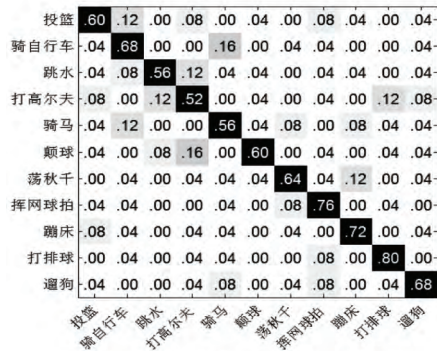
在 3.0 GHz CPU、32 位 Windows 操作系统、Matlab 2012a 的实验环境下, 在 UCF-YouTube、KTH 两个数据库上对文中方法进行验证。两种数据库均采用 5-折叠交叉验证。

4.2 UCF YouTube 数据库

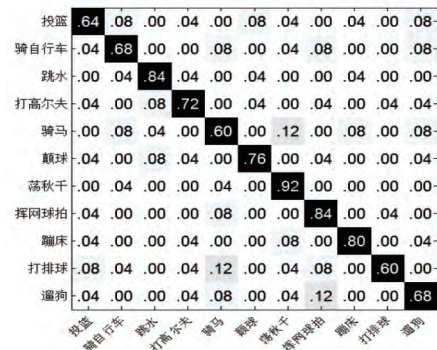
对于 UCF-YouTube^[13] 体育活动数据集, 其数据具有复杂的环境和场景变化, 还有视角、尺度、光照等的变化, 是一个极具挑战的行为识别数据库。该数据库包含 11 种行为, 每种行为在 25 种不同的场景下完成。实验中训练集的大小为 $11 \times 25 \times 100$, 采用 5-折叠交叉验证, 采用迭代 200 次的 12×12 的 SOM 网络进行测试。分别用 HOG3D、DPHOF 以及混合两种特征在数据库上进行测试, 结果如图 2 所示。



(a) HOG3D (平均准确率 56.72%)



(b) DPHOF (平均准确率 64.7%)



(c) HOG3D 和 DPHOF 融合 (平均准确率 73.45%)

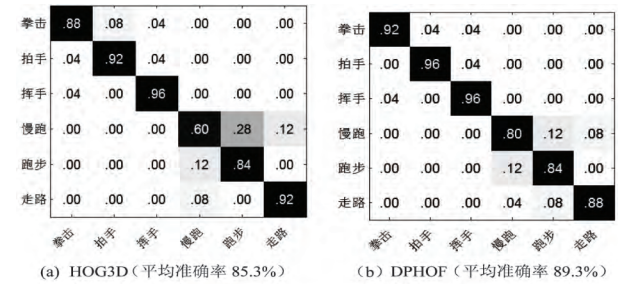
图2 不同方法在UCF-YT数据集上的混淆矩阵

从图2中可看出,提出的DPHOF特征对于复杂的UCF-YT数据集更具有辨别性,能大大地提高行为识别精度。这是因为对于UCF-YT数据库,由于其复杂的背景,加上相机运动会造成背景中许多不感兴趣的STIP,从而影响了SOM构造的全局视频描述符的准确性,而HOG3D描述子易受相机运动的影响,会给识别过程带来许多干扰。而DPHOF描述子作为一种优越的运动特征描述方法,对光照、相机运动的干扰有很好的鲁棒性。并且多向投影方法使得不同行为的光流特征更具辨别力。所以文中的描述方法可以更准确地描述兴趣点特征,而且使用SOM训练击中率直方图来表示视频,不仅具有局部特征,还包含全局特征。所以文中方法取得了更好的识别效果。

4.3 KTH 数据库

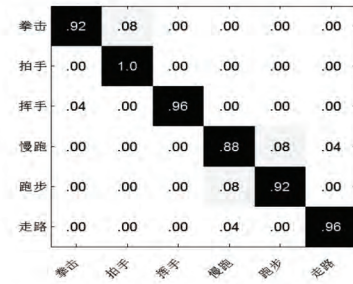
KTH数据库包含6种行为,每种行为在25种不同的场景下完成。实验中训练集的大小为 $6 \times 25 \times 100$,

采用5-折叠交叉验证,采用迭代100次的 10×10 的SOM网络进行测试,结果如图3所示。



(a) HOG3D (平均准确率 85.3%)

(b) DPHOF (平均准确率 89.3%)



(c) HOG3D 和 DPHOF 融合 (平均准确率 94%)

图3 不同方法在KTH数据集上的混淆矩阵

图3表明,在同一数据库下,使用DPHOF描述方法要比单独使用HOG3D的效果好很多,且两种局部描述子与SOM结合构造的全局描述符更能高效表示视频特征。能取得较好的识别率,一方面是由于提出的DPHOF描述子能高效表示空间局部特征;另一方面是与SOM结合构造的全局视频描述符能更好地表示视频特征。使用全局和局部混合特征来进行人体行为识别可以达到更好的识别效果。

5 结束语

提出一种基于混合时空特征和SOM网络的新的行为识别框架,该框架不需要人体检测、跟踪等复杂的预处理步骤。提出一种新的时空特征描述方法(DPHOF),用HOG3D和DPHOF来描述局部空间信息,并结合SOM来构造全局的视频描述符。实验结果表明,提出的DPHOF描述符能高效表示时空兴趣点,且由SOM构造出的全局视频描述子可以高效地表示视频特征。基于SOM的识别框架在识别精确度上取得了很好的效果。

参考文献:

- [1] 李瑞峰,王亮亮,王珂.人体动作行为识别研究综述[J].模式识别与人工智能,2014,27(1):35-48.
- [2] WANG H, SCHMID C. Action recognition with improved trajectories[C]//IEEE international conference on computer vision, [s.l.]: IEEE, 2013: 3551-3558.
- [3] ZHEN X T, SHAO L. A local descriptor based on Laplacian pyramid coding for action recognition[J]. Pattern Recognition

(下转第118页)

参考文献:

- [1] LE T M ,LIU R P ,HEDLEY M.Rogue access point detection and localization[C]//23rd international symposium on personal ,indoor and mobile radio communications. [s.l.]: IEEE 2012: 2489-2493.
- [2] 陈 潮 靳慧云.无线局域网中非法 AP 的定位问题研究[J].信息安全 2010(10): 72-73.
- [3] JIN Y ,SOH W S ,WONG W C.Indoor localization with channel impulse response based fingerprint and nonparametric regression[J].IEEE Transactions on Wireless Communications 2010 9(3): 1120-1127.
- [4] SEN S ,RADUNOVIC B ,CHOUDHURY R R ,et al.You are facing the mona lisa: spot localization using PHY layer information[C]//Proceedings of the 10th international conference on mobile systems ,applications ,and services. [s.l.]: ACM 2012: 183-196.
- [5] WU Kaishun ,XIAO Jiang ,YI Youwen ,et al.CSI-based indoor localization[J].IEEE Transactions on Parallel and Distributed Systems 2013 24(7): 1300-1309.
- [6] 朱 荣 ,白光伟 ,沈 航 ,等.基于贝叶斯过滤法的 CSI 室内定位方法[J].计算机工程与设计 ,2015 ,36(3): 567-571.
- [7] 邓晓华.基于 CSI 的被动式室内定位与目标计数方法研究[D].杭州: 杭州电子科技大学 2014.
- [8] XIA P F ,ZHOU S L ,GIANNAKIS G B.Adaptive MIMO-OFDM based on partial channel state information[C]//IEEE signal processing workshop on signal processing advances in wireless communications. [s.l.]: IEEE 2004: 551-555.
- [9] CORREA J ,ADAMS M ,PEREZ C.A Dirac delta mixture-based random finite set filter[C]//International conference on control ,automation and information sciences. [s.l.]: [s.n.] 2015: 231-238.
- [10] ABBASI A ,LIU Huaping.Improved line-of-sight/non-line-of-sight classification methods for pulsed ultra-wideband localization[J].IET Communications 2014 8(5): 680-688.
- [11] ZHOU Zimu ,YANG Zheng ,WU Chenshu ,et al.WiFi-based indoor line-of-sight identification[J].IEEE Transactions on Wireless Communications 2015 14(11): 6125-6136.
- [12] CONTI A ,DARDARI D ,GUERRA M ,et al.Experimental characterization of diversity navigation [J]. IEEE Systems Journal 2014 8(1): 115-124.
- [13] GUVENC L ,CHONG C C ,WATANABE F ,et al.NLOS identification and weighted least-squares localization for UWB systems using multipath channel statistics [J].EURASIP Journal on Advances in Signal Processing 2008 2008: 36.
- [14] MUCCHI L ,MARCOCCI P.A new parameter for UWB indoor channel profile identification[J].IEEE Transactions on Wireless Communications 2009 8(4): 1597-1602.
- [15] BENEDETTO F ,GIUNTA G ,TOSCANO A ,et al.Dynamic LOS/NLOS statistical discrimination of wireless mobile channels[C]//65th vehicular technology conference. [s.l.]: IEEE 2007: 3071-3075.
- [16] TEPEDELENIOLU C ,ABDI A ,GIANNAKIS G B.The Ricean K factor: estimation and performance analysis [J]. IEEE Transactions on Wireless Communications ,2003 ,2(4): 799-810.
- [17] WU Kaishun ,XIAO Jiang ,YI Youwen ,et al.FILA: fine-grained indoor localization[C]//Proceedings of IEEE INFOCOM. [s.l.]: IEEE 2012: 2210-2218.
- [18] HALPERIN D ,HU Wenjun ,SHETH A ,et al.Tool release: gathering 802.11n traces with channel state information[J]. ACM SIGCOMM Computer Communication Review ,2011 ,41(1): 53.
- [19] LAPTEV I.On space-time interest points [J].International Journal of Computer Vision 2005 64(2-3): 107-123.
- [20] DALLAR P ,RABAUD V ,COTTRELL G ,et al.Behavior recognition via sparse spatio-temporal features [C]//IEEE international workshop on performance evaluation of tracking and surveillance.Beijing ,China: IEEE 2005: 65-72.
- [21] KLASER A ,MARSZALEK M ,SCHMID C.A spatio-temporal descriptor based on 3D-gradients [C]//British machine vision conference. [s.l.]: [s.n.] 2008.
- [22] BOUGUET J Y.Pyramidal implementation of the Lucas Kanade feature tracker: description of the algorithm[R]. [s.l.]: Intel Corporation Microprocessor Research Labs 2000.
- [23] LIU J ,LUO J ,SHAN M.Recognizing realistic actions from videos "in the wild" [C]//Proceedings of the computer vision and pattern recognition. [s.l.]: [s.n.] 2009.
- [24] MOTA V F ,PEREZ E A ,MACIEL L M ,et al.A tensor motion descriptor based on histograms of gradients and optical flow [J].Pattern Recognition Letters 2014 39(4): 85-91.
- [25] TANG X Q ,XIAO G Q.Action recognition based on maximum entropy fuzzy clustering algorithm [M]//Foundations of intelligent systems.Berlin: Springer 2014: 155-164.
- [26] LAPTEV I ,MARSZALEK M ,SCHMID C ,et al.Learning realistic human actions from movies [C]//26th IEEE conference on computer vision and pattern recognition.Anchorage , AK ,United States: IEEE 2008: 1-8.
- [27] 张飞燕 ,李俊峰.基于光流速度分量加权的人体行为识别[J].浙江理工大学学报 2015 33(1): 115-123.
- [28] LI Nijun ,CHENG Xu ,ZHANG Suofei ,et al.Realistic human action recognition by Fast HOG3D and self-organization feature map[J].Machine Vision and Applications 2014 25(7): 1793-1812.

(上接第 101 页)

Letters 2013 34(15): 1899-1905.

- [4] MOTA V F ,PEREZ E A ,MACIEL L M ,et al.A tensor motion descriptor based on histograms of gradients and optical flow [J].Pattern Recognition Letters 2014 39(4): 85-91.
- [5] TANG X Q ,XIAO G Q.Action recognition based on maximum entropy fuzzy clustering algorithm [M]//Foundations of intelligent systems.Berlin: Springer 2014: 155-164.
- [6] LAPTEV I ,MARSZALEK M ,SCHMID C ,et al.Learning realistic human actions from movies [C]//26th IEEE conference on computer vision and pattern recognition.Anchorage , AK ,United States: IEEE 2008: 1-8.
- [7] 张飞燕 ,李俊峰.基于光流速度分量加权的人体行为识别[J].浙江理工大学学报 2015 33(1): 115-123.
- [8] LI Nijun ,CHENG Xu ,ZHANG Suofei ,et al.Realistic human action recognition by Fast HOG3D and self-organization feature map[J].Machine Vision and Applications 2014 25(7):