结合码本优化和特征融合的人体行为识别方法

石爱辉[1],程 勇[2],曹雪虹[1,2]
(1.南京邮电大学 通信与信息工程学院,江苏 南京 210003;
2.南京工程学院 通信工程学院,江苏 南京 211167)

摘 要: 为了提高视频序列中人体行为识别的正确率,提出了一种结合两层 K-means 聚类优化码本和视频表达级特征融合的行为识别方法。首先对训练集视频提取出的时空兴趣点利用梯度方向直方图(HOG)和光流直方图(HOF)进行描述,并对属于不同视频以及不同种类动作视频的描述子使用两层 K-means 聚类形成各自更具有代表性的视觉词汇,从而提高码本的表达能力。然后将表示每个视频的 HOG 和 HOF 描述子分别作为码本优化后的词袋模型的输入,得到两种不同的视频全局表达并进行特征融合,由于 HOG 和 HOF 描述子在形成视频表达级特征时相关性较大,融合后的特征更具区分性和分类鲁棒性。最后采用支持向量机对融合后的特征进行分类识别。实验结果表明,该方法能够有效地提高识别率。

关键词: 词袋模型; 两层 K-means 聚类; 视频表达级特征融合; 行为识别
中图分类号: TP391 **文献标识码:** A **文章编号:** 1673-629X(2018)02-0107-05
doi: 10.3969/j.issn.1673-629X.2018.02.024

A Human Action Recognition Method Combined with Codebook Optimization and Feature Fusion

SHI Ai-hui[1],CHENG Yong[2],CAO Xue-hong[1,2]
(1.School of Communications and Information Technology,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
2.School of Communication Engineering,Nanjing Institute of Technology,Nanjing 211167,China)

Abstract: In order to improve the accuracy of human actions recognition in video sequence,we present an actions recognition method which combines two-level K-means clustering with video-level descriptor feature fusion.Firstly the space-time interest points extracted by video in training set are described by histogram of oriented gradient (HOG) and histograms optical flow (HOF),and the descriptors of different video and different kinds of motion video are formed their representative visual vocabulary respectively through K-means clustering with two levels,thus improving the expression of the codebook.Taking the descriptors of HOF and HOG as the input of the bag of word model respectively,the two different global expressions of video are obtained and fused in features.Due to the high correlation when the descriptors of HOG and HOF forming the characteristics of the video expression level,the fused features are distinguishing and robust in classification.Finally,the support vector machine (SVM) is adopted for classification and recognition to characteristics of fusion.The experiments show that the proposed method can improve the accuracy of recognition effectively.

Key words: bag of word model; two-level K-means clustering; video representation-level feature fusion; action recognition

0 引 言

人体行为识别研究在智能监控、人机交互等领域具有广阔的应用前景,因而受到越来越多的研究者关注。在实际应用中,由于视频中人体行为动作的多变性、复杂的背景以及摄像机的视角变化等因素,人体行为识别仍是计算机视觉领域的难点和热点问题[1-2]。

近些年涌现了大量的人体行为识别算法,例如基于模板匹配的方法,其主要思路是将不同种类行为视频序列提取的特征数据建立相应的模板,识别时将待测视频提取的特征数据与模板进行比较匹配。这种方法虽然计算量小,实现相对简单,但需要存储各种动作视频的特征数据作为模板,存储代价较大[3]。基于光

收稿日期: 2017-03-14 **修回日期:** 2017-07-20 **网络出版时间:** 2017-11-15
基金项目: 江苏省科技计划项目(2016008-06);闽江学院福建省信息处理与智能控制重点实验室开放课题(MJUKF201712)
作者简介: 石爱辉(1992-),男,硕士研究生,研究方向为现代通信中的智能信号处理;程 勇,副教授,博士,研究方向为图像处理与模式识别;曹雪虹,教授,博导,研究方向为现代通信中的智能信号处理。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20171115.1436.058.html

流的方法，主要利用光流这种基于视频中帧与帧之间变化的运动信息。文献[4]提出在基于视频的人体行为识别过程中，可以将视频序列中的光流信息转化为更能明显区分不同动作差异的运动特征，利用不同层面的运动特征参数表示视频序列中的光流信息。基于兴趣点的方法是利用 histogram of oriented gradient (HOG)[5] 和 histograms optical flow (HOF)[5] 等描述子对视频中检测到的时空兴趣点局部区域进行描述。由于时空兴趣点是对视频中运动显著区域的描述，包含了丰富的动作细节信息，因此具有较好的抗噪声性能。

文献[6]提出基于密集轨迹的人体行为识别方法，即通过跟踪光流场密集采样的特征点来获得轨迹，并计算轨迹位移向量及其轨迹中子时空块的梯度方向直方图(HOG)、光流直方图(HOF)和运动边界直方图(MBH)[7]作为视频序列的底层局部特征描述子，然后将这些局部特征描述子作为视觉词袋模型(BoVW)[8]的输入获得视频序列的全局表达，最后将这种视频全局表达作为支持向量机的输入进行分类识别，取得了较好的识别效果。

在目前的行为识别算法中，基于视觉词袋模型的方法是研究热点之一。在传统的视觉词袋模型中，对所有视频的一部分局部特征描述子进行一次 k-means 聚类而形成的码本，其视觉词汇并不具有很好的代表性。而有效的字典学习是视觉词袋模型的关键步骤，文中提出对视频中提取的局部特征描述子根据取自不同视频和不同种类动作进行两层 k-means 聚类，形成更有代表性和区分度的码本。特征融合是一种使得特征描述鲁棒性更强的有效方法，对于视频中提取的两种局部特征描述子 HOG 和 HOF，在分别形成全局视频表达后进行融合，融合后的全局视频表达特征更具有区分性和鲁棒性。

1 文中算法

1.1 算法框架

文中算法框架如图 1 所示。首先对视频中的时空兴趣点进行检测，然后利用 HOG 和 HOF 作为局部特征描述子对兴趣点进行描述，接着将 HOG 和 HOF 描述子分别作为词袋模型的输入，得到两种不同的视频级全局表达，将这两种视频级全局表达进行融合作为最终的视频级表达特征，最后将其代入到支持向量机中对行为动作进行分类。

1.2 局部特征描述子

在人体行为识别的课题研究中，由于进行实验仿真所使用到的数据库中的视频相对简单和稳定，因而不需要对其中的人体进行跟踪和检测，所以对视频提取局部特征是一种常见的方法。

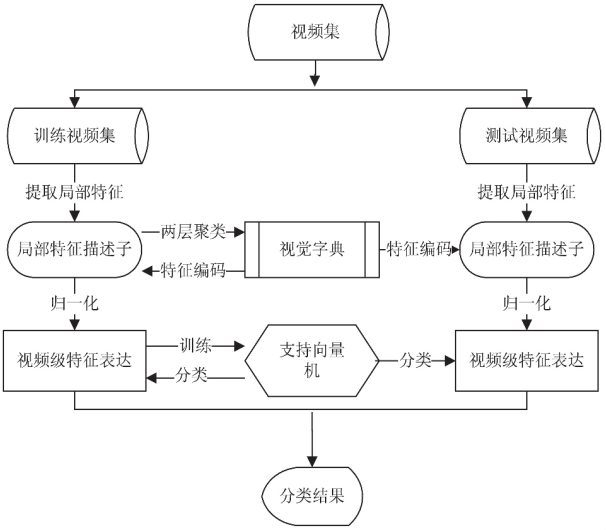


图 1 算法框架

对于视频中的时空兴趣点，一种具有鲁棒性好、适应性强的时空特征被广泛使用，其通过 Harris 角点检测[9]扩展到包括时间维的三维空间即 Harris-3D 获得。Harris 角点检测的基本原理是选择不同尺度的局部空间，计算其中每个像素二阶矩阵的特征值，对于某个像素点的特征值为局部最大值时被视为角点。对于包括时间维的三维空间，其中的尺度空间包括空间尺度和时间尺度，对于被认为是时空兴趣点的像素点在空间域和时间域会同时有显著的变化，因此在时空域上表示一个图像序列 $V(\cdot)$，利用其与高斯核函数作卷积获得其尺度空间表示：

$$L(\cdot;\sigma_l^2,\tau_l^2) = g(\cdot;\sigma_l^2,\tau_l^2) * V(\cdot) \tag{1}$$

$$g(x,y,t;\sigma_l^2,\tau_l^2) = \frac{1}{\sqrt{(2\pi)^2\sigma_l^4,\tau_l^2}}\exp(-(x^2+y^2)/2\sigma_l^2 - t^2/2\tau_l^2) \tag{2}$$

其中，$V(x,y,t)$ 表示视频中的像素点；g 表示高斯核；σ_l^2,τ_l^2 表示空间和时间上的尺度因子。

参照 Harris 角点检测的像素二阶矩阵，在时空尺度空间的二阶矩阵可表示为：

$$\mu(\cdot;\sigma_i^2,\tau_i^2) = g(\cdot;\sigma_i^2,\tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \tag{3}$$

其中，$g(\cdot;\sigma_i^2,\tau_i^2)$ 为平滑函数，其在空间域和时间域的平滑尺度分别为 $\sigma_i^2 = s\sigma_l^2$，$\tau_i^2 = s\tau_l^2$；一阶导数 L 是在用高斯核模糊后的视频帧 f 上求得，定义为：

$$L_x(\cdot;\sigma_i^2,\tau_i^2) = \partial_x(g * f) \tag{4}$$

$$L_y(\cdot;\sigma_i^2,\tau_i^2) = \partial_y(g * f) \tag{5}$$

$$L_t(\cdot;\sigma_i^2,\tau_i^2) = \partial_t(g * f) \tag{6}$$

假设 $\lambda_1,\lambda_2,\lambda_3$ 为 μ 矩阵的特征值，则时空兴趣

点由推广的 Harris 响应函数的局部最大值处定义:

$$H = \det(\boldsymbol{\mu}) - k\,\mathrm{trace}^3(\boldsymbol{\mu}) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3 \quad (7)$$

视频中兴趣点可以根据参数进行多尺度提取,检测到兴趣点后,为了在兴趣点处提取 HOG/HOF 特征,文献 [10] 在特征点处抽取大小为 $(2k\sigma_i; 2k\sigma_i; 2k\tau)$ 的局部视频块 $(k=9)$,然后将其分成空时为 3* 3* 2 的网格。对于每个网格,将梯度方向量化为 4,光流方向量化为 5(其中包含一个静止方向),从而一个时空兴趣点可以通过 72 维的 HOG 和 90 维的 HOF 来加以描述。

1.3　码本优化

传统的 BoVW 中利用全局描述子对视频进行描述,主要分成三个步骤: 首先利用 k-means 聚类算法对视频中获取的局部特征描述子进行聚类形成字典,然后根据底层特征描述子和字典形成频率直方图对视频进行描述,最后对直方图进行归一化处理后作为视频的中层表达。在视觉词袋模型中对视频提取的特征描述子进行聚类形成字典时,文中提出对视频中提取的特征描述子进行两层聚类优化码本,提高码本的表达能力。其中两层 k-means 聚类的过程如图 2 所示,首先对训练集中的每一个视频提取的 HOG 和 HOF 特征描述子分别进行 k-means 聚类,聚类数目为视频中,兴趣点数目的 25%,然后对同种行为动作的视频的聚类结果再进行 k-means 聚类,聚类数目大小为 K,最后将所有动作种类的聚类结果作为视觉词汇连接成码本,这样的码本更有代表性和区分度。除此之外,两层 k-means 聚类还能够降低对实验仿真内存的要求并减少聚类所花的时间。其中 K 的大小可以根据仿真实验的效果在一个范围内进行选择。

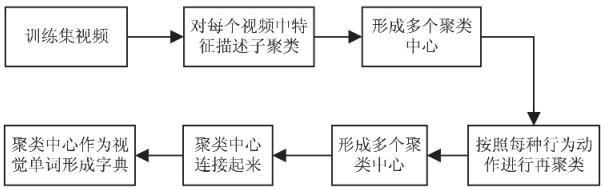


图 2　对每个视频以及每种动作进行聚类的流程

图 3 是分别使用传统聚类方法形成的码本和优化码本在 KTH 数据库中鼓掌和挥手动作的直方图表示。利用以上构建的字典,视觉单词的位置与相应的行为动作之间有了对应关系,从而改变了直方图的分布情况。与传统词袋模型中使用的码本相比,在一定程度上提高了同种动作视频的视觉单词直方图分布的相似程度,而使得不同动作类别的视觉直方图分布的差异明显。

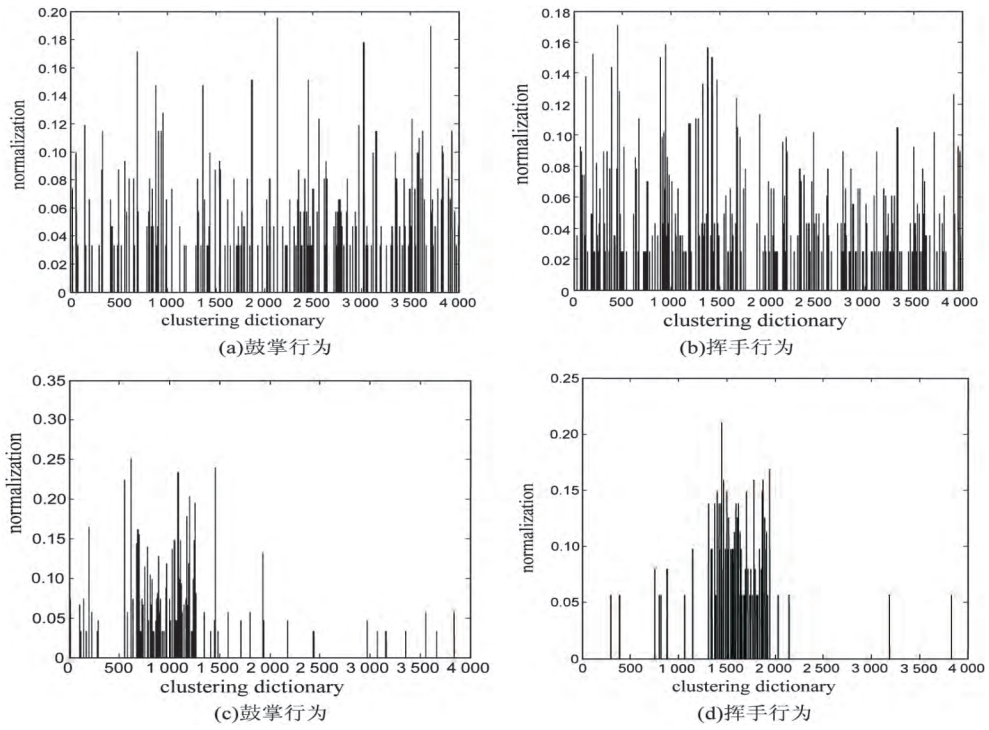


图 3　两种不同行为直方图的表示

1.4　视频表达级特征融合

HOG 特征描述子包含了视频图像中的表观形状信息,而 HOF 特征描述子包含了视频图像中的运动信息。但文献 [11] 的实验结果表明,仅仅使用 HOF 特征描述子比使用 HOF 和 HOG 特征描述子在描述子级融合的效果好,对于描述子级融合是将描述视频中局部特征的多个描述子串联在一起形成单个的描述子,然后将其送入到 BoVW 框架中获取全局视频表达。

针对这种情况,文中将 HOF 和 HOG 描述子在视频表达级层面进行融合,其过程如图 4 所示。视频表达级的融合是将描述视频中局部特征的 HOF 和 HOG 描述子分别送入 BoVW 框架中获取到两种不同的视频全局表达,然后对这两种视频全局表达进行融合作为最终的视频表达级特征。对于 HOG 和 HOF 这两种不同的特征描述子,在形成视频表达级描述子相关性较大时,视频表达级层面上的特征融合比在局部特征描述子级层面上的直接融合效果要好。

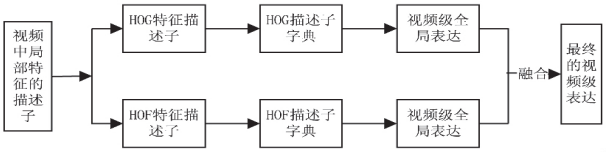


图 4 视频级表达特征融合方法

1.5 支持向量机分类器

使用支持向量机(SVM)分类器进行分类识别。SVM 的主要思想:在空间 H 中,如果要将训练数据集 (x_1,y_1),(x_2,y_2),…,(x_m,y_m) 分成两类 $y_i \in \{-1,+1\}$,对于所有能将数据集分成两类的超平面 $wx+b=0$,选择一个最优决策超平面使得该平面两侧距离该平面最近的两类样本之间的距离最大化,其中 w 和 b 的值可以通过 Lagrange 乘数 α_i 求解约束条件下的极小值问题求得[13]。

$$f(x)=\mathrm{sgn}\left(\sum_i \alpha_i y_i K(\boldsymbol{x}_i,x)+b\right) \tag{8}$$

其中,对应非零 α_i 的 $\boldsymbol{x}_i$ 向量称为支持向量。引入核函数 $K(x,y)$ 巧妙地解决了在高维空间中的内积运算,较好地解决了非线性分类问题。文中使用的是线性核函数。

2 实验仿真分析

2.1 数据集

为了验证文中算法的有效性,选择两个比较经典的数据集(KTH 和 Weizmann)进行仿真实验。

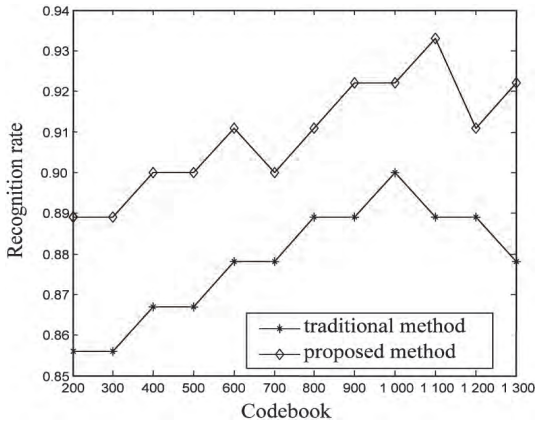
KTH 数据集包括 6 类行为动作(walking、jogging、running、boxing、hand waving、hand clapping),是由 25 个不同的人在 4 种不同场景下(室内、室外、尺度变化和衣着变化)采集完成。所有视频背景相对静止,摄像机的运动比较轻微,视频的帧率为 25 帧/s,分辨率为 160x120,整个数据集包含了 599 个视频文件。将其中 16 人的所有动作视频作为训练集,其余 9 人的所有动作视频作为测试集。最后的识别率是由测试集中所有被正确识别出的视频个数计算得到。

Weizmann 数据集包括 10 种不同类型的行为视频(bend、jump、jack、pjump、run、side、skip、walk、wave1、wave2),每种动作由 9 个不同的人所展示,采用的方法是将其中 1 人的所有动作视频作为测试集,其他人的所有动作视频作为训练集,循环 9 次,最后将平均正确率作为识别率。

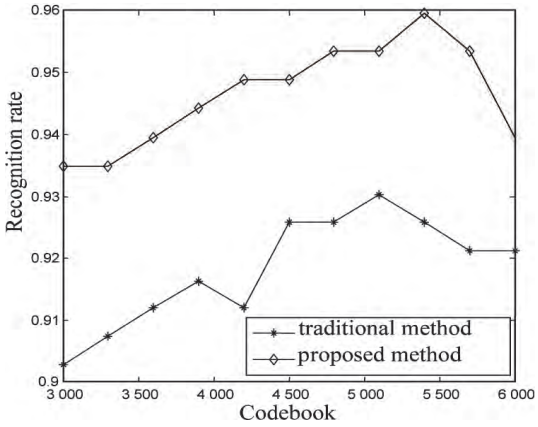
2.2 实验结果

图 5 分别是使用一次 k-means 方法和使用两次 k-means方法对局部特征描述子进行聚类,构建不同数量的视觉词汇在 KTH 和 Weizmann 数据集上的识别率对比曲线。可以看出,在不同数量的视觉词汇下,使用优化后的码本的识别率明显高于使用传统聚类方法形成的码本的识别率。Weizmann 数据集中的视频序列的长度较短,视频中提取的时空兴趣点的数目也较少,在构建码本时视觉词汇的数量也相应减少,导致在 Weizmann 数据集上的识别率明显低于 KTH 数据集上的识别率。

同时,识别率总体上是随着码本大小增加而不断提高,当码本到达一定的数目后识别率基本保持不变。而当字典过大时,一些视频中的时空兴趣点较少对应到码本上,词汇减少不能有效地描述视频。相较于使用传统的聚类方法形成的码本,码本优化后在 KTH 和 Weizmann 数据集上的识别率提升了 3%左右,证明了文中方法的有效性。



(a) 使用一次 k-means 方法



(b) 使用两次 k-means 方法

图 5 识别率对比曲线

使用单个 HOF 特征描述子以及优化后的码本形

成的频率直方图作为视频表达级描述子在 KTH 和 Weizmann 数据集上的识别率，分别为 95.8% 和 93.3%。而使用 HOG 和 HOF 特征描述子以及各自优化后的码本形成的两种频率直方图融合作为最终的视频表达级描述子在 Weizmann 与 KTH 数据集上仿真实验效果最好时的识别率，分别为 96.7% 和 94.4%。从实验结果可以看出，结合码本优化和视频表达级特征融合的方法与传统方法相比，在 KTH 与 Weizmann 数据集上的识别率均有不同程度的提升，表明了该方法的有效性。

表 1 列出了文中方法与近年来人体行为识别研究课题在 KTH 和 Weizmann 数据集上识别率的比较。与其他方法相比，文中方法在这两个数据库上均取得了较高的识别率。

表 1　各算法平均识别率对比　%

方法	KTH	Weizmann
文献 [11]	91.4	84.3
文献 [13]	91.5	93.5
文献 [14]	81.16	85.2
文献 [15]	90.6	80.2
文中	96.7	94.4

3　结束语

为了提高视觉词袋模型应用在人体行为识别研究课题的识别率，引入了一种结合多层 k-means 聚类与视频级表达特征融合的行为识别算法，降低了对内存的要求并减少了聚类所用的时间，可以更有效地描述视频。仿真结果表明，该方法在两个经典数据集上的识别率高于大多数算法。针对如何提高易混淆动作的识别率以及选用其他编码方法替代 VQ 编码将是下一步的研究工作。

参考文献:

[1] 王　博，李　燕. 视频序列中的时空兴趣点检测及其自适应分析 [J]. 计算机技术与发展，2014，24(4)：49-52.

[2] 刘雨娇，范　勇，高　琳，等. 基于时空深度特征的人体行为识别算法 [J]. 计算机工程，2015，41(5)：259-263.

[3] 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.

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

[5] WANG H，YI Y. Tracking salient key points for human action recognition [C] //IEEE international conference on systems，man，and cybernetics. [s.l.]：IEEE，2015：3048-3053.

[6] WANG Heng，KLASER A，SCHMID C，et al. Action recognition by dense trajectories [C] //Proceedings of IEEE international conference on computer vision and pattern recognition. Washington D C，USA：IEEE Press，2011：3169-3176.

[7] LI Q，CHENG H，ZHOU Y，et al. Human action recognition using improved salient dense trajectories [J]. Computational Intelligence & Neuroscience，2016，2016：6750459.

[8] FARAKI M，PALHANG M，SANDERSON C. Log-Euclidean bag of words for human action recognition [J]. IET Computer Vision，2015，9(3)：331-339.

[9] HARRIS C，STEPHENS M. A combined corner and edge detector [C] //Proceedings of alvey vision conference. [s.l.]：[s.n.]，1988：147-151.

[10] LAPTEV I，MARSZALEK M，SCHMID C，et al. Learning realistic human actions from movies [C] //IEEE computer society conference on computer vision and pattern recognition. [s.l.]：IEEE，2008：1-8.

[11] KLASER A，MARSZALEK M，SCHMID C. A spatio-temporal descriptor based on 3D-gradients [C] //British machine vision conference. [s.l.]：[s.n.]，2008：995-1004.

[12] 边肇祺，张学工. 模式识别 [M]. 第 2 版. 北京：清华大学出版社，2000：296-303.

[13] LU M，ZHANG L. Action recognition by fusing spatial-temporal appearance and the local distribution of interest points [C] //International conference on future computer and communication engineering. [s.l.]：[s.n.]，2014：75-78.

[14] DOLLAR P，RABAUD V，COTTRELL G，et al. Behavior recognition via sparse spatio-temporal features [C] //IEEE international workshop on visual surveillance & performance evaluation of tracking & surveillance. [s.l.]：IEEE，2005：65-72.

[15] TU H B，XIA L M，WANG Z W. The complex action recognition via the correlated topic model [J]. Scientific World Journal，2014，2014：810185.