基于兴趣点统计特征的双人交互行为预测算法

姬晓飞,谢 旋
(沈阳航空航天大学 自动化学院,辽宁 沈阳 110136)

摘 要:针对一些复杂敏感场景下需要快速及时地对人类交互行为做出预测的问题,提出了一种基于兴趣点统计特征的双人交互行为预测方法。该方法首先对动作视频提取时空兴趣点,并对其进行 3D-SIFT 描述,然后利用词袋方法对动作视频进行表示。在训练阶段,利用高斯模型建立不同时间比例下每个动作的预测模型。在动作预测阶段,对于一个未知长度的动作视频,提取其词袋表示,并将其与所建立的不同时间长度的预测模型进行比较,得到与各模型之间的预测相似概率,最终实现对该交互行为的识别预测。利用 UT-interaction 数据库对该方法进行测试的实验结果表明,该方法易于实现,实时性好,并具有较好的预测效果。

关键词:双人交互预测;兴趣点统计特征;词袋;高斯模型;概率预测
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2019)07-0039-04
doi:10.3969/j.issn.1673-629X.2019.07.008

Human Interaction Prediction Algorithm Based on Statistical Features of Interest Points

JI Xiao-fei, XIE Xuan
(School of Automation, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: A human interaction prediction algorithm based on statistical features of interest points method is proposed to solve the problem that human interaction needs to be predicted quickly and timely in some complex and sensitive scenarios. First, the spatio-temporal interest points are extracted and performed 3D-SIFT description, then the bag of words is used to represent the action video. In the training, Gaussian models are used to establish the action model for each action at different time scales. In the prediction, the bag of words representation is extracted and compared with the established prediction models of different time lengths to obtain similar prediction probabilities between the models for an action video of unknown length. Finally, the recognition and prediction of interaction prediction is completed. The experiment on UT-interaction dataset demonstrates that the proposed approach is easy to implement with better real-time performance and predictive effect.

Key words: interaction prediction; statistical features of interest points; bag of words; Gaussian models; probability prediction

1 概 述

基于视频的双人交互行为识别与理解是图像处理与计算机视觉领域中备受关注的前沿方向,它利用视频分析技术从包含人的图像序列或视频中检测、跟踪、识别人体及动作对象,并对其行为进行理解和描述[1-3]。目前大部分研究关注的都是对发生行为的事后检测,而在很多现实场景中,需要系统能够对正在执行的、未完成的行为进行提早的预测。人类动作预测与动作分类不同,动作预测系统需要在动作执行过程中做出“哪些动作行为发生”的决定。双人交互行为预测具有重大的现实意义,如:在两个人的打架行为恶化之前对其进行检测,将使视频监控系统具有阻止犯罪行为发生的能力,使视频资源发挥更大的作用。

Ryoo[1]率先提出动态 BoW(bag of word)的概率统计方法解决双人交互行为的预测问题,采用时空特征的整体直方图形式对动作进行表示,而后有效的建模特征随时间变化的分布情况实现动作预测。该方法不仅简单易行,并且提供了双人交互行为预测基本框架。在此基础上,Yu 等[4]提出了一种时空隐式形状模型(spatial-temporal implicit shape model, STISM)表示

收稿日期:2018-07-12 **修回日期**:2018-11-15 **网络出版时间**:2019-03-06
基金项目:辽宁省自然科学基金(201602557);辽宁省科技公益研究基金(2016002006);辽宁省高等学校优秀人才支持计划项目(LR2015034)
作者简介:姬晓飞(1978-),女,副教授,硕导,研究方向为视频分析与处理、模式识别;谢 旋(1994-),女,硕士,研究方向为模式识别、视频分析。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190306.0901.004.html

局部时空特征的时空结构,并采用多类平衡的随机森林方法匹配实现双人交互行为预测。Li 等[5]通过监测运动速度,将长时间活动分解编码成有意义的行动单元序列,然后引入概率后缀树(probabilistic suffix tree,PST)表示动作单元之间的马尔可夫依赖关系;最后利用预测累积函数(predictive accumulative function,PAF)描述各种活动的可预测性。以上均是采用传统的概率模型统计的方法,识别与预测的准确率均不是很高。近些年,基于卷积神经网络(convolutional neural network,CNN)的方法开始被应用到双人交互行为预测领域。Ke 等[6]从部分序列的连续视频帧计算光流图像,以分别捕获每个 RGB 帧和每个光流图像对全局和局部上下文的依赖性,然后利用长时短期记忆(long short term memory,LSTM)网络学习包括空间和时间信息的结构模型,最后引入排名分数融合方法预测交互类别。但是最佳权重的选择具有随机性。Ke 等[7]将 CNN 应用于视频流编码图像以学习人类交互预测的时间信息,利用几个连续的光流图像的特征来学习随时间变化的规律。但是这种方法只利用了时间特征,缺乏人体姿态的空间特征信息描述,并且需要学习大量的样本,计算复杂度相对较高。

基于以上分析,文献[1]提出的方法简单有效,且可实现性强,其不足之处是实现双人交互行为预测与识别一体化效果不理想,预测与识别准确率较低。文中在文献[1]的基础上,提出了词袋模型与多时间比例动作模型概率融合的方法,以实现双人交互行为预测与识别一体化。文中算法的具体流程见图 1。

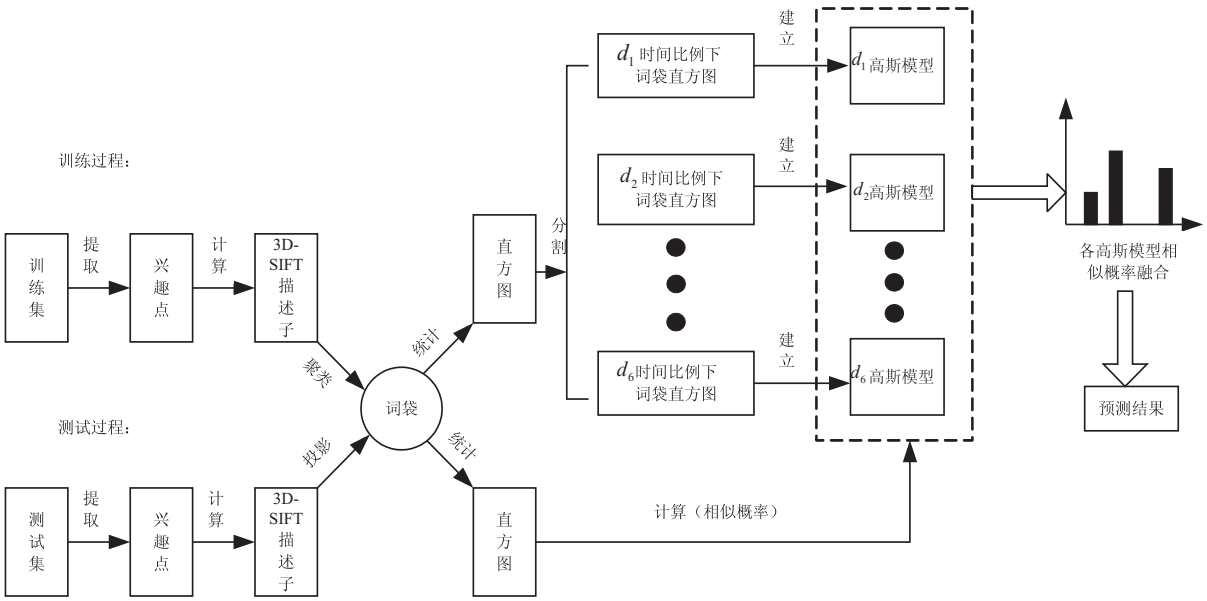

图 1 算法流程

首先,训练集视频进行兴趣点提取和 3D-SIFT 特征描述[8];然后将训练数据分割成不同时间比例数据,并用词袋方法得到在不同时间比例下的视频直方图表示;最后,利用高斯模型建立各不同时间比例数据下的动作模型。当给定未知长度测试视频,进行特征描述后形成一个词袋直方图表示,计算其与训练好的不同时间比例下各高斯模型的相似概率,判别测试视频所属动作类别。通过大量实验验证,该方法在保证一定预测准确率的同时,也得到了较好的识别效果。

2 兴趣点特征提取与描述

局部特征具有可以描述具有显著变化运动信息的优点[9],兴趣点是目前比较常用的一种局部特征[10],因此文中采用其作为基础特征。通过对兴趣点的邻域进行有效描述,能够得到代表此图像序列的局部信息特征[11]。3D-SIFT 描述算子是一种三维时空梯度方向直方图,能准确地捕捉到视频数据的时空特性的本质[12]。为了充分利用上下文的运动信息,文中对整个视频采集兴趣点,然后进行 3D-SIFT 特征描述,即:在兴趣点邻域内建立 3D 球形体积块,在每个体积块里进行梯度累积。

3 动作预测模型的建立

兴趣点特征与词袋模型[13]相结合可以方便地得到动作视频的词袋直方图表示。通常采用 k-means 方法对训练数据的所有局部特征表示进行聚类形成词典,然后将一个动作视频的所有局部特征向词典投影,最终统计词典中视觉单词在视频中出现的频率,形成动作视频的统计直方图表示。该方法已广泛用于人体动作识别。为了将此框架运用于双人交互行为预测,将训练视频按照一定比例分割成不同时间长度的子训练集,即将兴趣点 3D-SIFT 描述的训练数据,按照不

同时间比例分割成不同时间长度的子训练数据,每个子训练数据用一个词袋直方图表示。令 O_l 表示动作视频, d_i 表示其时间比例, $h_{d_i}(O_l)$ 表示 d_i 时间比例下 O_l 的子训练数据直方图。v_w 表示第 w 个视觉单词,则每一个特征直方图 $h_{d_i}(O_l)$ 的第 w 个词袋的值为:

$$h_{d_i}(O_l)[w] = |\{f | f \in v_w \wedge t_f < d_i\}| \quad (1)$$

其中, f 表示视频 O_l 提取的特征; t_f 表示其时间位置。

每一个 $h_{d_i}(O_l)$ 描述了时间比例为 d_i 的时空特征直方图随时间变化的分布情况,示例如图 2 所示。

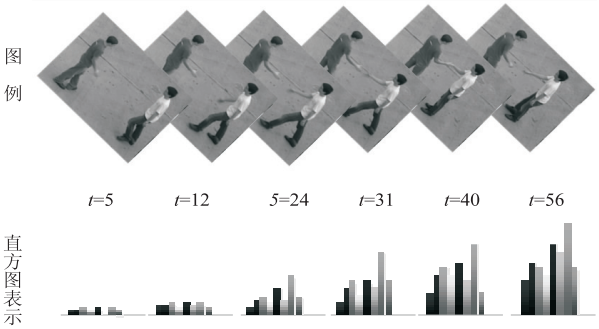

图 2　握手动作整体直方图变化趋势

经大量实验数据发现,同类别动作视频在相同时间比例下的词袋直方图符合一定参数下的高斯分布。因此可以采用高斯模型建立不同时间比例下同类别动作模型,该模型可以较好地反映某种动作执行到某个时间节点时,其词袋直方图的表现形式。

文中将训练视频按不同时间比例 $d_1, d_2, \cdots, d_6$ 分割,并对分割后的训练视频分别建立高斯模型。记 $h^{(1,d)}, h^{(2,d)}, \cdots, h^{(A,d)}$ 分别为当前时间比例为 d 的第 A 类动作的高斯模型,当给定一未知动作视频 O^{test} 时,计算 O^{test} 与当前时间比例 d 下各个动作高斯模型的似然概率,即: $p(O^{\text{test}} | h^{(1,d)}), p(O^{\text{test}} | h^{(2,d)}), \cdots, p(O^{\text{test}} | h^{(A,d)})$。

$$p(O^{\text{test}} | h^{(a,d)}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-[h(O^{\text{test}}) - h_d(a)]^2}{2\sigma^2}} \quad (2)$$

其中, d 为训练视频的当前时间比例; A 为 d 时间比例下动作模型类别数; $h^{(a,d)}$ 为动作 a 的直方图对应的高斯模型; $h(O^{\text{test}})$ 为未知动作视频 O^{test} 的直方图高斯模型; σ^2 描述的是动作 a 在时间比例 d 下高斯模型的相同变量。

4　概率预测策略

给定一时间长度为 t 的测试视频 O^{test} (t 未知),计算 O^{test} 与各类动作在不同时间比例下高斯模型的相似概率,依据概率值大小判别测试视频与训练集中各类动作的相似程度。最终将未知动作判别为与其相似度最高的动作模型所属类别。即:

$$A^* = \underset{1 \leqslant a \leqslant A}{\operatorname{argmax}} \left(\sum_{1 \leqslant d \leqslant D}^{D} \frac{p(O^{\text{test}} | h^{(a,d)})}{\sum_{1 \leqslant a \leqslant A}^{A} p(O^{\text{test}} | h^{(a,d)})} \right) \quad (3)$$

其中, d 为当前视频的时间比例; a 为 d 时间比例下当前动作模型类别。

5　实验结果与分析

5.1　数据库

文中采用词袋模型与多时间比例高斯模型相结合的预测方法,原理上,这种方法可以对未知双人交互行为动作进行基本的预测。本次实验采用的数据库来自于 UT-interaction 数据库[14],该数据库广泛用于双人交互行为识别与预测算法研究中。实验在主频为 2.40 GHz,内存 2 GB,32 位 win7 操作系统下 Matlab 2014a 软件平台上完成。实验中采用留一交叉验证对数据库进行测试,聚类觉单词 $k = 800$, $d_1, d_2, \cdots, d_6$ 分别为 50%,60%,70%,80%,90%,100%。由于双人交互行为预测需要丰富的行为信息,所以数据库中除去“指”动作,如图 3 所示。

图 3　UT-interaction 数据库图例

5.2　概率预测结果

在本次实验中,利用 UT-interaction 数据库对词袋结合多时间比例高斯模型的方法进行了测试。以 handshake 测试视频为例,示例如图 4 所示。

由图 4 的实验结果可以看出,随着测试视频时间比例的逐渐增大,同类别动作相似概率值逐渐增大,进一步验证提出的方法可以实现双人交互行为预测与识别一体化。通过对大量不同时间比例的测试视频实验,得到的实验预测结果如表 1 所示。

	50%	60%	70%	80%	90%	100%
handshake	0.67	0.5	0.4	0.35	0.3	0.2
hug	0.16	0.2	0.1	0.1	0.2	0.2
kick	0	0	0.1	0.2	0	0.1
punch	0	0.2	0.2	0.15	0.3	0.2
push	0.17	0.1	0.2	0.2	0.2	0.3

(a) 测试视频时间比例 $d_1 = 50\%$

	50%	60%	70%	80%	90%	100%
handshake	0.5	0.71	0.45	0.42	0.3	0.2
hug	0.26	0.1	0.1	0.1	0.2	0.3
kick	0	0	0.1	0.1	0	0.2
punch	0	0	0.25	0.15	0.3	0
push	0.3	0.19	0.1	0.2	0.2	0.3

(b) 测试视频时间比例 $d_2 = 60\%$

	50%	60%	70%	80%	90%	100%
handshake	0.4	0.5	0.7	0.4	0.35	0.2
hug	0.1	0.1	0.1	0.1	0.1	0.3
kick	0	0.1	0.1	0.1	0	0.2
punch	0.2	0.1	0	0.2	0.35	0
push	0.3	0.2	0.1	0.2	0.2	0.3

(c) 测试视频时间比例 $d_3 = 70\%$

	50%	60%	70%	80%	90%	100%
handshake	0.4	0.35	0.5	0.8	0.6	0.3
hug	0	0.2	0.1	0.1	0.1	0.3
kick	0.1	0	0	0.1	0	0.2
punch	0.3	0.25	0.3	0	0.1	0
push	0.2	0.2	0.1	0	0.2	0.2

(d) 测试视频时间比例 $d_4 = 80\%$

	50%	60%	70%	80%	90%	100%
handshake	0.3	0.4	0.55	0.6	0.83	0.70
hug	0.1	0.2	0.1	0.1	0	0.1
kick	0	0.2	0	0.1	0	0.1
punch	0.3	0	0.25	0	0.17	0
push	0.3	0.2	0.1	0.2	0	0.1

(e) 测试视频时间比例 $d_5 = 90\%$

	50%	60%	70%	80%	90%	100%
handshake	0.2	0.3	0.4	0.5	0.7	0.86
hug	0.2	0.1	0.15	0.1	0.1	0
kick	0	0.2	0.1	0.2	0	0
punch	0.3	0.2	0.25	0	0.1	0
push	0.3	0.2	0.1	0.2	0.1	0.14

(f) 测试视频时间比例 $d_6 = 100\%$

图 4 handshake 测试视频在不同时间比例下相似概率混淆矩阵

表 1 不同时间比例测试视频最终预测结果

比例/%	最大概率叠加	比例/%	最大概率叠加
50	0.67	80	0.78
60	0.70	90	0.83
70	0.73	100	0.86

5.3 不同预测方法比较

表 2 给出了近年来在公开数据库中进行双人交互行为预测与识别结果,将文中提出的方法与其他作比较。结果表明,采用的方法预测与识别率均高于文献[1]和文献[5],虽然文献[4]和文献[7]预测与识别率较高,但是算法复杂度很高并且需要大量学习样本。而提出的词袋与多时间比例模型结合的方法并不需要建立复杂的预测模型,处理速度可达到 15 fps,且预测与识别准确率较高。

表 2 不同方法的识别结果 %

识别方法	Accracy w. half videos	Accuracy w. full videos
3D-SIFT+Integral Bow +Gaussian model	67.0	86.0
Integral Bow[1]	65.0	81.7
PST+PAF[5]	55.0	65.0
STISM[4]	80	91.7
CNN[7]	88.3	93.0

6 结束语

提出的基于兴趣点统计特征的方法,实现了对不同交互行为动作的预测。从特征描述与多时间比例模型概率预测的角度出发,采用词袋与高斯模型相结合的方法,很好地处理了对于未知时间长度的未知动作的预测和识别问题。该方法实现简单,满足实时性要求,具有较好的应用背景。但是,对于具有相似动作区间的动作预测存在一定的误差。因此,下一步的研究重点将放在预测模型优化上,以提高双人交互行为的预测率。

参考文献:

[1] RYOO M S. Human activity prediction: early recognition of ongoing activities from streaming videos[C]//International conference on computer vision. Barcelona, Spain: IEEE, 2011:1036-1043.

[2] 吴联世,夏利民,罗大庸. 人的交互行为识别与理解研究综述[J]. 计算机应用与软件,2011,28(11):60-63.

[3] HASSNER T. A critical review of action recognition benchmarks[C]//IEEE conference on computer vision and pattern recognition workshops. Portland, OR, USA: IEEE, 2013:245-250.

[4] YU G, YUAN J, LIU Z. Predicting human activities using spatio-temporal structure of interest points[C]//ACM international conference on multimedia. [s. l.]: ACM, 2012: 1049-1052.

[5] LI Kang, HU Jie, FU Yun. Modeling complex temporal composition of actionlets for activity prediction[C]//European conference on computer vision. Florence, Italy: Springer-Verlag, 2012:286-299.

[6] KE Qiuhong, BENNAMOUN M, AN Senjian, et al. Leveraging structural context models and ranking score fusion for human interaction prediction[J]. IEEE Transactions on Multimedia, 2018, 20(7):1712-1723.

[7] KE Qiuhong, BENNAMOUN M, AN Senjian, et al. Human interaction prediction using deep temporal features[C]//European conference on computer vision. [s. l.]: Springer, 2016:403-414.

[8] WEINLAND D, BOYER E, RONFARD R. Action recognition from arbitrary views using 3D exemplars[C]//IEEE international conference on computer vision. Rio de Janeiro, Brazil: IEEE, 2007:1-7.

[9] 姬晓飞,王昌汇,王扬扬. 分层结构的双人交互行为识别方法[J]. 智能系统学报,2015,10(6):893-900.

[10] SCHMID C, MOHR R. Local grayvalue invariants for image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(5):530-535.

[11] 王 博. 基于时空兴趣点的人体行为识别方法研究[D]. 南京:南京邮电大学,2014.

[12] 刘 懿,王 敏. 基于时空域 3D-SIFT 算子的动作识别[J]. 华中科技大学学报:自然科学版,2011,39:134-136.

[13] LI Feifei, PRONA P. A Bayesian hierarchical model for learning natural scene categories[C]//IEEE computer society conference on computer vision and pattern recognition. San Diego, CA, USA: IEEE, 2005:524-531.

[14] RYOO M S, AGGARWAL J K. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities[C]//IEEE international conference on computer vision. [s. l.]: IEEE, 2009:1593-1600.