文章编号：1672-6952(2019)06-0091-06 投稿网址：http://journal.lnpu.edu.cn

基于整体和个体分割融合的双人交互行为识别

魏 鹏[1]，曹江涛[1]，姬晓飞[2]
（1.辽宁石油化工大学，辽宁 抚顺 113001；2.沈阳航空航天大学，辽宁 沈阳 110136）

摘 要： 在人类交互行为识别领域，基于RGB视频的局部特征往往不能有效区分近似动作，将深度图像（Depth）与彩色图像（RGB）在识别过程中进行融合，提出一种融合Depth信息的整体和个体分割融合的双人交互行为识别算法。该算法首先分别对RGB和Depth视频进行兴趣点提取，在RGB视频上采用3DSIFT进行特征描述，在Depth视频上利用YOLO网络对左右两人兴趣点进行划分，并使用视觉共生矩阵对局部关联信息进行描述。最后使用最近邻分类器分别对RGB特征和Depth特征进行分类识别，进一步通过决策级融合两者识别结果，提高识别准确率。结果表明，结合深度视觉共生矩阵可以大大提高双人交互行为识别准确率，对于SBU Kinect interaction数据库中的动作可以达90%的正确识别率，验证了所提算法的有效性。

关键词： 交互行为； 局部特征； 视觉共生矩阵； 深度特征； YOLO； 决策级融合
中图分类号： TP301.6 **文献标志码：** A doi：10.3969/j.issn.1672-6952.2019.06.016

Double Human Interaction Recognition Based on Integration of Whole and Individual Segmentation

Wei Peng[1]，Cao Jiangtao[1]，Ji Xiaofei[2]
（1.Liaoning Shihua University，Fushun Liaoning 113001，China；2.Shenyang Aerospace University，Shenyang Liaoning 110136，China）

Abstract: In the field of human interaction recognition，local features based on RGB video often cannot effectively distinguish approximate actions. The Depth image information and the color image information are merged in the recognition process，and a two-person interactive behavior recognition algorithm that integrates the depth information and the individual segmentation fusion is proposed. The algorithm firstly extracts the points of interest for RGB and Depth video，then uses 3DSIFT to describe the features on RGB video. The YOLO network is introduced into divide the left and right points of interest on the Depth video，and the local co-occurrence matrix is used for local correlation information description. Finally，the nearest neighbor classifier is used to classify the RGB features and Depth features，and further the recognition results are obtained by the decision-level fusion，which improves the accuracy of recognition. The results show that the combination of depth visual co-occurrence matrix can greatly improve the recognition accuracy of double interaction behavior，and the correct recognition rate of 90% of the actions in SBU Kinect interaction database can verify the effectiveness of the proposed algorithm.

Keywords: Interaction behavior；Local features；Co-occurrence visual matrix；Depth features；YOLO；Decision level fusion

人类交互行为的认知和理解是计算机视觉领域的一个重要研究课题。相关技术广泛应用于智能监控、人机交互、基于内容的视频检索等领域[1]。基于RGB视频的双人交互行为识别已经取得了一定的进展[2-3]，但是由于其缺乏维度信息，导致复杂交互行为识别准确率不高。随着深度信息应用的普及，一些研究者已经将深度信息引入到双人交互行为识别的研究中，并取得了一定的进展[4-6]。

收稿日期：2018-12-05 **修回日期：**2018-12-20
基金项目：辽宁省自然科学基金项目(201602557)；辽宁省科技公益研究基金项目(2016002006)；辽宁省教育厅科学研究服务地方项目(L201708)；辽宁省教育厅科学研究青年项目(L201745)。
作者简介：魏鹏(1993-)，男，硕士研究生，从事基于深度信息的双人交互行为识别方向研究；E-mail:2513021587@qq.com。
通信联系人：曹江涛(1978-)，男，博士，教授，从事智能方法及其在工业控制信息处理上的应用，视频分析与处理等领域研究；E-mail:cigroup@126.com。

基于RGB和RGBD(RGB＋Depth)的双人交互行为识别主要分为两种处理方式，第一类将交互行为双方看作一个整体来识别[7]。T.H.Yu等[8]采用语义基元生成字典来描述视频中的局部时空体，并引用金字塔时空关系匹配的方法来识别交互动作。金壮壮等[9]使用两种不同的信息源和特征组合来进行编码，使用直方图投影的方式进行视觉词包构建，最后将最近邻分类器得到的概率矩阵进行加权融合，进而获得最终的识别结果。O.Oreifej等[10]采用HON4D(Histogram of Oriented 4D)特征来描述深度序列，该特征可以捕获4D空间中时间、深度和空间坐标的表面法线方向的分布，更好地获取深度序列中的关节形状运动线索。基于整体的交互行为识别方法是将交互双方作为一个整体，无需对双人交互视频的动作进行单独分割，该类方法的处理过程比较简单。因为基于整体的方法无法准确地表达交互动作中双方内在的相关性，所以该类方法的识别准确率略低，因此通常需要十分复杂的特征表示和匹配方法来保证交互行为识别的准确率。第二类基于个体分割的方法[11]，该方法将交互双方分解为单个原子动作或者动作模块，并将人与人之间的单个原子运动关系结合起来，进行交互行为的识别与理解。在检测人关节点位置信息的基础上，L.Meng等[12]提出语义空间关系描述方法用于表述个体动作姿态及不同动作执行人之间交互行为。该方法基于姿态估计算法，对交互行为双方人体交互中的内部空间信息和外部空间信息进行建模。基于共生原子动作匹配识别方法[13-15]对来自不同个体成对出现的共生原子动作进行模板表示，然后使用模板匹配进行识别。基于个体分割的互动行为识别方法需要跟踪和检测人体肢体部分，或者需要对原子动作进行识别[16-18]。因为现实中的交互行为场景比较复杂，而且受人体本身的遮挡、光照条件、噪声等因素的影响，所以无法保证准确地得到人体部分区域以及准确地识别单一的原子动作。

根据以上分析可知，基于整体或基于个体分割的处理框架，在处理双人交互行为识别中均具有一定的优势，也存在一定缺陷。本文综合利用两种框架的优势，在RGB视频上采用基于整体的方法对动作视频进行表示，在Depth视频上采取基于个体分割的动作视频表示，算法框图如图1所示。

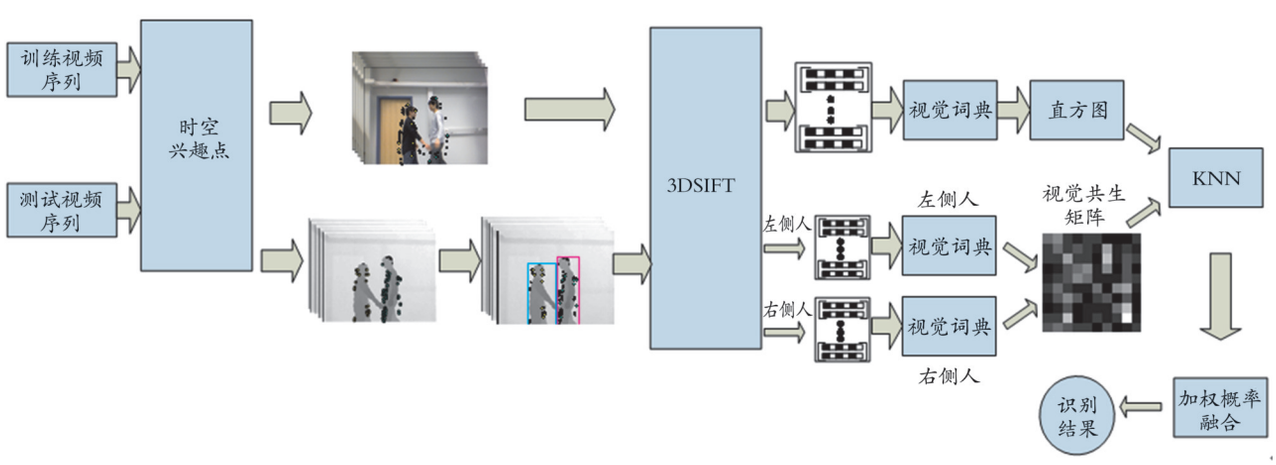

图1 算法结构框图

算法流程为：

(1)在RGB视频中将两个人看成一个整体，提取时空兴趣点。对RGB视频提取的兴趣点进行3DSIFT描述，利用词袋模型构建RGB动作视频的特征表示。

(2)在Depth视频上采用深度学习网络YOLOV3(You Only Look Once Version3)[19]对两个人进行分割，并在每个人的区域内分别提取时空兴趣点。采用视觉共生矩阵对Depth视频中属于每个人相关联的兴趣点进行特征描述。

(3)将RGB视频基于整体识别概率矩阵和Depth视频基于个体分割的识别概率矩阵进行融合，对双人交互进行识别。

本文在公开的SBU Kinect interaction数据库上进行算法测试，取得了比较好的实验结果，验证了本文方法的正确性。

1 个体分割

采用个体分割的框架的首要条件是能够对两个个体进行精确的分割。YOLOV3是一种基于深度学习的简单高效的目标检测算法，其最大的特点是运行速度非常快，YOLOV3把一幅图片划分成S×S的网格，根据网格坐标和内容来预测目标物体的位置。YOLOV3广泛应用在目标物体检测与行

人检测领域[20-22]，本文通过迁移学习利用预设YOLOV3网络模型，将Depth视频中两个人分别分割出来得到每个人的坐标信息，再进行下一步处理，检测结果如图2所示。

图2　网络个体分割结果

2　视频特征表示

基于时空兴趣点表征交互行为的算法被广泛应用于人类交互行为识别领域，本文使用时空兴趣点与视觉词包(Bag of Word, BOW)模型相互结合的方法对视频进行特征表示。

2.1　RGB视频特征表示

截止2019年，应用最广泛的是使用兴趣点检测的方法，此方法是由P.Dollár等[23]提出的基于Gabor滤波器和高斯滤波器相结合的计算函数响应值的方法。该方法分别使用空间和时间两个相互独立的线性滤波器进行组合，从视频序列中提取出大量详细的时空兴趣点，用于提取时序视频中人体变化的明显行为特征。对于提取的兴趣点进行3DSIFT描述，结合词袋模型对动作视频进行特征表示。通常采用k-means方法对训练数据的所有局部特征进行聚类形成词典，然后将一个动作视频的所有局部特征向词典投影，最终统计聚类词典中视觉单词在每类动作视频中出现的频率，进而形成动作视频的统计直方图。

2.2　Depth视频特征表示

Depth视频的基本特征和RGB视频一样使用了兴趣点提取和3DSIFT特征描述，不同之处在于受K.Slimani等[24]的启发，本文提出了利用Depth视频的视觉共生矩阵来表示人类交互行为。

在交互过程中，将交互双方看成不同的个体，提取两个人各自的兴趣点，基于兴趣点进行3DSIFT特征描述，使用视觉共生矩阵将视频中的两个人看成一个相关联的整体。

该算法首先利用YOLOV3网络进行检测并分割视频中的人，以便为每一个人创建3D-XYT时空体。然后，在两个3D-XYT时空体中提取时空兴趣点并对其进行3DSIFT特征描述。仍然采用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，最后构建视觉共生矩阵来统计左右动作执行人之间视觉单词共同出现的频数，形成Depth视频的特征表示。

3　实验验证与结果分析

3.1　数据库介绍

SBU双人交互数据库[22]是由Stony Brook Univer-sity于2012年运用微软Kinect传感器建立的，此数据库是一个不仅拥有深度图像和彩色图像，还拥有骨架图像(skeleton)的双人交互动作类的数据库。SBU数据库包含八种交互动作类别，如靠近(approaching)、远离(departing)、踢打(kicking)、击打(punching)、推(pushing)、拥抱(hugging)、握手(shaking hands)、交换(exchanging)，所有视频类都使用相同的室内环境背景进行录制，每一类动作视频都由不同的人来完成，增加数据集的多样性。整个数据库拥有260组双人交互动作视频，数据库被分割成21个集合，每个集合包含一个或两个类别的序列。在每个类别中，一个人正在行动，另一个人正在做出反应。本文将SBU双人交互数据库的RGB视频以动作类进行划分，每个动作的80%作为训练集，20%作为测试集，进行分析和实验。本文在训练集上对模型进行训练，将训练好的模型应用在测试集上进行测试，得到模型的性能等指标。

3.2　RGB和Depth视频测试结果

对于RGB视频而言，首先得到兴趣点，基于兴趣点和词袋模型结合对动作视频进行表示，然后使用KNN进行分类，正确识别率为78.33%，基于RGB视频识别的混淆矩阵如图3所示。

图3　基于RGB视频识别的混淆矩阵

从图3可以看出，混淆主要集中在靠近和远离两类动作上。通过分析可知，靠近和远离相互交互动作中交互体存在主动和被动运动模式，被动个体

基本保持不动，而主动个体的运动模式均为走这个基本动作，两者存在极大的相似性，因此识别的准确率较低。

在 Depth 视频上使用视觉共生矩阵来表示双人交互行为模型，并对其在 SBU Two-person interaction 动作类数据库进行实验，其准确率为 75.83%，基于 Depth 视频识别的混淆矩阵如图 4 所示。

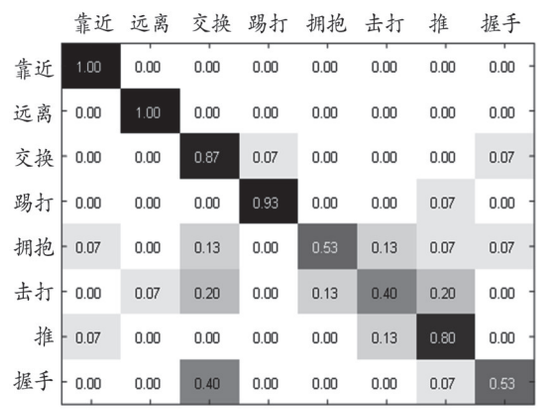

图 4　基于 Depth 视频识别的混淆矩阵

从图 4 可以看出，基于 Depth 图像的视觉共生矩阵，对于存在主动和被动运动模式的靠近和远离动作取得了很好的识别准确性，因此基于 Depth 图像的视觉共生矩阵可以作为 RGB 视频信息的有效补充，进一步提高识别的准确性。

3.3　两种特征融合结果

基于 RGB 视频的交互动作识别方法对于靠近和远离容易混淆的问题，Depth 视频识别方法能很好地解决此类问题，将 RGB 视频和 Depth 视频方法的优点相结合进行决策级融合。根据决策级融合所得到的结果，运用遍历的方法得到 Depth 视频和 RGB 视频最优权值分别为 0.36 和 0.64，最优权值测试结果如图 5 所示。

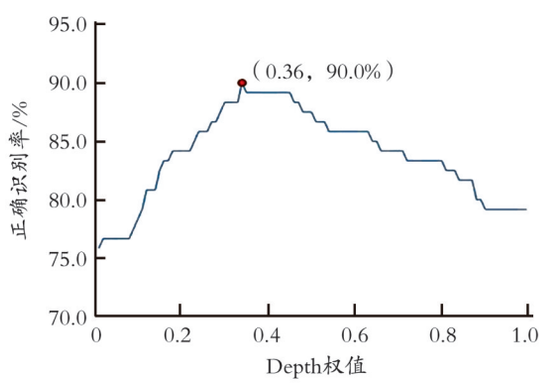

图 5　最优权值测试结果

最终的识别结果采用归一化后的混淆矩阵表示，决策级融合后的混淆矩阵如图 6 所示。从图 6 可以看出，RGB 视频和 Depth 视频概论矩阵经过加权决策及融合后的正确识别率提高到 90.00%，其中“靠近”、“远离”、“交换”和“踢打”识别效果得到明显提高。

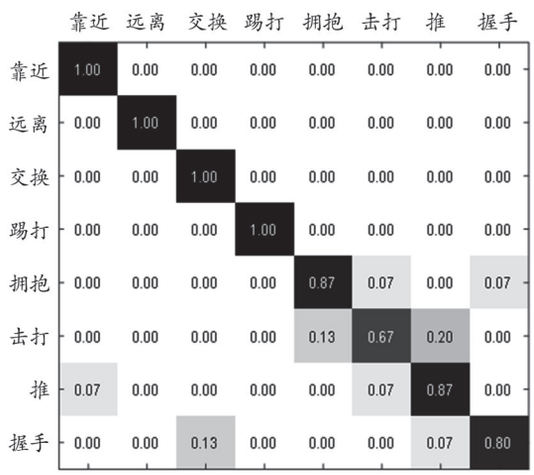

图 6　决策级融合后的混淆矩阵

4　实验结果

为了验证所提算法的有效性，将同样在 SBU Two-person interaction 动作类数据库上进行过算法测试的文献方法与本文方法进行比较与分析。其中，K.Yun 等[25]采用基于单原子关节点之间距离的物理空间特征来对双人交互行为进行建模，获得了很好的效果，通过实验证明此方法优于平面特征和速度等特征。Y.Ji 等[26]提出了一个交互式的身体部分模型，它连接不同参与者的交互肢体，以表示交互式身体部位的关系。对 8 个交互肢体进行时空运动特征运算，并将得到的时空运动特征进行词典构建，使用径向基（RBF）核的 SVM 进行识别。文献[27]通过提取交互双方关节点空间的距离、运动特征和单个人关节点运动特征来进行交互身体部位描述，然后将这些特征构建对比特征分布模型（CFDM）实现双人交互行为识别。文献[25]利用 LSTM 网络结构，并对每帧输入的不同关节点上加入注意机制，使每个关节在不同帧内拥有不同的权重，对动作时序拥有很好的认知和识别。

不同算法的识别结果如表 1 所示。从表 1 可以看出，本文方法将 RGB 视频和 Depth 视频两者相结合得到较高正确识别率，其识别结果优于文献[25—27]的识别率。文献[28]的正确识别率略高于本文结果，但是该方法引入了深度学习对人体骨架进行识别的方法，依赖于骨架信息的准确提取，且算法较复杂，需要大量的样本进行学习，存在对硬件要求高的问题。

表 1　不同算法的识别结果

算法	识别方法	识别率/%
本文算法	局部和全局信息融合	90.0
文献[25]算法	Raw skeletons	79.7
	Joint distance features	80.3
文献[26]算法	Raw position+Contrast Mining	79.4
	Joint features+Contrast Mining	86.9
文献[27]算法	Joint features+BOW	82.5
	Joint features +CM	84.6
	Joint features +CFDM	89.4
文献[28]算法	SA-LSTM	88.0
	STA-LSTM	91.5

5　结　论

根据 RGB 图像与 Depth 图像特性，提出一种融合 Depth 信息的整体和个体分割的双人交互行为识别算法，该算法结合 Depth 图像对 RGB 图像的互补特性，充分利用 Depth 图像的视觉共生矩阵特征，大大提高了识别的准确率。但是，对于存在较高的相似性动作来说还存在一定误识率。下一步的研究尝试在局部信息表示的基础上，融合光流的整体信息，进一步提高算法的识别准确性。

参　考　文　献

[1] Weinland D, Ronfard R, Boyer E. A survey of vision-based methods for action representation segmentation and recognition [J]. Computer Vision and Image Understanding, 2011, 115(2):224-241.

[2] 俞浩.基于时空兴趣点的人体行为识别研究[J].南京:南京师范大学, 2017.

[3] 王佩瑶, 曹江涛, 姬晓飞.基于改进时空兴趣点特征的双人交互行为识别[J].计算机应用, 2016, 36(10): 2875-2879.

[4] Ohn-Bar E, Trivedi M. Joint angles similarities and HOG2 for action recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland:IEEE,2013: 465-470.

[5] 金宏硕, 刘振宇.基于Kinect的手势图像识别研究[J].微处理机, 2018, 39(3):47-54.

[6] Uddin M Z, Torresen J, Jabid T. Human activity recognition using depth body part histograms and hidden markov models [C]//2016 International Conference on Innovations in Science, Engineering and Technology (ICISET). Dhaka:IEEE, 2016: 1-4.

[7] Li N, Cheng X, Guo H, et al. A hybrid method for human interaction recognition using spatio-temporal interest points [C]// 2014 22nd International Conference on Pattern Recognition. Stockholm:IEEE, 2014: 2513-2518.

[8] Yu T H, Kim T K, Cipolla R. Real time action recognition by spatio-temporal semantic and structural forests [C]// Proceedings of the 21st British Machine Vision Conference. BMVC.United Kingdom:[S.l.], 2010:1-12.

[9] 金壮壮,曹江涛,姬晓飞.多源信息融合的双人交互行为识别算法研究[J].计算机技术与发展,2018,28(10):32-36.

[10] Oreifej O, Liu Z. Hon4dHistogram of oriented 4d normals for activity recognition from depth sequences[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland:IEEE,2013: 716-723.

[11] Kong Y, Jia Y, Fu Y. Interactive phrases: Semantic descriptionsfor human interaction recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(9): 1775-1788.

[12] Meng L, Qing L, Yang P, et al. Activity recognition based on semantic spatial relation[C]//Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012).Tsukuba:IEEE, 2012: 609-612.

[13] 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. Kyoto:IEEE, 2015.

[14] Zhang X, Cui J, Tian L, et al. Local spatio-temporal feature based voting framework for complex human activity detection and localization[C]//The First Asian Conference on Pattern Recognition. Beijing:IEEE, 2011: 12-16.

[15] Yuan F, Yuan J. Middle-Level representation for human activities recognition: The role of spatio-temporal relationships [C]// European Conference on Computer Vision. Berlin Heidelberg:Springer, 2010:168-180.

[16] Shahroudy A, Ng T T, Gong Y, et al. Deep multimodal feature analysis for action recognition in RGB+D videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(5): 1045-1058.

[17] Ijjina E P, Chalavadi K M. Human action recognition in RGB+D videos using motion sequence information and deep learning[J]. Pattern Recognition, 2017,72: 504-516.

[18] Wang P, Li W, Gao Z, et al. Scene flow to action map: A new representation for RGB+D based action recognition with

convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii:IEEE,2017: 595-604.

[19] Redmon J, Farhadi A. Yolov3: An incremental improvement[EB/OL].(2018-04-08)[2018-12.01]. https://pjreddie.com/media/files/papers/YOLOv3.pdf.

[20] 高宗, 李少波, 陈济楠, 等. 基于 YOLO 网络的行人检测方法[J]. 计算机工程, 2018, 44(5): 215-219.

[21] 郭英强, 曹江涛. 多相机条件下的行人再识别方法研究[J]. 辽宁石油化工大学学报, 2018, 38(5): 77-81.

[22] 黄杰军, 呼吁, 周斌, 等. 基于改进YOLO的双模目标识别方法研究[J]. 计算机与数字工程, 2018, 46(4): 808-811.

[23] Dollár P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features[C]// Joint IEEE International Workshop on Visual Surveillance & Performance Evaluation of Tracking & Surveillance. Beijing:IEEE, 2006.

[24] Slimani K N E H, Benezeth Y, Souami F. Human interaction recognition based on the co-occurence of visual words[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.[S.l.]:IEEE, 2014: 455-460.

[25] Yun K, Honorio J, Chattopadhyay D, et al. Two-person interaction detection using body-pose features and multiple instance learning[C]//2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence:IEEE, 2012: 28-35.

[26] Ji Y, Ye G, Cheng H. Interactive body part contrast mining for human interaction recognition[C]//2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Chengdu:IEEE, 2014: 1-6.

[27] Ji Y, Cheng H, Zheng Y, et al. Learning contrastive feature distribution model for interaction recognition[J]. Journal of Visual Communication and Image Representation, 2015, 33: 340-349.

[28] Song S, Lan C, Xing J, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]//Thirty-first AAAI conference on artificial intelligence. New Orleans: AAAI Press,2017:4263-4270.

(编辑 陈 雷)