

# 基于多通道信息融合的双人交互动作识别算法

黄菲菲<sup>1</sup>, 曹江涛<sup>1</sup>, 姬晓飞<sup>2</sup>

(1. 辽宁石油化工大学 信息与控制工程学院, 辽宁 抚顺 113000;  
2. 沈阳航空航天大学 自动化学院, 辽宁 沈阳 110136)

**摘要:**基于视频的双人交互行为识别是计算机视觉领域中一个富有挑战性的研究课题。针对基于整体的双人交互动作识别方法的特征表示复杂度高及匹配方法难以确定的问题,文中提出了一种基于多通道信息融合的双人交互动作识别算法。该方法首先采用更符合人类视觉系统的HSI颜色空间模型,分别通过H、S、I三个通道来提取HOG特征并进行直方图统计表示,使用最近邻分类器分别获得三通道下的识别结果,然后对识别结果进行等比例融合得到待测视频的最终识别结果。该方法在UT-interaction上进行了测试,得到了81.7%的识别率,证明了该方法的有效性及其可行性。将其与相同数据库下的其他方法进行比较,结果表明该方法特征易于提取,计算效率高,避免了复杂的运算,具有一定的应用价值。

**关键词:**HOG特征;HSI颜色空间;等比例融合;行为识别

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2016)03-0058-05

doi:10.3969/j.issn.1673-629X.2016.03.014

## Two-human Interaction Recognition Algorithm Based on Multi-channels Information Fusion

HUANG Fei-fei<sup>1</sup>, CAO Jiang-tao<sup>1</sup>, JI Xiao-fei<sup>2</sup>

(1. School of Information and Control Engineering, Liaoning Shihua University, Fushun 113000, China;  
2. School of Automation, Shenyang Aerospace University, Shenyang 110136, China)

**Abstract:** Two-human interaction recognition based on video is a challenging research topic in computer vision. Aiming at the problem of high complexity for feature representation and matching method hard to determine for the two-human interaction recognition method, a two-human interaction recognition algorithm based on multi-channels information fusion is proposed in this paper. Firstly, HSI color space model which is more fit for the human visual system is used. Respectively by H, S, I three channels to extract the HOG feature for histogram statistics representation, the nearest neighbor classifier is used to require the identification results respectively under the three-channel, then the results is integrated with equal ratio to obtain the overall recognition rate. The proposed method is tested on UT-interaction which has achieved recognition ratio of 81.7%, proving the validity and feasibility of this method. Compared with other methods, the proposed method has higher calculation efficiency and recognition accuracy with increasing number of potential applications.

**Key words:** HOG features; HSI color space model; equal ratio fusion; interaction recognition

## 1 概述

双人交互动作在日常生活中非常普遍,如握手、拥抱等。基于视频的双人交互行为识别与理解在智能视频监控、人机交互、体育赛事检索、虚拟现实等领域有着广泛的应用前景。与单人动作相比,双人交互动作往往更加复杂,完成双人动作所涉及到的肢体动作种类更多,肢体之间的配合及排列方式也更加多样化。如何有效提取运动特征以及建立合理的交互模型是双

人交互行为识别与理解的两个重要研究内容。大量的国内外科研工作者及相关商家对此产生了浓厚的兴趣,尤其在美国、英国等国已经展开了大量相关项目的研究<sup>[1-4]</sup>。然而,由于光照条件的变化、背景的混乱干扰、运动目标的影子、物体与环境之间的遮挡等,使得双人交互行为识别仍然是一个富有挑战的课题<sup>[5]</sup>。

目前基于视频的双人交互行为识别方法大致分为两大类:基于个体分割的交互动作识别和基于整体的

收稿日期:2015-06-07

修回日期:2015-09-14

网络出版时间:2016-02-18

基金项目:国家自然科学基金资助项目(61103123, 61203021)

作者简介:黄菲菲(1990-),女,硕士研究生,研究方向为图像处理与识别;曹江涛,博士,教授,通讯作者,研究方向为智能方法及其在工业控制和视频信息处理上的应用;姬晓飞,博士,副教授,研究方向为视频处理及模式识别理论。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160218.1630.026.html>

交互动作识别。

基于个体分割的交互动作识别往往由动作执行个体的具有时间顺序的多个子动作在高层次结合而成,将交互动作分解为单个人的子动作并结合考虑人与人之间的运动关系进行识别与理解。Vahdat 等<sup>[6]</sup>提出基于样例的关键姿态图模型对双人交互行为进行建模与识别的方法。Yuan 等<sup>[7]</sup>将视频序列用一系列具有一致空间结构和一致运动的组件表示,通过对比这些成对组件的时空关系对双人交互行为进行识别。韩磊等<sup>[8-9]</sup>提出了一种基于时空单词的两人交互行为识别方法。把从行为视频中提取到的时空兴趣点划分给不同的人体,并在兴趣点样本空间聚类生成时空码本实现双人交互动作的识别。该类方法依赖于个体的正确分割,但在复杂的交互行为场景下,因遮挡等因素的影响,人体区域的正确分割很难保证。

基于整体的交互动作识别与理解方法通常将交互动作表示为包含所有动作执行人的一个整体时空描述

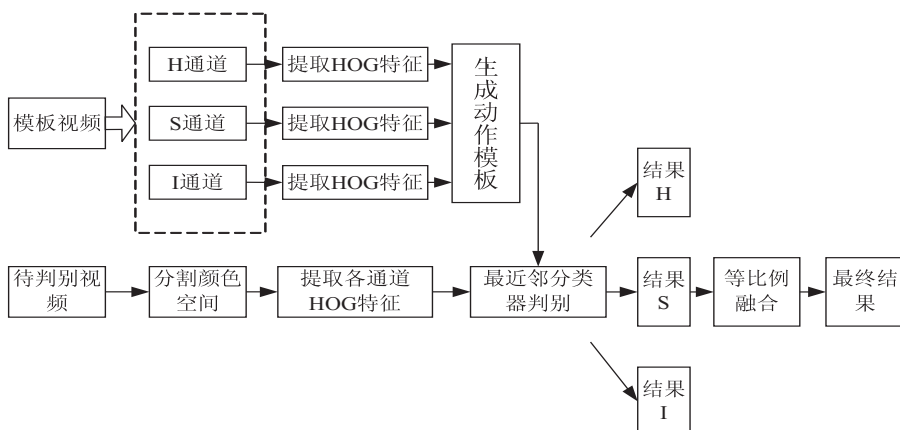


图1 算法结构框图

算法的实现过程分为训练和测试两个过程。在训练过程中,先将视频序列由RGB颜色空间转到HSI空间,然后在H、S、I三通道下对图像中的运动区域进行HOG特征的提取和表示,生成交互动作的模板。在测试过程中,也分别在H、S、I三通道下提取待测试视频每帧中运动区域的HOG特征,然后在三个通道下分别采用最近邻判别计算待测试视频帧与动作模板的相似性概率,最后等比例融合三个通道下的识别结果,得到待测视频的最终识别结果。

该方法仍然采用基于整体的双人交互动作识别方法,但特征提取和匹配算法简单,容易实现,且识别的准确率较高。

## 2 特征表示与提取

### 2.1 交互运动的检测与分割

文中利用帧差法进行交互运动的检测和分割,其原理就是在图像序列相邻的两帧图像间采用基于像素

形式,然后通过度量待识别交互动作时空特征表示与训练模板的匹配程度,对交互行为进行识别和理解<sup>[10]</sup>。Kong 等<sup>[11]</sup>采用语义基元森林(Semantic Texton Forest)生成词典对视频中的局部时空体进行描述,并引入金字塔时空关系匹配核对交互动作进行识别。Li 等<sup>[12]</sup>结合运动上下文(Motion Context)的全局特征和局部时空兴趣点的时空特征相关性(Spatio-Temporal Correlation),对双人交互行为进行描述,并分别提出了基于GA训练的随机森林方法及有效的时空匹配方法实现交互行为的识别与理解。该类方法无需对交互动作的特征进行动作个体的分割,处理思路简单;但是该类方法无法准确地表示交互动作中交互的内在属性,因此其识别的准确性有限,往往需要十分复杂的特征表示及匹配方法来保证识别的准确性。

根据以上分析,文中提出了一种基于多通道信息融合的双人交互动作识别算法,算法结构如图1所示。

的时间差分,并且通过阈值化去除静止的物体,提取图像中的运动区域,如图2所示。

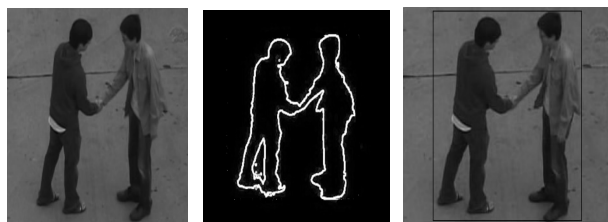


图2 运动分割效果

为了提高算法的识别准确性,文中以交互行为双方为主的两个感兴趣区域间的距离为0时,交互行为执行双方身体出现接触,双人交互行为进入执行阶段。在这一个阶段中,文中将双人交互行为整体所在区域作为感兴趣区域,进行分割提取操作,基于剪影特征的边界信息分割提取双人交互行为的感兴趣区域。而当交互行为开始及结束时,两个交互对象的位置是由远及近和由近及远的,在这两个过程中可以通过帧间差

分的方式,获得交互双方两个剪影的边界信息,分别获得以交互行为双方为主,冗余信息极少的感兴趣区域,将两个感兴趣区域的外边界合并,得到的就是未发生明显交互时的感兴趣区域。与此同时,在开始和结束阶段剔除交互双方距离较远时的图像。

## 2.2 RGB 颜色空间与 HSI 颜色空间的转换

人的视觉对亮度的敏感程度远强于对颜色浓淡的敏感程度,因此 HSI 色彩空间比 RGB 色彩空间更符合人的视觉特性。因此文中将采集到的视频序列转换到 HSI 色彩空间下进行后续处理,使其更加符合人眼的处理机制。RGB 模式到 HSI 模式的常用转换公式如下:

$$\begin{cases} H = \begin{cases} \theta & B > G \\ 2\pi - \theta & B \leq G \end{cases} \\ S = 1 - \frac{3[\min(R, G, B)]}{R + G + B} \\ I = \frac{R + G + B}{3} \end{cases} \quad (1)$$

其中,  $\theta =$

$$\arccos\left[\frac{2R - G - B}{2\sqrt{(R - G)^2 + (R - B)(G - B)}}\right]。$$

如式(1)所示,HSI 颜色空间对  $R, G, B$  三个分量重新编码。其中,色度分量在  $[0, 2\pi]$  范围内;饱和度分量和亮度分量在  $[0, 1]$  范围内。文中将  $R$  通道减半,  $G$  通道翻倍,  $B$  通道设为 0, 实现由 RGB 色彩空间向 HSI 空间的转换。其转化前后的对比图如图 3 所示。



(a) RGB 原图像



(b) 转换后的各通道图像

图 3 转换前后对比图

图中显示的 HSI 彩色模型可在彩色图像中从携带的彩色信息(色调和饱和度)里消去强度分量的影响。

## 2.3 HOG 特征提取

方向梯度直方图(Histogram of Oriented Gradient, HOG)特征最初是由 Dalal 等<sup>[13]</sup>提出的一种在计算机视觉和图像处理中用来进行物体检测的特征描述子。它通过计算和统计图像局部区域的梯度方向直方图来

构成特征。梯度提取操作不仅能够捕捉轮廓,人影和一些纹理信息,还能进一步弱化光照的影响。HOG 特征是一种不需要在相邻帧间进行处理的简单全局特征表示法,只需要在当前帧像素点间求取梯度的幅值和方向,并在不同方向上对像素点幅值大小进行直方图统计即可。因此文中采用 HOG 对每帧中的运动区域进行特征表示。

图像梯度的计算可以分解为图像横坐标和纵坐标方向的梯度,图 4 中像素点  $(x, y)$  的梯度为:

$$\begin{aligned} G_x(x, y) &= H(x + 1, y) - H(x - 1, y) \\ G_y(x, y) &= H(x, y + 1) - H(x, y - 1) \end{aligned} \quad (2)$$

式中,  $G_x(x, y)$ ,  $G_y(x, y)$  分别表示输入图像中像素点  $(x, y)$  处的水平方向梯度和垂直方向梯度。

像素点  $(x, y)$  处的梯度幅值和梯度方向为:

$$\begin{aligned} G(x, y) &= \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \\ \partial(x, y) &= \tan^{-1}\left(\frac{G_y(x, y)}{G_x(x, y)}\right) \end{aligned} \quad (3)$$

图像 HOG 特征的表示通常先将图像分成小的连通区域,然后采集细胞单元中各像素点的梯度或边缘的方向直方图,最后把这些直方图组合起来就可以构成特征描述器。文中将每一幅运动区域做  $4 * 4$  的分割,每个分割出的区块提取 12 维的 HOG 特征,那么最终特征的长度为  $16 * 12 = 192$  维,如图 4 所示。



图 4 HSI 空间下分通道提取 HOG 特征

由于 HOG 是在图像的局部方格单元上操作,所以它对图像的几何和光学的形变都能保持很好的不变性。其次,在粗的空域抽样,精细的方向抽样以及较强的局部光学归一化等条件下,只要行人大体上能够保持直立的姿势,是可以容许行人有一些细微的肢体动作,这些细微的动作可以被忽略而不影响检测的结果。

## 3 识别方法与等比例融合

### 3.1 识别方法

文中选用最简单的最近邻分类器<sup>[14]</sup>。具体算法如下:

(1) 找到测试序列每一帧的最近邻。设测试样本序列第  $t$  帧的特征向量为  $M'_0(t = 1, 2, \dots, T)$ , 训练样本所对应的第  $n$  帧特征向量为  $M'_n$ 。用欧几里德距离来测试  $M'_0, M'_n$  的相似性,与距离最小的训练样本帧就是测试样本序列第  $t$  帧的最近邻,如式(4)所示。



$$S_q = \min \| \mathbf{M}_Q^t - \mathbf{M}_T^n \| \quad (n = 1, 2, \dots, N) \quad (4)$$

(2)将测试帧对应的最近邻的训练帧所属动作的标号赋给当前的测试帧,这样测试序列的每一测试帧都将得到一个动作的标号。

(3)将测试序列每一帧的动作标号进行统计,测试序列类别对应为票数最多的标号对应的动作。例如,handshake\_002序列中有76帧,每帧都用最近邻动作标号标记,统计结果为[63,5,2,0,6,0],即有63帧被标记为1号动作,5帧被标记为2号动作,以此类推。票数最多为63,其对应的动作标号为1,则此序列将被识别为1号动作。

3.2 等比例融合

利用最近邻分类器得到测试视频序列分别在H、S、I三个通道上的分类投票直方图,然后将三个通道的分类投票直方图进行归一化处理,产生三个通道的分类概率直方图。最后识别结果通过三个通道的等比例融合而产生。

4 实验

4.1 数据库介绍

实验采用的数据库是UT视频数据库,该数据库是公开可下载的([http://cvrc. Ece. Utexas. edu/SDHA2010/Human\\_Interaction. html](http://cvrc. Ece. Utexas. edu/SDHA2010/Human_Interaction. html))。该数据库是由德州大学奥斯汀分校(University of Texas Austin)提供的。不同于对简单的周期性行为进行分类,该数据库包含了不同的时空条件下,连续视频流中的各种行为。UT交互动作数据库包含六大类人体交互行为的连续视频序列,分别是握手(hand shake)、拥抱(hug)、脚踢(kick)、指向(point)、猛击(punch)和推搡(push),每类动作下包含10个动作视频,一共60个视频,这60个视频也是已经标记好的,如图5所示。

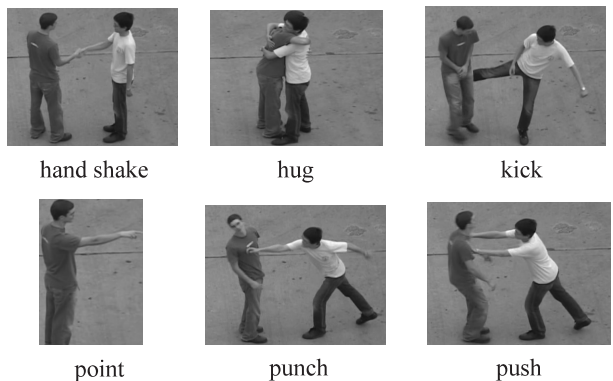


图5 UT数据库中六种动作

整个数据库由15个人在真实场景下两两完成,该数据库中的视频场景内大多包含杂乱的场景,相机的抖动,变化的光照等挑战因素。视频的分辨率是720\*480,刷新率20 fps,其中人的高度约为200像素。因

此在该视频上进行双人交互动作的检测与识别是十分具有挑战性的。

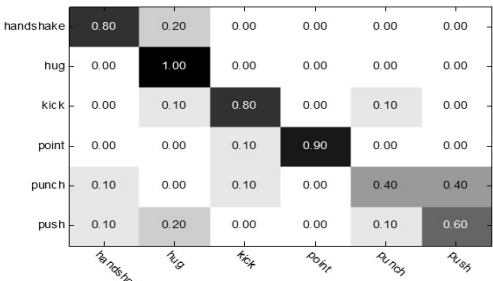
4.2 实验结果分析

文中采用留一法(leave one out)来验证算法的有效性,即每次实验选择数据库中的一个人的所有动作为测试样本集,而余下的作为训练样本集。然后循环,每个人的动作都将作为测试样本进行测试,并统计识别结果。在RGB图像上直接提取HOG特征以及在HSI颜色空间上分通道提取HOG特征识别结果如表1所示。

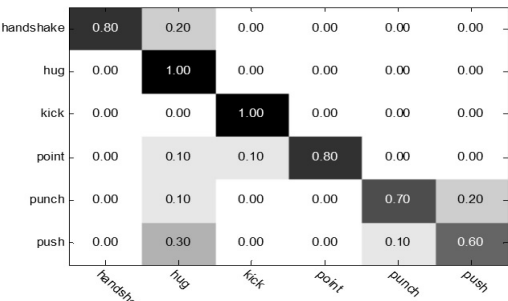
表1 识别结果	
各颜色空间	识别结果/%
RGB 图像	75
H 通道	55
S 通道	61.67
I 通道	71.67
三通道混合	81.7

由表1可以看出,在RGB图像上直接提取HOG特征结果为75%,在H、S、I三个通道上分别提取HOG特征的结果为55%、61.67%、71.67%,均低于在RGB图像上直接提取HOG特征的识别结果。然而将三个通道的识别结果进行等比例融合得到的最终识别率为81.7%,其识别的准确性有了大幅度的提高。

在RGB颜色空间上直接提取HOG特征和三个通道分别提取HOG特征并进行融合得到的识别混淆矩阵分别如图6(a)、(b)所示。



(a) RGB 空间下的混淆矩阵



(b)三通道融合后的识别混淆矩阵

图6 识别混淆矩阵

从图6(a)可以看出,识别结果中有较多的错误识

别,如 hand shake 被误判为 hug,punch 误判为 push 的机率较高;完全正确识别动作只有 hug。

从图(b)中可以看出,将三个通道混合后识别准确性有了大幅度的提升,完全正确识别动作有 hug 和 kick 两个动作,且对于 kick 和 punch 两类动作的识别结果有了显著提高。

将文中的识别方法与近期基于 UT 数据库的其他方法进行比较。实验结果如表 2 所示。

表 2 识别结果比较

文献	识别方法	识别结果/%
文献[12]	Plsa+SVM	79
文献[7]	Co-component + $\gamma^2$ SVM	78.2
文献[7]	GA search based random Forest ST correlation	85
文中方法	HSI 三通道 HOG 提取+最近邻判别+ 等比例融合	81.7

从表 2 可以看出,文中方法除了略差于文献[7]中的基于 GA 训练的随机森林方法的识别结果,而优于其他方法。但文献[7]的方法融合了运动上下文的全局特征和局部时空兴趣点的时空特征对双人交互行为进行描述,较文中方法复杂,计算复杂度高。因此文中方法特征易于提取,避免了复杂运算,并具有较高的识别准确性。

5 结束语

文中提出一种基于 HSI 颜色空间多通道的信息融合的双人交互动作识别方法。实验结果表明,用该方法提取 HOG 特征,在 UT-interaction 上得到了 81.7% 的识别率,证明了该方法的有效性及其可行性。此外,文中方法对差别较大的行为识别效果较好,对相似行为的识别效果还有待于进一步提高。

参考文献:

[1] Slimani K N E H, Benezeth Y, Souami F. Human interaction recognition based on the co-occurrence of visual words[C]//Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops. Columbus, Ohio, USA; IEEE, 2014; 461-466.

[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] Mukherjee S, Biswas S, Mukherjee D P. Recognizing interac-

tion between human performers using “key pose doublet” [C]//Proceedings of the ACM multimedia conference. Scottsdale, AZ, United States; ACM, 2011; 1329-1332.

[4] Ryoo M S. Human activity prediction: early recognition of ongoing activities from streaming videos[C]//Proceedings of the IEEE international conference on computer vision. Barcelona, Spain; IEEE, 2011; 1036-1043.

[5] Kantorov V, Laptev I. Efficient feature extraction, encoding and classification for action recognition[C]//Proceedings of the IEEE computer society conference on computer vision and pattern recognition. Columbus, OH, United States; IEEE, 2014; 2593-2600.

[6] Vahdat A, Gao Bo, Ranjbar M, et al. A discriminative key pose sequence model for recognizing human interactions[C]//Proceedings of the IEEE international conference on computer vision. Barcelona, Spain; IEEE, 2011; 1729-1736.

[7] Yuan Fei, Prinett V, Yuan Junsong. Middle-level representation for human activities recognition; the role of spatio-temporal relationships[C]//Proceedings of the 11th European conference on computer vision. Heraklion, Crete, Greece: [ s. n. ], 2010; 168-180.

[8] 韩磊, 李君峰, 贾云得. 基于时空单词的两人交互行为识别方法[J]. 计算机学报, 2010, 33(4): 776-784.

[9] 李君峰. 基于视觉的人与人交互动作分析[D]. 北京: 北京理工大学, 2010.

[10] Burghouts G J, Schutte K. Spatio-temporal layout of human actions for improved bag-of-words action detection[J]. Pattern Recognition Letters, 2013, 34(15): 1861-1869.

[11] Kong Yu, Jia Yunde, Fu Yun. Interactive phrases; semantic descriptions for human interaction recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(9): 1775-1788.

[12] Li Nijun, Cheng Xu, Guo Haiyan, et al. A hybrid method for human interaction recognition using spatio-temporal interest points[C]//Proceedings of the 22nd international conference on pattern recognition. Stockholm, Sweden: [ s. n. ], 2014; 2513-2518.

[13] Navneet D, Bill T. Histograms of oriented gradients for human detection[C]//Proc of IEEE computer society conference on computer vision and pattern recognition. San Diego, CA, USA; IEEE, 2005; 886-893.

[14] Wang Liang, Geng Xin, Leckie C, et al. Moving shape dynamics; a signal processing perspective[C]//Proceedings of the IEEE computer society conference on computer vision and pattern recognition. [ s. l. ]; IEEE Press, 2008; 1649-1656.