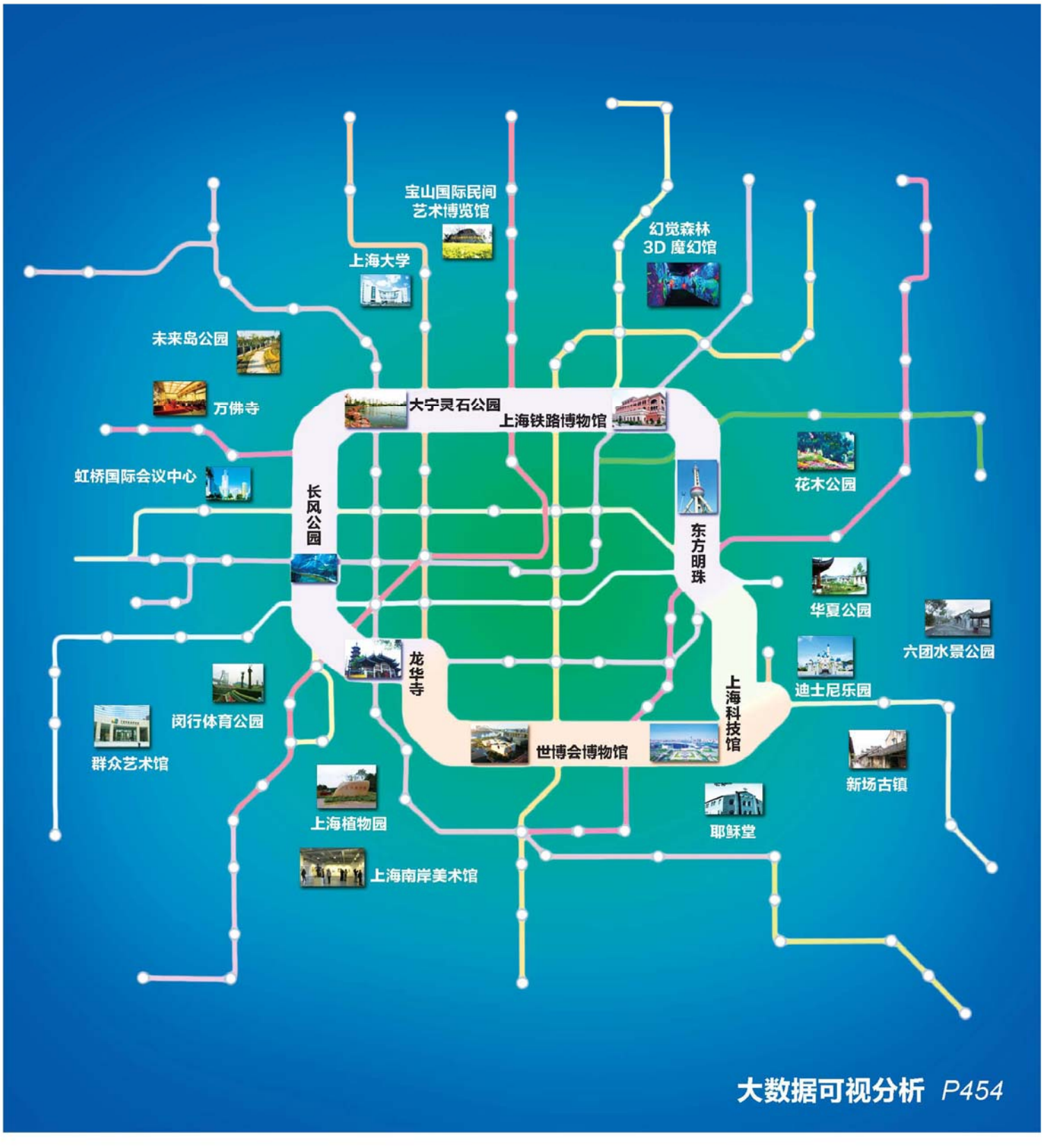
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2015
04
VOL.20

ISSN1006-8961
CN11-3758/TB



大数据可视分析 *P454*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 李小文
编辑出版 《中国图象图形学报》编辑出版委员会
邮政信箱 北京9718信箱
邮　　编 100101
电子信箱 jig@radi.ac.cn
电　　话 010-64807995
网　　址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

第20卷第4期（总第228期）
2015年4月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
P.O.Box 9718, Beijing, P.R.China
Zip code 100101
E-mail jig@radi.ac.cn
Telephone 010-64807995
Website www.cjig.cn

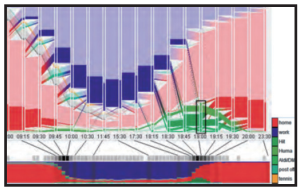
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

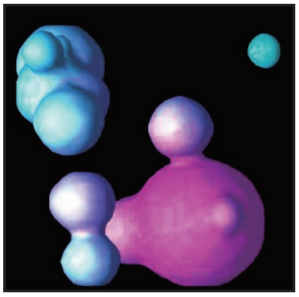
国外发行代号 M1406
国内邮发代号 82-831
国内定价 50.00元

2015年4月 第20卷 第4期（总第228期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

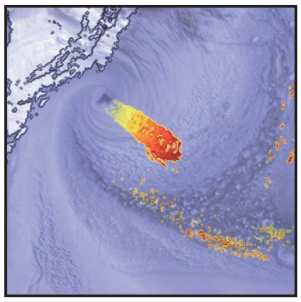
目次



城市道路交通数据可视分析综述(第0454页)



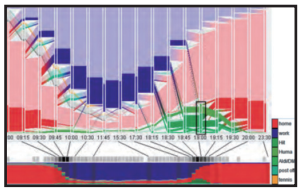
可视数据清洗综述(第0468页)



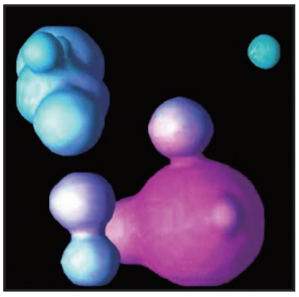
非对称高斯函数的时变体数据特征跟踪及可视化(第0576页)

Volume 20, Number 4
Published April 16, 2015

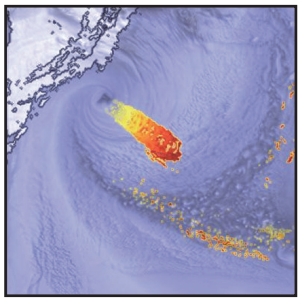
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Visual analytics of urban road transportation data: a survey (P0454)



Survey of visualization data cleaning(P0468)



Time-varying volume visualization and feature tracking based on asymmetric Gaussian function(P0576)

E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号：TP391.4　　**文献标识码**：A　　**文章编号**：1006-8961(2015)04-0520-07

论文引用格式：Wang T Q, Wang S J. Human action recognition using mid-level spatial-temporal features[J]. Journal of Image and Graphics, 2015, 20(4):0520-0526.[王泰青,王生进. 基于中层时空特征的人体行为识别[J]. 中国图象图形学报,2015,20(4):0520-0526.][DOI:10.11834/jig.20150408]

基于中层时空特征的人体行为识别

王泰青，王生进

清华大学电子工程系，智能技术与系统国家重点实验室，清华信息科学与技术国家实验室，北京　100084

摘　要：目的　人体行为识别是计算机视觉领域的一个重要研究课题,具有广泛的应用前景。针对局部时空特征和全局时空特征在行为识别问题中的局限性,提出一种新颖、有效的人体行为中层时空特征。**方法**　该特征通过描述视频中时空兴趣点邻域内局部特征的结构化分布,增强时空兴趣点的行为鉴别能力,同时,避免对人体行为的全局描述,能够灵活地适应行为的类内变化。使用互信息度量中层时空特征与行为类别的相关性,将视频识别为与之具有最大互信息的行为类别。**结果**　实验结果表明,本文的中层时空特征在行为识别准确率上优于基于局部时空特征的方法和其他方法,在KTH数据集和日常生活行为(ADL)数据集上分别达到了96.3%和98.0%的识别准确率。**结论**　本文的中层时空特征通过利用局部特征的时空分布信息,显著增强了行为鉴别能力,能够有效地识别多种复杂人体行为。

关键词：行为识别;时空兴趣点;中层时空特征;点互信息

Human action recognition using mid-level spatial-temporal features

Wang Taiqing, Wang Shengjin

School of Electronic Engineering, Tsinghua University, State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Beijing　100084, *China*

Abstract: **Objective** Human action recognition is an important research topic in the field of computer vision; this method has promising potential applications. On the basis of the limitations of local and global spatial-temporal features, a novel and effective middle-level spatial-temporal feature is proposed for action recognition. **Method** The proposed feature encodes the structural distribution of local features in the neighborhood of the spatial-temporal interest point (STIP), there by improving the discriminative power of STIP. This feature can model the flexible intra-action variations. Pointwise mutual information is introduced to measure the correlation between the mid-level feature and the action. The video clip is finally classified as the action category that has the greatest mutual information with the mid-level features. **Result** Experimental results validated the advantage of the proposed mid-level feature over the local-feature-based baseline methods and other published results. The mid-level feature achieved 96.3% and 98.0% recognition accuracies on the KTH and ADL (Activities of daily living) action datasets, respectively. **Conclusion** The proposed mid-level spatial-temporal feature enhances the discriminative power of actions by harnessing the spatial-temporal distribution of local spatial-temporal features. Consequently, this feature is capable of recognizing realistic human actions.

Key words: action recognition; spatial-temporal interest point; mid-level spatial-temporal feature; pointwise mutual information

收稿日期:2014-11-20;**修回日期**:2014-12-15

基金项目:国家高技术研究发展计划(863)基金项目(2012AA011004);国家科技支撑计划基金项目(2013BAK02B04)

第一作者简介:王泰青(1987—),男,清华大学电子工程系攻读信息与通信工程专业博士研究生,主要研究方向为人体行为分析。E-mail: wangtq09@mails.tsinghua.edu.cn

通信作者:王生进,教授,Email: wgsgj@tsinghua.edu.cn

Supported by: National High Technology Research and Development Program of China (2012AA011004); National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2013BAK02B04)

0　引　言

利用计算机自动识别视频中的人体行为是计算机视觉领域的一个重要研究课题，在智能视频监控、基于内容的视频检索、自然人机交互等领域具有广泛的应用前景[1-2]。例如，利用人体行为识别技术对地铁中的监控视频数据进行分析，识别斗殴、晕倒、丢弃包裹等行为，帮助警方及时处置。此外，人的表情识别也可以视为一种面部行为模式的识别[3]。

根据使用特征的不同，行为识别方法可分为以下两类[1-2]：

1）基于全局时空特征的方法。全局时空特征将人体行为表示为视频空间（即 XYT 空间）中 3 维物体，对 3 维物体的整体信息进行描述。文献[4]将人体行为的前景图像序列表示为 XYT 空间中的时空形状，使用泊松方程计算时空形状的全局特征，通过计算测试视频与训练视频中时空形状的相似度识别人体行为。文献[5]将人体行为视频作为模板，使用有向时空能量描述子作为模板的全局特征，通过模板匹配的方法进行行为识别。由于全局时空特征对人体行为的整体进行描述，适用于对刚性物体进行描述，不能够灵活描述人体关节变化导致的行为类内差异。

2）基于局部时空特征的方法。局部时空特征对视频中兴趣点邻域内的时空信息进行描述。文献[6]在视频中检测时空角点，使用梯度方向直方图和光流直方图对角点的邻域信息进行描述。文献[7]通过空域高斯滤波器和时域 Gabor 滤波器检测时空兴趣点，使用时空梯度方向直方图等特征对时空兴趣点进行描述。基于局部时空特征的行为识别方法一般通过聚类算法将局部时空特征量化为视觉单词，将行为视频表示为视觉单词分布直方图进行分类[6-7]。由于局部时空特征仅对视频中兴趣点的局部视频区域进行描述，缺乏足够的行为鉴别能力。

针对全局时空特征和局部时空特征在行为识别中的局限性，本文提出一种基于中层时空特征的人体行为识别方法。该特征通过对时空兴趣点较大邻域内的局部特征的结构化分布进行描述，增强时空兴趣点的行为鉴别能力；同时，由于避免了对人体行为的整体描述，该特征比全局时空特征能够更好地适应行为的类内变化。图 1 给出了本文方法的流程图。首先在视频中检测时空兴趣点（图 1（a）中圆圈所示），利用稀疏编码算法对时空兴趣点的局部特征进行语义编码（图 1（b）），然后计算时空兴趣点较大邻域内的语义编码结构化分布，生成中层时空特征（图 1（c）），最后通过构建随机森林学习中层时空特征与行为类别的互信息（图 1（d）），将视频识别为与之互信息最大的行为类别。

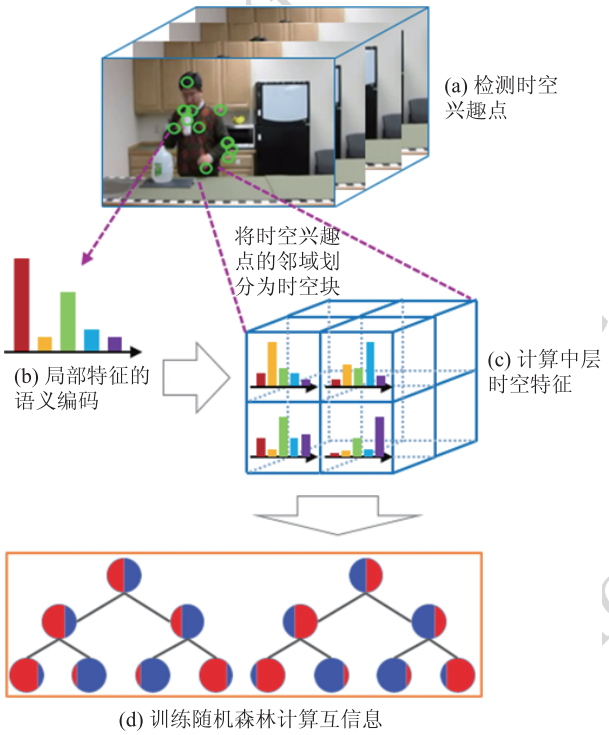


图 1　本文方法的流程图

Fig. 1　Framework of the proposed method

1　局部时空特征的语义编码

使用 Harris3D 检测子[6]在视频中检测时空兴趣点（图 1（a）），使用梯度方向直方图（HOG）和光流直方图（HOF）描述时空兴趣点的表观和运动信息[8]。由于 HOG-HOF 刻画了以时空兴趣点为中心的局部视频区域内的时空信息，将 HOG-HOF 特征称为局部时空特征（以下简称为局部特征）。由于 HOG-HOF 特征维度高，且对噪声敏感，因此利用稀疏编码算法[9]对局部特征在不同行为视频中的分

布特点进行抽取和描述,从而获得更为简洁和鲁棒的特征描述,即语义编码。首先利用稀疏编码算法为每类行为分别学习词典,然后将这些词典组合成联合词典,将局部特征分解为联合词典中单词的稀疏线性组合。

假设数据集有 C 类行为,$\boldsymbol{X}^c=[\boldsymbol{x}_1^c,\cdots,\boldsymbol{x}_{n_c}^c]\in\mathbf{R}^{m\times n_c}$表示第 c 类行为视频中的局部特征集合,$\boldsymbol{x}_i^c$ 表示局部特征,m 和 n_c 分别是局部特征的维度和数量。通过优化获得第 c 类行为的词典 $\boldsymbol{D}^c$,即

$$\boldsymbol{D}^c=\arg\min_{\boldsymbol{D},\boldsymbol{S}}\frac{1}{2}\|\boldsymbol{D}\boldsymbol{S}-\boldsymbol{X}^c\|_{l_2}^2+\lambda\|\boldsymbol{S}\|_{l_1}\quad(1)$$

式中,$\boldsymbol{D}\in\mathbf{R}^{m\times k_c}$是由 k_c 个视觉单词组成的词典,$\boldsymbol{S}\in\mathbf{R}_+^{k_c\times n_c}$是 $\boldsymbol{X}^c$ 中的局部特征在字典 $\boldsymbol{D}$ 上的稀疏编码系数,使用 $\boldsymbol{S}$ 的 1 范数 $\|\boldsymbol{S}\|_{l_1}$约束编码系数的稀疏性,λ 为稀疏正则项。

将所有行为类别的词典组合为一个联合词典 $\hat{\boldsymbol{D}}=[\boldsymbol{D}^1,\cdots,\boldsymbol{D}^C]$,将局部特征 $\boldsymbol{x}\in\mathbf{R}^{m\times 1}$表示为联合词典 $\hat{\boldsymbol{D}}$ 中单词的稀疏线性组合,即

$$\boldsymbol{s}^*=\arg\min_{\boldsymbol{s}}\frac{1}{2}\|\hat{\boldsymbol{D}}\boldsymbol{s}-\boldsymbol{x}\|_{l_2}^2+\lambda\|\boldsymbol{s}\|_{l_1}\quad(2)$$

式中,$\boldsymbol{s}\in\mathbf{R}_+^{k\times 1}$是 x 的稀疏编码稀疏,$k=\sum_{c=1}^{C}k_c$ 是 $\hat{\boldsymbol{D}}$ 中的单词数量。将局部特征 $\boldsymbol{x}$ 的编码系数 $\boldsymbol{s}^*$ 按照对应的行为类别的词典分为 C 组$\{\boldsymbol{s}^1,\cdots,\boldsymbol{s}^C\}$,将 $\boldsymbol{x}$ 的语义编码定义为

$$\boldsymbol{d}=[\|\boldsymbol{s}^1\|_{l_1},\cdots,\|\boldsymbol{s}^c\|_{l_1}]^{\mathrm{T}}\in\mathbf{R}^{C\times 1}\quad(3)$$

式中,$\|\boldsymbol{s}^i\|_{l_1}$表示 $\boldsymbol{s}^i$ 的 1 范数。图 1(b)形象地展示了一个局部特征的语义编码,横轴表示行为类别,纵轴表示局部特征在各个行为词典上的编码系数的 1 范数。如果局部特征 $\boldsymbol{x}$ 是第 c 类行为所独有的,则 $\boldsymbol{x}$ 可表示为词典 $\boldsymbol{D}^c$ 中单词的线性组合,即 $\|\boldsymbol{s}^c\|_{l_1}=1$,而对于任意的 $i\neq c$,$\|\boldsymbol{s}^i\|_{l_1}=0$,因此它的语义编码仅在第 c 维存在一个非零项;反之,如果 $\boldsymbol{x}$ 在多类行为视频中均有出现,则它的语义编码有多个非零项。因此,语义编码简洁地刻画了局部特征在不同行为类别中的分布特点。将位于视频中第 t 帧的(x,y)坐标处的时空兴趣点 $\boldsymbol{p}$ 记为 $\boldsymbol{p}=(\boldsymbol{v}_p,\boldsymbol{d}_p)$,其中 $\boldsymbol{v}_p=[x,y,t]^{\mathrm{T}}$ 是 p 的位置信息,$\boldsymbol{d}_p$ 是 $\boldsymbol{p}$ 的语义编码。

2 中层时空特征的计算方法

通过对时空兴趣点较大邻域内的局部特征语义编码的结构化分布进行描述,生成时空兴趣点的中层时空特征(以下简称为中层特征),增强其对人体行为的鉴别能力。为了在中层特征中引入局域特征的时空位置信息,将时空兴趣点的邻域离散化为若干个时空块,分别计算每个时空块内语义编码的分布信息,从而描述时空兴趣点周围不同时空区域中的局部特征语义编码的分布信息。

如图 1(c)所示,将时空兴趣点的邻域划分为 $N=n_x\times n_y\times n_t$ 个时空块,每个时空块的大小为 $w_x\times w_y\times w_t$。假设时空兴趣点 p 邻域内的第 i 个时空块的中心坐标为 $\boldsymbol{o}_i$。视频中的任意一个时空兴趣点 $q=(\boldsymbol{v}_q,\boldsymbol{d}_q)$对该时空块的隶属度函数 $m_i(\boldsymbol{q})$定义为基于两者距离的截断高斯函数

$$m_i(\boldsymbol{q})\propto\psi(\boldsymbol{v}_q,\boldsymbol{o}_i)\exp\left(-\frac{1}{2}(\boldsymbol{v}_q-\boldsymbol{o}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{v}_q-\boldsymbol{o}_i)\right)\quad(4)$$

式中,$\boldsymbol{\Sigma}=\mathrm{diag}(\sigma_x^2,\sigma_y^2,\sigma_t^2)$,$\sigma_x^2$,$\sigma_y^2$ 和 σ_t^2 是高斯核函数在 x、y 和 t 方向上的标准差。$\psi(\boldsymbol{v}_q,\boldsymbol{o}_i)$是决定时空块作用范围的窗函数

$$\psi(\boldsymbol{v},\boldsymbol{o}_i)=\begin{cases}1 & |\boldsymbol{v}-\boldsymbol{o}_i|\leqslant\eta[w_x,w_y,w_t]^{\mathrm{T}}\\ 0 & \text{其他}\end{cases}\quad(5)$$

式中,$\boldsymbol{a}\leqslant\boldsymbol{b}$ 表示向量 $\boldsymbol{a}$ 中的每个元素均小于向量 $\boldsymbol{b}$ 中的对应元素。这里取 $\eta=1$,从而使得相邻时空块的作用范围有一定程度的重叠,可以增强中层特征对位置和尺度变化的适应性。第 i 个时空块的描述子 $\boldsymbol{h}_p^i$ 表示为该时空块作用范围内局部特征语义编码的加权平均,即

$$\boldsymbol{h}_p^i=\sum_{\boldsymbol{q}}m_i(\boldsymbol{q})\boldsymbol{d}_q\quad(6)$$

时空兴趣点 p 邻域内的 N 个时空块的描述子构成 $\boldsymbol{p}$ 的中层时空特征的 N 个特征通道:$\boldsymbol{H}_p=(\boldsymbol{h}_p^1,\cdots,\boldsymbol{h}_p^N)$。可以看到,中层时空特征描述了时空兴趣点邻域内不同时空位置上的局部特征语义编码的分布,利用了多个局部特征的结构化信息,具有比局部特征更强的行为鉴别能力。

3 基于中层时空特征的行为识别方法

为了识别视频中人体行为的种类,使用互信息度量视频与行为类别的相关性,并将测试视频 $\boldsymbol{V}$ 识别为与之互信息最大的行为类别,即

$$c^*=\arg\max_{c\in\{1,\cdots,C\}}pmi(\boldsymbol{V},c)\quad(7)$$

式中，$pmi(\boldsymbol{V},c)=\log(P(\boldsymbol{V}|c)/P(\boldsymbol{V}))$ 为视频 $\boldsymbol{V}$ 与行为类别 c 的点互信息（本文简称为互信息）。这里将视频 $\boldsymbol{V}$ 表示为它所包含的中层时空特征的集合，即 $\boldsymbol{V}=\{\boldsymbol{H}_p\}$，其中 p 为 $\boldsymbol{V}$ 中的时空兴趣点。假设中层特征之间是统计独立的，则 $pmi(\boldsymbol{V},c)$ 可以表示为

$$pmi(\boldsymbol{V},c)=\log\frac{\prod_{\boldsymbol{H}_p\in\boldsymbol{V}}P(\boldsymbol{H}_p|c)}{\prod_{\boldsymbol{H}_p\in\boldsymbol{V}}P(\boldsymbol{H}_p)}=\sum_{\boldsymbol{H}_p\in\boldsymbol{V}}pmi(\boldsymbol{H}_p,C) \tag{8}$$

式中，$pmi(\boldsymbol{H}_p,c)=\log(P(\boldsymbol{H}_p|c)/P(\boldsymbol{H}_p))$ 为时空兴趣点 $\boldsymbol{p}$ 的中层特征 $\boldsymbol{H}_p$ 与第 c 类行为的互信息。因此，视频与行为类别的互信息可以表示为视频中的中层特征与行为类别的互信息之和。为了便于求解，将 $pmi(\boldsymbol{H}_p,c)$ 表示为

$$pmi(\boldsymbol{H}_p,c)=\log\frac{P(c|\boldsymbol{H}_p)}{P(c)} \tag{9}$$

式中，$P(c)$ 为行为类别 c 的先验概率，$P(c|\boldsymbol{H}_p)$ 为 $\boldsymbol{H}_p$ 关于行为类别 c 的后验概率，本节使用随机森林对后验概率 $P(c|\boldsymbol{H}_p)$ 进行建模。随机森林由多棵统计独立的决策树组成（如图 1（d）所示），下面介绍决策树的生成算法。

将所有训练视频中的中层特征记为集合 $\{(\boldsymbol{H}_p,y_p)\}$，其中 $\boldsymbol{H}_p=(\boldsymbol{h}_p^1,\cdots,\boldsymbol{h}_p^N)$ 为时空兴趣点 p 的中层特征，$y_p\in\{1,\cdots,C\}$ 为 $\boldsymbol{p}$ 所在的视频的行为类别。以构建行为类别 c 的决策树为例，将 $y_i=c$ 的中层特征视为正样本，将 $y_i\neq c$ 的中层特征视为负样本。在根节点处设计二值函数

$$t_{\rho,r,\delta}(\boldsymbol{H}_p)=\begin{cases}1 & \boldsymbol{H}_p^{\rho}(r)>\delta\\ 0 & \text{其他}\end{cases} \tag{10}$$

式中，$\rho\in\{1,\cdots,N\}$ 表示中层特征 $\boldsymbol{H}_p$ 的第 ρ 个特征通道 $\boldsymbol{h}_p^{\rho}$，$\boldsymbol{H}_p^{\rho}(r)$ 表示 $\boldsymbol{h}_p^{\rho}$ 的第 r 个元素，δ 为偏置。将满足 $t_{\rho,r,\delta}(\boldsymbol{H}_p)=1$ 的样本加入集合 $\boldsymbol{X}_{\mathrm{L}}$，将满足 $t_{\rho,r,\delta}(\boldsymbol{H}_p)=0$ 的样本加入集合 $\boldsymbol{X}_{\mathrm{R}}$。新建根节点的左子节点 n_{L} 和右子节点 n_{R}，将 $\boldsymbol{X}_{\mathrm{L}}$ 和 $\boldsymbol{X}_{\mathrm{R}}$ 分别作为 n_{L} 和 n_{R} 的样本集。在每个节点处随机生成 γ 个 (ρ,r,δ)，选择使得正负样本分离度最大[10]的 (ρ^*,r^*,δ^*) 作为该节点的二值函数的参数。重复上述过程，直到节点的样本数量小于某个阈值。图 1（d）中给出了决策树的示意图，每个节点中红色和蓝色区域面积的比例表示该节点的样本集合中正负样本的比例。

假设行为类别 c 的随机森林 $\boldsymbol{F}_c=\{T_i,i=1,\cdots,K\}$ 由 K 个决策树组成。为了计算中层特征 $\boldsymbol{H}_p$ 与行为类别 c 的互信息，将 $\boldsymbol{H}_p$ 输入到 $\boldsymbol{F}_c$ 的所有决策树中。假设 $\boldsymbol{H}_p$ 在决策树 T_i 中到达的叶子节点有 N_i^+ 个正样本和 N_i^- 个负样本，则 $\boldsymbol{H}_p$ 的后验概率 $P(c|\boldsymbol{H}_p)$ 表示为

$$P(c|\boldsymbol{H}_p)=\frac{1}{K}\sum_{i=1}^{K}\frac{N_i^+}{N_i^++N_i^-} \tag{11}$$

从而可以根据式（8）（9）计算视频 $\boldsymbol{V}$ 与行为类别 c 的互信息 $pmi(\boldsymbol{V},c)$，利用式（7）识别 $\boldsymbol{V}$ 中的人体行为。

4　实验结果

在 KTH 行为数据集[8]和日常生活行为数据集（ADL）[11]上对本文方法进行了测试。为了深入分析本文方法的不同模块对实验结果的影响，设计了两个对照方法：

1）使用 HOG-HOF 特征作为时空兴趣点的描述子，利用随机森林学习 HOG-HOF 特征与行为类别的互信息进行行为识别；

2）使用语义编码作为时空兴趣点的描述子，利用随机森林学习语义编码与行为类别的互信息进行行为识别。

4.1　KTH 数据集上的实验结果

KTH 数据集[8]是行为识别领域识别最为广泛的行为数据集，由 6 种行为组成，分别是 box（拳击）、handclap（拍手）、handwave（挥手）、jog（慢跑）、run 和 walk，如图 2 所示。每种行为的视频数据采集自 25 个人在 4 种不同场景中的行为。在语义编码中每类行为的词典由 250 个单词组成。中层特征参数为 $(n_x,n_y,n_t)=(3,3,3)$ 和 $(w_x,w_y,w_t)=(15,15,8)$。每类行为的随机森林由 20 棵决策树组成。采用文献[8]的实验方法，将 25 个人中的 16 个人的行为视频作为训练集，将其他人的视频作为测试集。

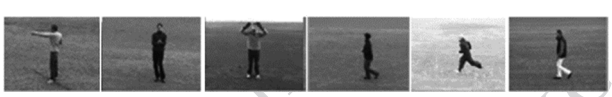


图 2　KTH 数据集样本示例

Fig. 2　Samples from the KTH dataset

本文的中层时空特征在 KTH 数据集上取得了 96.3% 的识别率，基于 HOG-HOF 特征的对照方法 1）和基于语义编码的对照方法 2）分别获得了 91.2% 和 94.9% 的识别率。这表明，语义编码通过描述 HOG-HOF 特征在不同行为类别中的分布信息获得了比底层的 HOG-HOF 特征更强的行为鉴别力。同时，通过描述时空兴趣点较大邻域内语义编码的结构化分布，本文的中层时空特征可以进一步增强时空兴趣点的行为鉴别力。

表 1 对比了本文方法与其他方法[8,12-15]在 KTH 数据集上的行为识别率。可以看到，本文方法优于大部分的其他方法。需要指出的是，文献[15]从 KTH 以及其他数据集中人工选择 205 个行为模板取得了 98.2% 的识别率，比本文方法使用了更多的训练样本。同时，由于依赖于人工选择，该方法难以用于较大规模的行为识别问题，而且实验结果具有一定的主观性。图 3 给出了本文方法在 KTH 数据集上的识别混淆矩阵，可以看出本文方法能够准确识别其中 4 种行为，而在 jog（慢跑）和 run 两种行为上产生一定的混淆，这是由于这两种行为本身具有较大的相似性。

表 1　KTH 数据集上的行为识别结果

Table 1　Action recognition results on the KTH dataset

行为识别方法	平均识别精度/%
文献[8]	71.7
文献[12]	88.3
文献[13]	93.8
文献[14]	94.4
文献[15]	98.2
对照方法 1）（HOG-HOF）	91.2
对照方法 2）（语义编码）	94.9
本文	96.3

4.2　ADL 数据集上的实验结果

ADL 数据集[11]由 10 种日常行为的视频组成，包括 answerPhone、dialPhone、lookupPhonebook、chopBanana、peelBanana、eatBanana、eatSnack、drinkWater、useSilverware 和 writeOnBoard。每种行为由 5 个人重复进行 3 次，共采集 150 段视频。该数据集最初用于研究老年人认知辅助系统。图 4 给出了该数据集的 loopupInPhonebook、chopBanana 和 drinkWater 样本示例。在语义编码中每类行为的词典由 375 个单词组成，其他参数与 KTH 数据集相同。本节采用留一法统计行为识别准确率[11]，即依次选择一个人的视频作为测试样本，其他人的视频作为训练样本，重复进行 5 次实验，统计平均识别率。

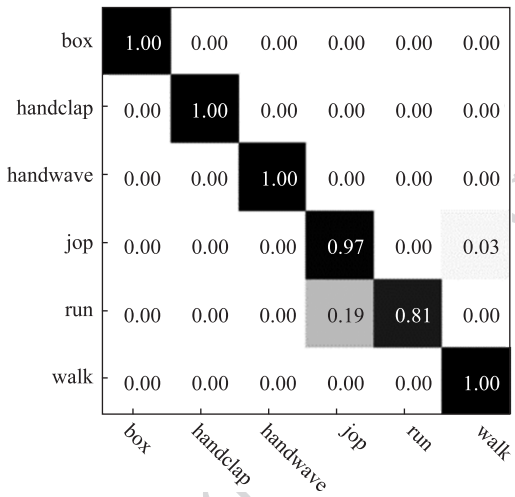


图 3　本文方法在 KTH 数据集上的分类混淆矩阵

Fig. 3　Action confusion matrix on the KTH dataset

图 4　ADL 数据集样本示例

Fig. 4　Samples from the ADL dataset

表 2 给出了本文方法与其他方法[11,13,16-18]在 ADL 数据集上的行为识别准确率。本文方法取得了 98.0% 的识别率，优于文献方法在该数据集上的最好识别率 96.0%[13]。图 5 给出了本文方法的识别混淆矩阵，显示了本文方法的有效性。文献[13]使用聚类算法将时空兴趣点的 HOG-HOF 特征量化为视觉单词，使用 4 种网格对其邻域内的视觉单词分布进行描述，利用多核学习算法学习 4 种网格的最优组合。本文使用的 3×3×3 的网格可以组成 $2^{27}-1\approx10^8$ 种形状，利用随机森林从这些形状中选择最优的网格组合；同时，使用语义编码对 HOG-HOF 特征进行“软量化”，避免了聚类算法的“硬量化”误差。

表 2　ADL 数据集上的行为识别结果

Table 2　Action recognition results on the ADL dataset

行为识别方法	平均识别精度/%
文献[16]	70.0
文献[17]	82.7
文献[11]	89.0
文献[18]	91.0
文献[13]	96.0
对照方法 1)(HOG-HOF)	86.0
对照方法 2)(语义编码)	90.7
本文	98.0

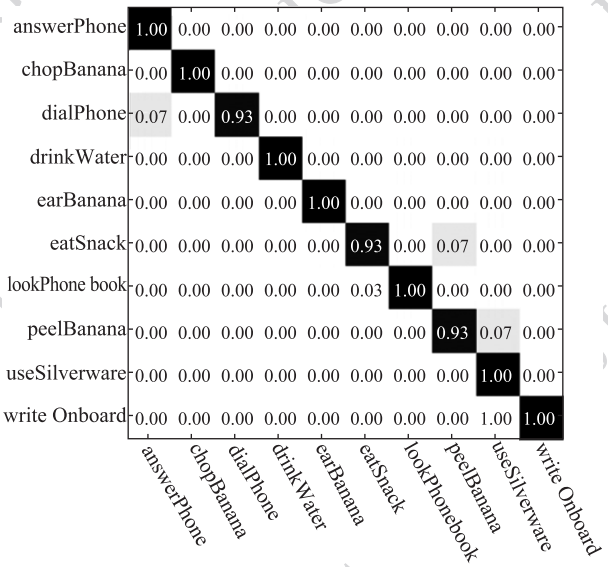


图 5　本文方法在 ADL 数据集上的分类混淆矩阵

Fig. 5　Action confusion matrix on the ADL dataset

基于 HOG-HOF 的对照方法 1)和基于语义编码的对照方法 2)分别获得了 86.0% 和 90.7% 的识别率。在 KTH 数据集上，本文方法的识别率比对照方法 1)和对照方法 2)分别提高了 5.1 和 1.4 个百分点，而在人体行为更为复杂的 ADL 数据集上，本文方法的识别率比两者分别提高 12.0 和 7.2 个百分点。这表明，中层时空特征通过描述局部特征的结构化分布信息，能够更好地识别复杂人体行为，而 HOG-HOF 和语义编码仅对时空兴趣点自身的时空信息进行描述，鉴别信息有限，缺乏对复杂人体行为的鉴别能力。

5　结　论

针对局部时空特征和全局时空特征在行为识别问题中的局限性，提出了一种新颖的中层时空特征用于行为识别。该特征通过描述时空兴趣点较大邻域内局部特征语义编码的结构化分布，增强了时空兴趣点的行为鉴别能力，并利用随机森林挖掘该特征的行为鉴别信息。实验结果表明，本文的中层时空特征在行为识别准确率上优于基于局部时空特征的对照方法以及主要文献方法。在人体行为更为复杂的 ADL 数据集上，中层时空特征在识别准确率上的优势更为显著，这说明研究中层时空特征对于识别复杂人体行为具有重要意义。目前，本文方法是判断一段视频中是否发生某种人体行为，未能给出行为发生的起止时间和在图像中的空间位置，因此，下一步将研究如何利用中层时空特征对人体行为的时空位置进行分析。

参考文献(References)

[1] Zhang Y J. Understanding spatial-temporal behaviors [J]. Journal of Image and Graphics, 2013, 18(2): 141-151. [章毓晋. 时空行为理解[J]. 中国图象图形学报, 2013, 18(2): 141-151.][DOI:10.11834/jig.20130203.]

[2] Gu J X, Ding X Q, Wang S J. A survey of activity analysis algorithms [J]. Journal of Image and Graphics, 2009, 14(3):377-387. [谷军霞,丁晓青,王生进. 行为分析算法综述[J]. 中国图象图形学报, 2009, 14(3): 377-387.][DOI:10.11834/jig.20090301.]

[3] Li Y, Wang S, Ding X. Eye/eyes tracking based on a unified deformable template and particle filtering [J]. Pattern Recognition Letters, 2010, 31(11): 1377-1387.

[4] Gorelick L, Blank M, Shechtman E, et al. Actions as space-time shapes [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2247-2253.

[5] Derpanis K G, Sizintsev M, Cannons K, et al. Efficient action spotting based on a spacetime oriented structure representation [C]// Proceedings of CVPR 2010. San Francisco, CA: IEEE Press, 2010: 1990-1997.

[6] Laptev I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64(2-3): 107-123.

[7] Dollár P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features [C]// Proceedings of VS-PETS 2005. Beijing, China: IEEE Press, 2005: 65-72.

[8] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local svm approach [C]// Proceedings of ICPR 2004. Singapore: IEEE Press, 2004: 3: 32-36.

[9] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227.

[10] Duda R O, Hart P E, Stork D G. Pattern Classification [M]. 2nd ed. New York: Wiley-Interscience, 2001: 398.

[11] Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints [C]// Proceedings of ICCV 2009. Kyoto, Japan: IEEE Press, 2009: 104-111.

[12] Rapantzikos K, Avrithis Y, Kollias S. Dense saliency based spatiotemporal feature points for action recognition [C]// Proceedings of CVPR 2009. Miami, USA: IEEE Press, 2009: 1454-1461.

[13] Wang J, Chen Z, Wu Y. Action recognition with multiscale spatio-temporal contexts [C]// Proceedings of CVPR 2011. Providence, USA: IEEE Press, 2011: 3185-3192.

[14] Wang T, Wang S, Ding X. Detecting human action as the spatio-temporal tube of maximum mutual information [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(2): 277-290.

[15] Sadanand S, Corso J J. Action bank: a high-level representation of activity in video [C]// Proceedings of CVPR 2012. Providence, USA: IEEE Press, 2012: 1234-1241.

[16] Matikainen P, Hebert M, Sukthankar R. Representing pairwise spatial and temporal relations for action recognition [C]// Proceedings of ECCV 2010. Berlin: Springer Press, 2010: 508-521.

[17] Raptis M, Soatto S. Tracklet descriptors for action modeling and video analysis [C]// Proceedings of ECCV 2010. Berlin: Springer Press, 2010: 577-590.

[18] Yuan F, Xia G S, Sahbi H, et al. Mid-level features and spatio-temporal context for activity recognition [J]. Pattern Recognition, 2012, 45(12): 4182-4191.