ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.10, October 2009, pp.2679−2691
doi: 10.3724/SP.J.1001.2009.03499

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于动作图的视角无关动作识别*

杨跃东[1+]，郝爱民[1]，褚庆军[2]，赵沁平[1]，王莉莉[1]

[1](北京航空航天大学 虚拟现实技术与系统国家重点实验室,北京 100191)

[2](国家教育部考试中心,北京 100084)

View-Invariant Action Recognition Based on Action Graphs

YANG Yue-Dong[1+], HAO Ai-Min[1], CHU Qing-Jun[2], ZHAO Qin-Ping[1], WANG Li-Li[1]

[1](State Key Laboratory of Virtual Reality Technology, Beihang University, Beijing 100191, China)

[2](National Education Examinations Authority Ministry of Education of China, Beijing 100084, China)

+ Corresponding author: E-mail: yyuedong@gmail.com

Yang YD, Hao AM, Chu QJ, Zhao QP, Wang LL. View-Invariant action recognition based on action graphs. *Journal of Software*, 2009,20(10):2679−2691. http://www.jos.org.cn/1000-9825/3499.htm

Abstract: This paper proposes a weighted codebook vector representation and an action graph model for view-invariant human action recognition. A video is represented as a weighted codebook vector combining dynamic interest points and static shapes. This combined representation has strong noise robusticity and high classification performance on static actions. Several 3D key poses are extracted from the motion capture data or points cloud data, and a set of primitive motion segments are generated. A directed graph called Essential Graph is built of these segments according to self-link, forward-link and back-link. Action Graph is generated from the essential graph projected from a wide range of viewpoints. This paper uses Naïve Bayes to train a statistical model for each node. Given an unlabeled video, Viterbi algorithm is used for computing the match score between the video and the action graph. The video is then labeled based on the maximum score. Finally, the algorithm is tested on the IXMAS dataset, and the CMU motion capture library. The experimental results demonstrate that this algorithm can recognize the view-invariant actions and achieve high recognition rates.

Key words: action recognition; view-invariant; action graph; interest point; Naïve Bayes

摘 要: 针对视角无关的动作识别,提出加权字典向量描述方法和动作图识别模型.将视频中的局部兴趣点特征和全局形状描述有机结合,形成加权字典向量的描述方法,该方法既具有兴趣点抗噪声强的优点,又可克服兴趣点无法识别静态动作的缺点.根据运动捕获、点云等三维运动数据构建能量曲线,提取关键姿势,生成基本运动单元,并通过自连接、向前连接和向后连接 3 种连接方式构成有向图,称为本质图.本质图向各个方向投影,根据节点近邻规则建立的有向图称为动作图.通过 Naïve Bayes 训练动作图模型,采用 Viterbi 算法计算视频与动作图的匹配度,根据最大匹配度标定视频序列.动作图具有多角度投影和投影平滑过渡等特点,因此可识别任意角度、任意运动方向的视频序列.实验结果表明,该算法具有较好的识别效果,可识别单目视频、多目视频和多动作视频.

* Supported by the National High-Tech Research and Development Plan of China under Grant Nos.2006AA01Z333, 2007AA01Z337 (国家高技术研究发展计划(863)); the China High-Tech Olympics Project under Grant No.Z0005191041211 (中国科技奥运专项)

Received 2008-07-31; Accepted 2008-10-17; Published online 2009-06-09

关键词: 动作识别;角度无关;动作图;兴趣点;Naïve Bayes
中图法分类号: TP391 文献标识码: A

在计算机视觉领域,人体动作识别是一个非常活跃的研究课题,其研究成果广泛应用于视频监控、视频检索和人机交互等方面,文献[1−4]给出了较为全面的总结和分析.针对视角无关的动作识别,本文提出基于兴趣点和形状的加权字典向量描述方法和基于三维动作库的动作图(action graph)识别模型.视频描述和识别模型是动作识别的两个研究重点.

视频中运动信息可以通过光流、全局时空特征和局部兴趣点等来描述.Ahmad 等人采用局部-全局光流(combined local-global optic flow)描述人体运动[5,6].Efros 等人基于光流法提出一种时空运动描述子[7].此外,运动历史图(motion-history images)[8]、时空体(spatio-temporal volumes)[9]、时空形状(space-time shape)[10]等全局时空信息也被用于描述人体运动.

近年来,局部兴趣点被广泛应用于运动识别.相对于全局运动描述,局部兴趣点具有较好的旋转、平移和缩放等不变性,可有效降低复杂背景、人体形状和相机等带来的影响.Laptev 在图像 Harris 角点检测的基础上,提出视频 Harris 角点检测算法[11].针对 3D Harris 兴趣点数量少的问题,Oikonomopoulos[12],Dollár[13]等人进行了有效的扩展.基于局部兴趣点描述,SVM(support vector machines)[14],NNC(nearest neighbour classifier)[13,15],pLSA (probabilistic latent semantic analysis)[16−18]和 LDA(latent dirichlet allocation)[18]等分类方法都曾被使用,并取得了较好的识别效果.

对于一些静态动作,例如“站”和“坐”,由于其运动特征不明显,如果仅仅采用兴趣点表示人体运动,则识别效果较差.人体形状/轮廓与视频中运动信息无关,可以有效地弥补这个缺陷.本文在兴趣点的基础上,结合人体形状,采用加权字典的方法,构建可描述动态和静态动作的表示形式.

目前,角度无关性动作识别模型可粗略地分为基于跟踪、基于不变描述和基于轮廓投影比较 3 类.Ramanan 等人采用自动跟踪算法,将跟踪结果与已标注的运动捕获库进行比较,对视频进行动作识别和合成[19].Ikizler 等人基于人体骨架模型跟踪运动,结合运动捕获库重建人体运动序列,然后采用 FSM(finite state machine)方法识别和检索动作[20].基于跟踪的方法识别精度受限于跟踪结果,而且采用骨架表示人体运动容易受到噪声干扰.Daniel 等人提出运动历史体(motion history volumes)视角无关的动作表示,并用于 3D 运动识别.该算法对性别、体型和视点具有较好的鲁棒性,但没有给出识别单目视频的实验[21,22].Parameswaran 基于投影不变性提出的不变量描述[23]以及黄飞跃等人基于人体轮廓信息提出的“包容形状”表示[24],都得到了较好的识别效果,但没有给出在公共数据集上的实验效果.基于轮廓投影比较是一种常用的识别模型,Ogale 等人提取视频中的人体轮廓,构建 PCFG(probabilistic context-free grammar)语法来识别人体动作[25].Weinland 等人提出一种基于 3D 样例的 HMM(hidden Markov model)算法,可识别单目视频和多目视频[26].Lv 等人采用金字塔匹配核(pyramid match kernel)算法快速比较两个人体轮廓的相似性,并提出运动网(action net)识别单目视频[27].基于轮廓投影比较法仅采用轮廓表示动作,受轮廓提取算法的约束较大,识别精度较低.

受 Lv 算法[27]的启发,本文提出运动图(action graph)的识别模型.运动图中的每个节点表示一组视频序列,视频序列被表示为加权字典向量,节点之间的相似度不是简单地基于轮廓进行比较,而是采用 Naïve Bayes 学习法,将兴趣点和形状这两种描述有机结合在一起,以提高算法的鲁棒性和识别率.

总体而言,本文算法具有如下贡献:

(1) 提出加权字典描述方法,将局部兴趣点描述的动态特征和全局形状描述的静态特征有机结合,既具有兴趣点抗噪声强的优点,又克服了兴趣点无法识别静态动作的缺点.

(2) 提出运动图识别模型,在轮廓投影比较法的基础上,引入局部兴趣点的运动描述,可以识别任意角度的人体动作.

(3) 支持多种数据格式,在学习阶段,3D运动训练数据既可以是可见外壳等点云格式,也可以是BVH等运动捕获数据格式.本文算法可识别单目视频、多目视频和多动作视频.

本文第 1 节给出算法的整体框架.第 2 节描述人体动作表示.第 3 节介绍动作图的构建和训练.第 4 节给出动作图识别视频方法.第 5 节在公共数据集上验证本文算法,并与现有算法进行比较分析.最后对本文作出总结并提出将来的工作方向.

1 本文算法整体流程

本文算法具体流程如图 1 所示,主要包含学习阶段和识别阶段.

在学习阶段,3D 动作库和投影生成的视频流是学习阶段的输入.整个学习阶段包括以下两个主要步骤:

(1) 动作表示:视频流经过背景剔除、阴影消除和形态滤波等预处理后,生成人体形状序列以及人体轮廓边界.人体形状由 Zernike 矩描述,并根据形状代码字典将视频序列表示为形状单词序列.结合人体轮廓边界,采用 Dollár 提出的兴趣点检测方法,建立兴趣点代码字典和兴趣点单词序列.经过上述处理后,每个视频被表示为两个单词序列,再根据单词数量确定权重,将视频表示为加权字典向量.

(2) 动作图生成:根据 3D 运动序列生成运动能量曲线,提取 3D 关键姿势,将运动序列分为多个序列片段,称为基本动作单元.动作单元向各个方向投影,建立动作图.根据加权字典向量描述的视频训练数据,采用 Naïve Bayes 训练动作图每个节点的统计模型.

在识别阶段,待识别数据(测试序列)可以是单目视频、多目视频或者包含多个动作的长视频序列,视频中人体运动方向可以为任意角度.整个识别阶段也分为两个主要步骤:

(1) 动作表示:与学习阶段的动作表示步骤一样,测试序列也表示为加权字典向量.

(2) 动作识别:测试序列经过采样,形成多个采样点.采样点和动作图节点之间两两计算匹配度,生成匹配度矩阵,采用 Viterbi 算法寻找 Viterbi 路径,路径匹配度之和即为该测试序列与该动作图的匹配度大小.匹配度最大的动作图类别为该测试序列的动作类别.

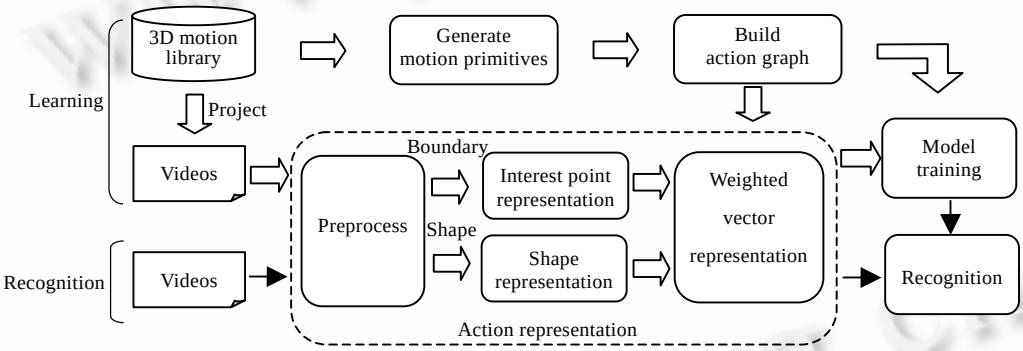


Fig.1 Flowchart of our approach

图 1 本文算法的主要流程图

2 人体动作表示

2.1 预处理

本文采用背景剔除、阴影消除和滤波提取视频前景,在人体动作视频序列中,假定视频前景为人体形状.

(1) 背景剔除

本文采用混合高斯背景模型[28,29],每个像素点颜色由多个高斯模型混合表示,均值和方差通过 N 个历史帧来计算.一般情况下,历史帧数 N 不应太小,本文选择 N=50(大约 2s).此外,为了提取“站立”等静止状态的形状,历史帧不包含与前一帧变化较小帧,防止出现前景空洞现象.

(2) 阴影消除

根据阴影与背景色调一致、亮度不同的特点,Horprasert 等人提出简单而有效的阴影消除方法[30].在 RGB 空间内,给定前景像素点 $I=[I_R,I_G,I_B]$和期望的背景像素值 $E=[\mu_R,\mu_G,\mu_B]$,定义亮度失真(bright distortion)α和颜色失真(color distortion)CD,归一化后,进一步消除非前景(背景、阴影和高亮)像素.

(3) 滤 波

经过背景剔除和阴影消除后,仍然会有少量噪声,本文采用形态滤波(腐蚀、膨胀)和中值滤波对提取的前景进行滤波处理.其中,中值滤波窗口大小为 5×5.

2.2 兴趣点表示

本文采用 Dollár[13]提出的时空兴趣点检测算法,对于给定的视频序列 I,在空域(x-y)上采用高斯核 $g(x,y;\sigma)$进行卷积,时域(t)上使用 Gabor 卷积,得到如下反应函数(response function):

$$R=(I\times g\times h_{ev})^2+(I\times g\times h_{od})^2 \tag{1}$$

其中,

$$h_{ev}(t;\tau,\omega)=-\cos(2\pi t\omega)\mathrm{e}^{-t^2/\tau^2},\ h_{od}(t;\tau,\omega)=-\sin(2\pi t\omega)\mathrm{e}^{-t^2/\tau^2} \tag{2}$$

参数σ和τ分别对应空域和时域的尺度因子.由于采用高斯和 Gabor 卷积,Dollár 反应函数对于时空角点和周期动作具有较强的反应值.与其他角点检测算法相比,Dollár 算法对“挥手”和“挠头”等周期运动具有较好的检测效果.本文使用 Dollár 推荐的参数值,$\sigma=2,\tau=4,w=4/\tau$.

在一定阈值范围内,经过非最大值抑制(non-maximal suppression)处理后,反应函数的局部极大值定义为兴趣点.以兴趣点为中心,沿 x,y,t 这 3 个方向取一些像素,构成一个小的像素体(cube),像素体的大小在各维度上为本维度尺度因子的 6 倍,即 cube 大小为 12×12×24.因此,视频序列 I 可以表示为一个像素体序列.计算像素体的梯度并将其自然展开,形成一个高维向量,采用 PCA 降维后的向量为该像素体的数值描述.除了梯度以外,还可以采用其他描述方法,例如,光流描述和 SIFT(scale-invariant feature transform)描述等.

定义 1(兴趣点单词). 在所有训练视频构成的像素体集合中,随机选择一些像素体,采用 K-均值算法,形成一个全局的兴趣点代码字典(interest-point codebook).每个像素体都对应一个与其距离最近的代码(code),称为兴趣点单词(interest-point word).

由定义 1 可知,视频序列可表示为兴趣点单词序列,例如,当代码字典为[1,2,3,4,5…,10]时,一个视频可能被表示为[1,4,5,3,1,5,2,…].

定义 2(兴趣点字典向量). 给定一个全局兴趣点代码字典 $\boldsymbol{U}=(u_1,u_2,\ldots,u_m)$以及一个兴趣点单词序列,则将该序列在代码字典 $\boldsymbol{U}$ 上的直方图定义为兴趣点字典向量,记为 $\boldsymbol{UP}=((u_1,p_1),(u_2,p_2),\ldots,(u_m,p_m))$,其中,元素$(u_i,p_i)$表示代码 u_i 在兴趣点单词序列中出现的次数为 p_i.

给定一个全局代码字典,任一视频序列经过特征点检测后,将生成一个像素体序列.根据定义 1,可转化为兴趣点单词序列;根据定义 2,该序列可进一步表示为兴趣点字典向量.

2.3 形状表示

本文采用 Zernike 前 10 阶矩描述人体形状.Zernike 矩具有较好的旋转不变性,为了保证其具有平移和缩放不变性,需利用几何矩(m_{00},m_{01},m_{10})对形状二值图像进行标准化处理,给定二值像素(x,y),标准化后为(x',y'):

$$(x',y')=(s(x-\bar{x}),s(y-\bar{y})),$$

其中,

$$\bar{x}=m_{10}/m_{00},\bar{y}=m_{01}/m_{00},s=1/\max(\sqrt{(x-\bar{x})^2+(y-\bar{y})^2}) \tag{3}$$

由于标准化后的图像 Zernike 矩前两个幅度值具有固定的理论值,因此,本文采用 34(36–2)维向量表示人体形状.

定义 3(形状单词). 与定义 1 类似,在所有训练视频的形状中随机选择一些形状,采用 K-均值算法形成一个全局的形状代码字典(shape codebook),每个形状都对应一个与其距离最近的代码,称为形状单词(shape word).

定义 4(形状字典向量). 与定义 2 类似,给定一个全局形状代码字典 $\boldsymbol{V}=(v_1,v_2,\ldots,v_n)$以及一个形状单词序列,

则该序列在代码字典 $\boldsymbol{V}$ 上的直方图定义为形状字典向量,记为 $\boldsymbol{VQ}=((v_1,q_1),(v_2,q_2),\ldots,(v_n,q_n))$.

形状字典向量和兴趣点字典向量都需依托于全局的代码字典,代码字典在整个训练和识别过程中保持不变,其大小会对识别结果产生一定的影响,需根据识别率进行选择.

2.4 加权字典向量

视频序列 $\boldsymbol{I}$ 可表示为兴趣点字典向量和形状字典向量.对于视频 $\boldsymbol{I}$ 的任意子视频序列 $\boldsymbol{I}'$,根据时序关系,可以得到对应的两个单词子序列,从而也可表示为两个字典向量.为了整合兴趣点特征和形状特征,本文提出加权字典向量的表示方法,

定义 5(加权字典向量). 任一视频序列(或其子序列)$\boldsymbol{I}$,令其兴趣点字典向量和形状字典向量分别为 $\boldsymbol{UP}=((u_1,p_1),(u_2,p_2),\ldots,(u_m,p_m))$和 $\boldsymbol{VQ}=((v_1,q_1),(v_2,q_2),\ldots,(v_n,q_n))$,则视频序列 $\boldsymbol{I}$ 的加权字典向量定义为

$$((u_1,\lambda p_1),\ldots,(u_m,\lambda p_m),(v_1,(1-\lambda)q_1),(v_n,(1-\lambda)q_n)),$$

其中,$\lambda=f\left(\sum_{i=[1:m]}p_i\Big/\sum_{i=[1:n]}q_i\right)$.权重曲线 $f(x)$为双曲正切函数,如公式(4)所示.

$$f(x)=\frac{e^{kx}-e^{-kx}}{e^{kx}+e^{-kx}} \tag{4}$$

其中,自变量 x 为视频序列 $\boldsymbol{I}$ 对应兴趣点单词数量与形状单词数量之比,理论上取值范围$[0,\infty)$.但实际上,兴趣点单词的数量最大值和形状单词的数量差别不大.

参数 k 控制加权向量对于兴趣点和形状的依赖性,如图 2 所示,k 越大,兴趣点字典向量得到权重的机会越大,意味着加权向量对兴趣点的依赖性越强;反之,k 越小,则对形状的依赖性越强.由于形状特征容易受背景剔除等算法的影响,而兴趣点特征具有体形无关性和鲁棒性高等优点,因此本文取 k=8,对视频的描述更多地依赖于兴趣点特征,只有当兴趣点数据较少时,形状特征才会取得较大的权重.

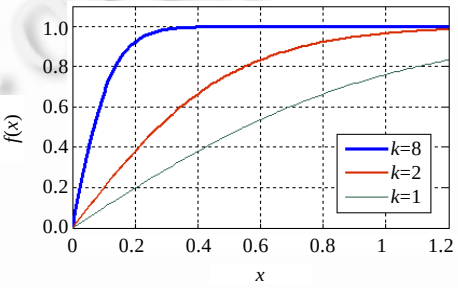


Fig.2 Weight curve

图 2 权重曲线

包含人体动作的一段视频经过兴趣点表示、形状表示和加权字典向量表示 3 步后,生成一个根据视频自动调整权重的加权字典向量,该向量具有固定的长度 $m+n$,即兴趣点代码字典和形状代码字典大小之和.

3 动作图

为了达到视角无关的动作识别,一种有效的方法是引入与视角无关的中间数据,3D 运动库是一个较好的选择.受 Lv[27]算法中 Action Net 的启发,本文提出动作图识别模型,为每个动作构建一个动作图,将 3D 动作库和 2D 视频流组织在一起.整个动作图构建过程主要分为 3 步:基本运动单元生成、动作图建立和节点概率模型计算.

3.1 基本运动单元生成

三维运动训练数据可以是运动捕获数据,也可以是可见外壳点云数据.这两种数据格式具有不同的数学表示.运动捕获数据采用人体骨架描述,每一个姿势 p 可表示为关节点列表.点云数据采用柱面坐标系离散傅里叶变换表示[22],这种表示具有平移和沿 z 旋转不变性等特点.

定义 6(基本运动单元). 具有单一运动特征的三维运动片段定义为基本运动单元,运动特征由段内关键姿势描述.表示为二元组$\langle keypose,segment\rangle$.

由三维运动序列生成基本运动片段包括两个核心步骤:运动分割和关键姿势提取.

(1) 运动分割本身为一个研究难点.本文采用一种简单的均匀分割方法,根据给定的片段数 k,将三维序列等分为 k 段.

(2) 使用运动能量曲线提取段内关键姿势,给定的三维运动片段 $segment=(p_1,p_2,\ldots,p_L)$,第 i 帧姿势的运动能量定义为 $E_i=\|p_i-p_{i-1}\|^2$,其中$\|\cdot\|$为欧式距离.运动变化越剧烈,则运动能量越大.因此,运动能量曲线中的最大值

$\max_{i\in[1:L]}(E_i)$ 对应的姿势定义为关键姿势,表示整个片段的运动特征.

3.2 运动图建立

定义 7(本质图). 给定一个动作类别及其多个三维运动训练序列,这些运动序列采用 DTW(dynamic time warping)对齐后,随机选择一个作为参考序列,根据片段数 k 生成 k 个基本运动单元,每个基本运动单元称为一个节点,节点之间构成的有向图定义为本质图.

本质图的两个最重要的元素是节点和连接.节点即三维参考序列的基本运动单元,根据 DTW 对应关系,每个节点都对应一组运动片段.节点之间的连接关系分为 3 种:自连接、向前连接和向后连接.如图 3 所示,每个节点都存在自连接,前后相邻的两个节点存在向前连接.末节点和首节点之间根据关键姿势建立向后连接,如果这两个节点关键姿势差别小于一个阈值(运动能量曲线均值/3),则建立向后连接,数学上,$\|kepose_{last}-kepose_{first}\|<\eta$,其中,阈值 $\eta=\sum_{i=1}^{L}E_i\Big/(3L)$,$L$ 为参考序列的总长度.向后连接可以表示"走"、"跑"和"挥手"等循环动作.

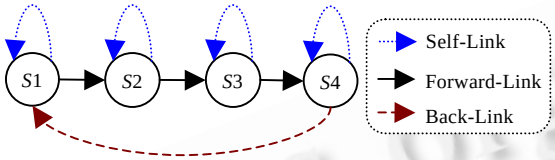


Fig.3 Types of linking

图 3 节点连接类型

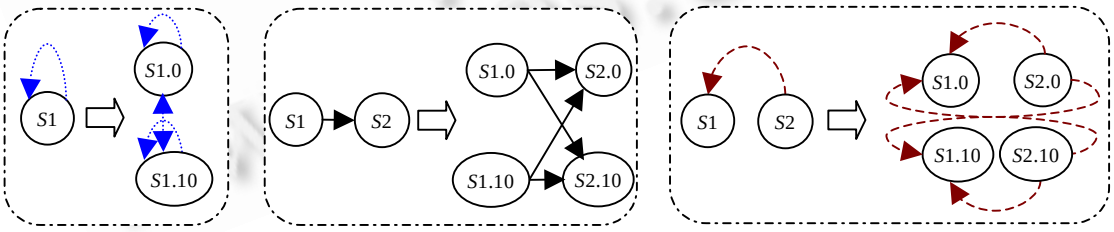
为了可以识别视频序列,需要将本质图沿各个观察角度投影,从三维降到二维,这样才可以判断测试视频序列与节点之间的匹配度,从而分类该视频.

定义 8(运动图). 本质图沿各个观察角度投影,本质图中每个节点投影为多个节点,连接关系按照扩展规则进行扩展相连,形成的新的有向图定义为运动图,记为 A.

在由本质图生成动作图的过程中需遵守两个规则:选择投影和相邻扩展.

(1) 选择投影.理论上,为了达到任意角度的识别,需要将本质图向整个观察球面进行投影.而实际上,向整个球面投影将导致运动图过于庞大,可以根据待识别视频的大概观察角度有选择性地投影.例如,如果待识别视频相机相对于人的高度已知(可根据视频进行估计)且观察方向与地面平行,人的运动朝向和相机旋转未知,则本质图只需在相机水平面上沿各个角度投影即可.

(2) 相邻扩展.如果两个投影角度相邻,运动序列可以在这两个投影角度之间较为平滑地过渡,则这两个投影为相邻投影,只有相邻投影节点之间才需扩展连接.相邻投影节点之间连接的扩展规则如图 4 所示,自连接节点投影后,相邻节点之间需建立互联;向前连接投影后,相邻节点需要建立交叉向前连接;向后连接投影后,末节点需要与相邻投影的头节点建立向后连接.



(a) Extend self-link (a) 扩展自连接

(b) Extend forward-link (b) 扩展向前连接

(c) Extend back-link (c) 扩展向后连接

Fig.4 Extending rules of linking

图 4 连接扩展规则

将图 3 所示的本质图沿一水平面每隔 10°投影一次,形成的动作图如图 5 所示.0°投影与 10°投影相邻,10°投影与 20°投影相邻,...,350°投影与 0°投影相邻,因此,36 个投影构成一个柱状动作图.

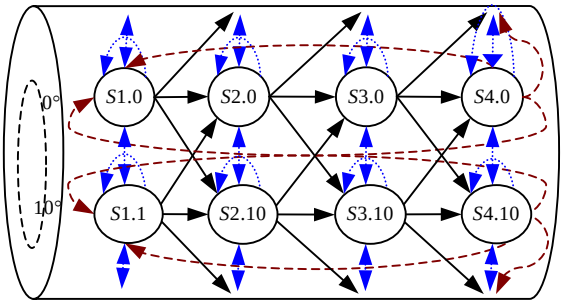


Fig.5 Example of action graph

图 5 动作图示例

运动图的每个节点由本质图节点沿某个角度的投影形成,因此,本质图节点对应的所有运动片段沿这个角度投影后形成的视频序列就成为运动图节点的训练数据.

动作图的节点数量决定了对视频中动作旋转的敏感程度,节点数量越多,识别过程中在各投影角度节点上跳转的机会也越多,意味着对于运动旋转越敏感,则越容易识别运动方向变化较快的动作,例如“转圈走”等.但节点数量越大,该节点对应的视频训练数据长度也越短,兴趣点单词和形状单词数量也越少,越容易受噪声干扰.因此,本文选择本质图节点的数量为 1~4 个.

3.3 节点概率模型计算

生成动作图后,每个节点都包含一组视频训练数据,这些训练数据都具有相同的运动特征和观察角度.基于这些训练数据,本文采用 Naïve Bayes 来训练节点概率模型,通过概率模型计算视频与节点之间的匹配度.给定一个视频序列 I,具有与节点 S 相同运动特征的概率为

$$p(S\mid I)=\frac{p(I\mid S)p(S)}{p(I)}\propto p(I\mid S)p(S) \tag{5}$$

由于运动图每个节点发生的概率都相等,因此公式(5)最关键的部分为条件概率 $p(I|S)$.视频序列 I 由加权字典向量$((u_1,\lambda p_1),\ldots,(u_m,\lambda p_m),(v_1,(1-\lambda)q_1),\ldots,(v_n,(1-\lambda)q_n))$表示,归一化后,令 I 表示为$((u_1,a_1),(u_2,a_2),\ldots,(u_m,a_m),(v_1,b_1),(v_2,b_2),\ldots,(v_n,b_n))$,其中 $\sum_{i=1}^{m}a_i\sum_{i=1}^{n}b_i=1$,假设 I 符合自然假设(naïve assumption),也即每个单词之间相互独立.于是,可以得到:

$$\begin{aligned}p(I\mid S)&=p((u_1,a_1),\ldots,(u_m,a_m),(v_1,b_1),(v_n,b_n))\mid S)\\&=\prod_{i=1}^{m}p((u_i,a_i)\mid S)\\&=\prod_{i=1}^{n}p((v_i,b_i)\mid S)\\&=\prod_{i=1}^{m}p(u_i\mid S)^{a_i}\prod_{i=1}^{n}p(v_i\mid S)^{b_i}\end{aligned} \tag{6}$$

其中,$p(u_i|S)$和 $p(v_i|S)$为在节点 S 中,兴趣点代码字典 U 的代码 u_i 和形状代码字典 V 的代码 v_i 出现的概率,这两个条件概率可由节点 S 对应的训练数据估计得到.为了防止出现“除 0”问题,采用拉普拉斯平滑(Laplacian smoothing)来处理:

$$p(u_i \mid S)=\frac{1+\sum_{I_j \in S} p_{ji}}{|U|+\sum_{k\in[1:|U|]}\sum_{I_j \in S} p_{jk}},\ p(v_i \mid S)=\frac{1+\sum_{I_j \in S} q_{ji}}{|V|+\sum_{k\in[1:|V|]}\sum_{I_j \in S} q_{jk}} \tag{7}$$

其中,p_{ji}表示在训练视频 I_j的兴趣点字典向量 ***UP*** 中,代码 u_i出现的次数;q_{ji}表示在视频 I_j的形状字典向量 ***VQ*** 中,代码 u_i出现的次数.

公式(5)得到的 $p(S|I)$为视频序列 I 与节点 S 之间的匹配度.在实际计算时,采用 log 值作为匹配度.

4 动作识别

假设已训练的动作图 A 有 N 个节点,测试序列(待识别视频序列)I 和该动作图的匹配度计算流程如下:

(1) 视频采样.从 I 中任意提取一些样本点,例如每一秒提取一个样本,假设样本点个数为 M.

(2) 计算匹配度矩阵.以样本点为中心,在测试序列中截取和节点基本运动单元相同长度的子序列,子序列与节点的匹配度定义为样本点与节点的匹配度.计算所有节点和样本点之间的匹配度,构成一个 $N\times M$ 矩阵.

(3) 最大值路径搜索.采用路径搜索算法(例如 HMM 向前算法和 Viterbi 算法等)在匹配度矩阵中寻找具有最大匹配度之和的路径.本文采用 Viterbi 算法,Viterbi 路径(Viterbi path)即为所求的路径.路径匹配度之和定义为视频与动作图的匹配度.

4.1 单目序列识别

给定 k 个运动类别,每个类别对应一个动作图 A,则待识别视频序列 I 的类别为

$$Category=\arg\max_{i\in[1:k]}(Score(A_i,I)) \tag{8}$$

其中,$Score(A_i,I)$表示视频序列 I 与第 i 个动作图的匹配度.

4.2 多目序列识别

给定 l 个待识别的视频$\{I_1,I_2,\ldots,I_l\}$,属于同一运动类别,可能有不同的观察角度,则这些视频的类别为

$$Category=\arg\max_{i\in[1:k]}\left(\sum_{j=1}^{l}Score(A_i,I_j)\right) \tag{9}$$

4.3 多动作序列识别

如果一个视频序列中包括多个运动动作,则需要采用滑窗识别的策略,根据滑窗内子序列类别设定滑窗中心的类别,随着滑窗的移动,生成一条类别曲线.由于视频序列中的动作变化平滑,因此可以采用滤波器(本文采用最大值滤波器)过滤类别曲线,降低噪声干扰,提高识别率.

5 实 验

为了验证本文算法的有效性,使用 Matlab 实现,并基于两个公共运动数据集进行实验、对比和分析,实验 1 在 IXMAS 数据集上计算单/多目识别率,并与使用同一数据集的最新相关文献 Weinland 算法[26]进行比较.实验 2 验证本文算法在识别多动作序列方面的有效性.

5.1 基于IXMAS数据集单/多目识别

IXMAS 数据集是一个公共开放的视频数据库**,包含 12 个采集者做 11 个动作(check watch,cross arms, scratch head,sit down,get up,turn around,walk,wave,punch,kick,pick up)的全过程,每人重复 3 次,整个运动过程从 5 个不同角度的相机进行观察,如图 6 所示.IXMAS 数据集提供点云运动序列、5 个同步的多角度视频序列、背景序列以及相机标定参数等.

我们将数据集手工分割组织,使每个动作类别包含 36(12×3)个三维点云序列以及 180(12×3×5)个视频序列.

** The dataset is obtained from the Perception website, http://charibdis.inrialpes.fr/html/non-secure-sequence.php?s=IXMAS.

随机选择 3 个采集者数据建立兴趣点代码字典 U 和形状代码字典 V、5 个采集者数据训练动作图,剩余 4 个采集者数据作为测试数据验证算法.为了节省内存和计算开销,所有视频都采样到 160×120.兴趣点代码字典 U 和形状代码字典 V 的大小都为 500,本质图的片段数 k 设置为 2,这 3 个参数的选择在本实验最后给出说明和依据.实验中识别率为运行 10 次的平均值.

(a) Camera 1 (b) Camera 2 (c) Camera 3 (d) Camera 4 (e) Camera 5
(a) 相机 1 (b) 相机 2 (c) 相机 3 (d) 相机 4 (e) 相机 5

Fig.6 Example of the IXMAS dataset

图 6 IXMAS 数据集示例

由于 IXMAS 数据集中只有 5 个相机,而且位置和方向全部已知,因此,动作图仅包含这 5 个投影方向.相机 5 的位置与其他 4 个相机的位置区别较大,运动无法平滑过渡到其他相机.因此,本文选择相邻规则如下:相机 1 与相机 2 相邻、相机 2 与相机 3 相邻、相机 3 与相机 4 相邻、相机 4 与相机 1 相邻,而相机 5 没有相邻投影方向.

5 个相机共有 31 种组合,每一种相机组合的平均识别率如图 7 所示,相机 1~相机 5 的识别率分别为 81.82%,78.66%,74.75%,75.76%,70.70%.由于相机 5 位于采集者头顶,自遮挡严重,导致兴趣点个数较少,形状区别不明显,因此识别率较低.这一点与 Weinland 得到的结论类似.多目相机平均识别率为 85.27%,4-目或者 5-目相机识别率普遍较高,多数达到识别率 87.88%.图 8 为单/多目相机下平均识别率曲线.由图 8 可知,随着待识别视频目数的增加,平均识别率也随之增加,这与直观上相机越多,识别信息越充分的推论相一致.

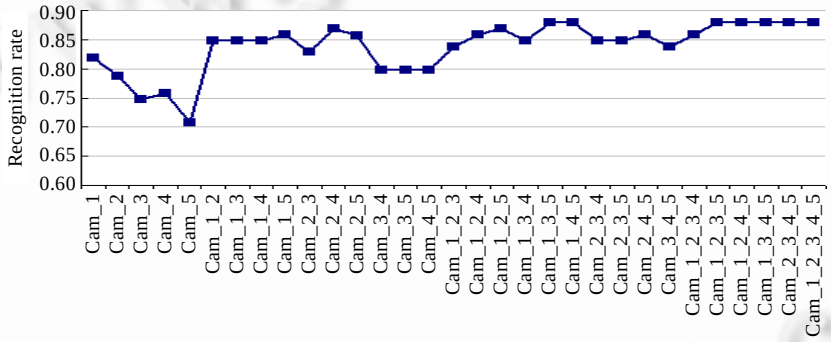


Fig.7 Recognition rates for all camera combinations

图 7 所有相机组合的识别率

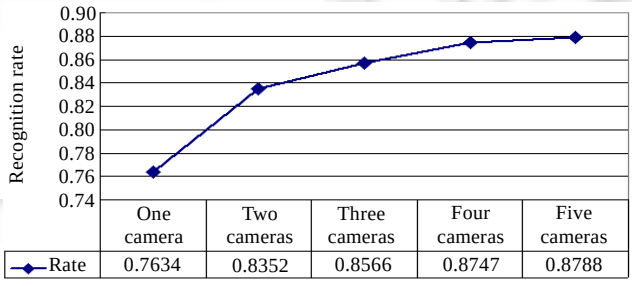


Fig.8 Relationship between recognition rate and camera number

图 8 识别率与目数的关系

图 9 给出单目相机 1 和多目情况(1,2,3,4,5)的识别率混淆矩阵(confusion matrix).由两个图可知,wave 与 scratch head 相似性较大,check watch 和 punch 相似性较大.从测试数据可以发现,前两个动作都是只有 1 只手在头部附近晃动,后两个动作都为抬手动作.因此,实验结果与实际数据相符.

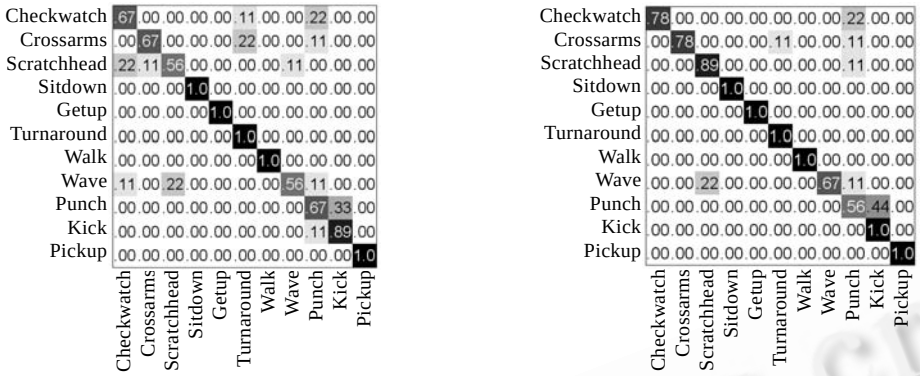


Fig.9 Confusion matrixes of camera 1 and multi-cameras 1, 2, 3, 4 and 5

图 9 相机 1 和多目相机 1-2-3-4-5 下的识别率混淆矩阵

表 1 给出了本文算法与 Weinland 算法在同一数据集下的结果对比.由于 Weinland 算法仅给出了 10 种相机组合的实验结果,因此无法对所有相机组合进行分析比较.由对比可以看出,在同一数据集下,本文算法比 Weinland 算法取得了更高的识别率.

Table 1 Comparison between our approach and Weinland approach

表 1 本文算法与 Weinland 算法对比

Camera combinations	Recognition rate (%)	
	Our approach	Weinland approach[26]
1	81.82	65.4
2	78.66	70.0
3	74.75	54.3
4	75.76	66.0
5	70.70	33.6
2 4	86.87	81.3
3 5	79.80	61.6
1 3 5	87.88	70.2
1 2 3 5	87.88	75.9
1 2 3 4	85.86	81.3

上述实验结果都是在 U,V 大小为 500、片段数 k 为 2 的情况下进行的.实际上,这 3 个因素都会对识别率造成影响,代码字典 U 和 V 的数量决定视频加权字典表示的维度,片段数 k 决定算法对旋转运动的敏感性.为了选择这 3 个参数,本文选择 U,V 大小范围为[100,200,300,400,500,600],k 的范围为[1,2,3,4,5],在所有的组合情况下计算单/多目识别率的平均值,得到 6×6×5 个识别率矩阵,当 U=500,V=500,k=2 时取得最大值.

5.2 基于CMU运动捕获数据集多动作序列识别

本实验用来验证本文算法识别多动作视频序列的有效性,待识别视频序列的动作可以为任意运动方向.我们从 CMU 运动捕获库[***]中选择 6 个动作库作为动作图的训练数据,这 6 个动作为 boxing,jumping,kicking,running,walking 和 standing,每个动作包含 5 种不同运动风格.

在与人体高度几乎相等的水平面上,每隔 45°设置一个观察相机,共 8 个相机,相邻规则为:相机 1 与相机 2 相邻,相机 2 与相机 3 相邻,相机 3 与相机 4 相邻,...,相机 7 与相机 8 相邻,相机 8 与相机 1 相邻,构成一个柱形动

*** The library is available at http://mocap.cs.cmu.edu/.

作图.因此,每个动作类别包含 5 个运动捕获序列以及 40(5×8)个视频序列.

测试序列(待识别视频序列)帧率为 36fps,包括 5 个运动类别(boxing,jumping,running,walking 和 standing),具有任意运动方向.滑动窗口的大小设置为 90,最大值滤波器窗口参数设置为 31.所有视频(训练视频和测试视频)都下采样到 160×120.选择与实验 1 相同的参数值 U,V 大小为 500,片段数 k 为 2.

过滤后,类别曲线如图 10 所示,横坐标为序列帧数,纵坐标为动作类别(1-boxing,2-jumping,3-kicking,4-running,5-walking 和 6-standing).序列中不同动作之间的过渡片段是影响识别的主要因素,本文假设两个动作 a 与 b 之间的过渡片段既可以属于 a 类,也可以属于 b 类.基于这个假设,与手工标注结果相比,过滤后识别正确率达到 90.54%.

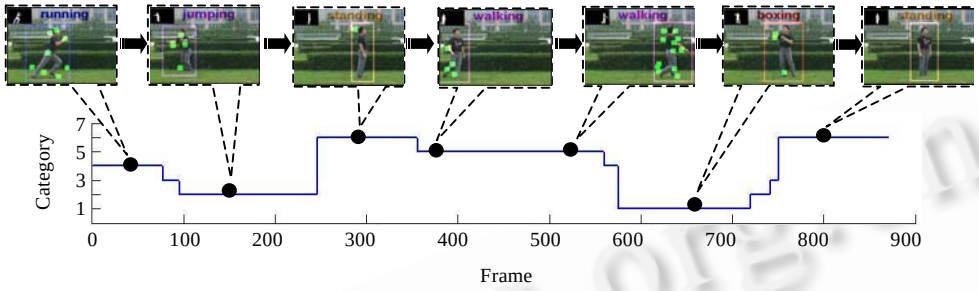


Fig.10 Recognition result of multi-action video

图 10 多动作视频识别结果

6 结束语

针对任意角度的动作识别,本文提出加权字典向量描述方法和动作图识别模型.将视频中的局部动态特征(兴趣点描述)和全局静态特征(形状描述)结合起来,形成加权字典向量的描述方法.基于这种描述方法提出动作图识别模型,采用 Naïve Bayes 和 Viterbi 算法计算视频与动作图之间的匹配度,最大匹配度的动作图类别即为该视频的动作类别.动作图实质上是本质图的投影,任意角度的视频总可以找到与之相匹配的投影方向,而且动作图的 3 种连接关系可以保证视频拍摄角度或者人体运动方向的任意切换.因此,如果在投影方向采样足够多的情况下,则可以识别任意角度、任意运动方向的视频.

在以后的工作中,我们将对本文方法进一步深入研究,主要包括:

(1) 动作图与 HMM 的关系分析:动作图与 HMM 具有较多相似之处,都采用有向图描述,而且匹配度计算方法类似.在 HMM 中加入投影信息并与动作图进行对比分析是本文的一个研究方向.

(2) 三维兴趣点描述:动作图仍然属于投影匹配法的范畴,直接在三维动作库中寻找投影不变性的三维兴趣点,并将其应用于动作和行为识别是本文的另一个研究方向.

致谢 非常感谢 Piotr Dollár 提供兴趣点检测 Matlab 工具箱.

References:

[1] Gavrila DM. The visual analysis of human movement: A survey. Computer Vision and Image Understanding, 1999,73(1):82−98.

[2] Wang L, Hu WM, Tan TN. Recent developments in human motion analysis. Pattern Recognition, 2003,36(3):585−601.

[3] Aggarwal JK, Park S. Human motion: Modeling and recognition of actions and interactions. In: Proc. of the 3D Data Processing, Visualization, and Transmission, the 2nd Int'l Symp. Washington: IEEE Computer Society, 2004. 640−647. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1335299

[4] Moeslund TB, Hilton A, Kruger V. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 2006,104(2):90−126.

[5] Ahmad M, Lee S. Human action recognition using shape and CLG-motion flow from multi-view image sequences. Pattern Recognition, 2008,41(7):2237−2252.

[6] Ahmad M, Lee S. HMM-Based human action recognition using multiview image sequences. In: Proc. of the 18th Int'l Conf. on Pattern Recognition, Vol.01. Washington: IEEE Computer Society, 2006. 263−266. http://dx.doi.org/10.1109/ICPR.2006.630

[7] Efros AA, Berg AC, Mori G, Malik J. Recognizing action at a distance. In: Proc. of the 9th IEEE Int'l Conf. on Computer Vision, Vol.2. Washington: IEEE Computer Society, 2003. 726. http://portal.acm.org/citation.cfm?id=946720

[8] Bobick AF, Davis JW. The recognition of human movement using temporal templates. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001,23(3):257−267.

[9] Yilmaz A, Shah M. Actions sketch: A novel action representation. In: Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005), Vol.1-Vol.01. Washington: IEEE Computer Society, 2005. 984−989. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=1467373&isnumber=31472

[10] Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(12):2247−2253.

[11] Laptev I. On space-time interest points. Int'l Journal of Computer Vision, 2005,64(2-3):107−123.

[12] Oikonomopoulos A, Patras I, Pantic M. Spatiotemporal salient points for visual recognition of human actions. IEEE Trans. on Systems, Man, and Cybernetics, 2006,36(3):710−719.

[13] Dollár P, Rabaud V, Cottrelln G, Belongie S. Behavior recognition via sparse spatio-temporal features. In: Proc. of the 14th Int'l Conf. on Computer Communications and Networks. Washington: IEEE Computer Society, 2005. 65−72. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=1570899&isnumber=33252

[14] Schuldt C, Laptev I, Caputo B. Recognizing human actions: A local SVM approach. In: Proc. of the 17th Int'l Conf. on Pattern Recognition. Washington: IEEE Computer Society, 2004. 32−36. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1334462

[15] Wong S, Cipolla R. Extracting spatiotemporal interest points using global information. In: Proc. of the 11th IEEE Int'l Conf. on Computer Vision. Los Alamitos: IEEE Computer Society, 2007. 1−8. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4408923

[16] Niebles JC, Wang H, Li FF. Unsupervised learning of human action categories using spatial-temporal words. In: Proc. of the British Machine Vision Conf. (BMVC). The British Machine Vision Association, 2006. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.83.8353

[17] Wong S, Kim T, Cipolla R. Learning motion categories using both semantic and structural information. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2007. 1−6. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4270330

[18] Niebles JC, Wang H, Li FF. Unsupervised learning of human action categories using spatial-temporal words. Int'l Journal of Computer Vision, 2008,79(3):299−318.

[19] Ramanan D, Forsyth DA. Automatic annotation of everyday movements. Technical Report, CSD-03-1262, UC Berkeley, 2003.

[20] Ikizler N, Forsyth D. Searching video for complex activities with finite state models. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2007. 1−8. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4270193

[21] Weinland D, Ronfard R, Boyer E. Automatic discovery of action taxonomies from multiple views. In: Proc. of the 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2006. 1639−1645. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1640952

[22] Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding, 2006,104(2):249−257.

[23] Parameswaran V, Chellappa R. View invariance for human action recognition. Int'l Journal of Computer Vision, 2006,66(1):83−101.

[24] Huang FY, Xu GY. Viewpoint independent action recognition. Journal of Software, 2008,19(7):1623−1634 (in Chinese with English abstract). http://www.jos.org.cn/ 1000-9825/19/1623.htm

[25] Ogale AS, Karapurkar A, Aloimonos Y. View invariant modeling and recognition of human actions using grammars. In: Proc. of the Int'l Conf. on Computer Vision, Workshop on Dynamical Vision (ICCV-WDM). Berlin, Heidelberg: Springer-Verlag, 2005. http://dx.doi.org/10.1007/978-3-540-70932-9_9

[26] Weinland D, Boyer E, Ronfard R. Action recognition from arbitrary views using 3D exemplars. In: Proc. of the 11th IEEE Int'l Conf. on Computer Vision. Los Alamitos: IEEE Computer Society, 2007. 1–7. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4408849

[27] Lv F, Nevatia R. Single view human action recognition using key pose matching and Viterbi path searching. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2007. 1–8. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=4270156&isnumber=4269956

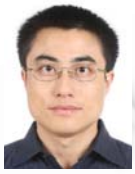
[28] Zivkovic Z, Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters, 2006,27(7):773–780.

[29] Zivkovic Z. Improved adaptive Gaussian mixture model for background subtraction. In: Proc. of the 17th Int'l Conf. on Pattern Recognition. Washington: IEEE Computer Society, 2004. 28–31. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1333992

[30] Horprasert T, Harwood D, Davis LS. A statistical approach for real-time robust background subtraction and shadow detection. In: Proc. of the IEEE ICCV Frame-Rate Workshop. 1999. 1–19. http://www.citeulike.org/user/nob/article/1402206

附中文参考文献:

[24] 黄飞跃,徐光祐.视角无关的动作识别.软件学报,2008,19(7):1623−1634. http://www.jos.org.cn/1000-9825/19/1623.htm



杨跃东(1980－),男,内蒙古集宁人,博士生,主要研究领域为人体动画,运动识别,虚拟现实.

赵沁平(1948－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为虚拟现实,分布式应用系统,人工智能.

郝爱民(1968－),男,博士,教授,CCF 高级会员,主要研究领域为虚拟现实,可视化,数据库.

王莉莉(1977－),女,博士,副教授,CCF 高级会员,主要研究领域为计算机图形学,三维逼真,实时绘制.

褚庆军(1966－),男,助理研究员,主要研究领域为计算机网络安全,软件开发,视频监控技术.