文章编号:1671-4598(2014)07-2148-03　　中图分类号:TP393　　文献标识码:A

一种改进的人类动作识别和定位算法研究

周晓青
(太原工业学院 计算机工程系,太原 030008)

摘要:针对现有人类动作识别和定位方法的不足,提出一种改进的人类动作识别和定位算法;该算法首先对每个视频帧进行分层分段,得到一组分段树,每颗树是身体分段树的候选;接着利用视频的轮廓、接合对象结构、全局前景色等信息对候选分段树进行修剪;最后在时域上对剩余分段层的每个分段进行前向和后向跟踪;基于难度较大的 UCF-Sports 和 HighFive 数据集对本文方法进行性能评估,实验结果表明,文章方法的性能要优于当前最新运动检测算法性能,运动定位性能与当前最新算法相当。

关键词:人类动作识别;定位;多分辨率;分段树;修剪

Research on An Improved Human Action Recognition and Positioning Algorithm

Zhou Xiaoqing
(Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan 030008, China)

Abstract: Aiming at the disadvantages of the existing human action recognition and positioning methods, an improved human action recognition and positioning algorithm is proposed. First, hierarchical segmentation is applied on each video frame to get a set of segment trees, each of which is considered as a candidate segment tree of the human body. Second, we prune the candidates by exploring several cues such as shape, articulated objects' structure and global foreground color. Finally, we track each segment of the remaining segment trees in time both forward and backward. The experimental results show that, the performance of our method is better than the state-of-art action recognition methods on two challenging benchmark datasets UCF-Sports and HighFive, and at the same time produce good action localization results.

Keywords: human action recognition; positioning; multi-grained representation; segment trees; prune

0 引言

人类动作识别是视频检索领域的主要研究课题[1],现有的动作识别方法大多重点关注视频的非静态部分而忽略大部分静态部分,从而影响了动作识别的精确性。实际上视频的非静态部分和相关静态部分对运动识别和定位均具有重要作用,应综合参考。文献[2]提出了一种人体模型动画自动生成方法。该方法首先自动提取和识别位于人体四肢和头顶末端的5个特征点,然后根据人体骨骼刚性运动特征和运动数据文件提供的骨骼信息,进行关节点精确定位,最后将提取的骨骼采用局部坐标架对齐的方法实现与运动数据匹配。文献[3]提出一种基于主题模型的人体动作识别方法。该方法首先提取时空兴趣点来描述人体运动,然后使用慢特征分析算法计算兴趣点梯度信息不变量最优解,最后使用概率潜在语义分析模型识别人体动作。另外,文献[4-5]提出的基于STIP或密集轨迹的运动识别方法在多种基准数据集上取得了优异性能。然而,STIP或密集轨迹中的空间-时间关系并没有明确说明。STIP或密集轨迹的高阶统计信息也被应用于运动识别,例如文献[6]的配对方法,但是这些研究没有同时考虑运动识别和运动定位。基于人体外形整体表示的运动识别方法可以实现运动者定位[7]。但是这些方法的稳健性不够,难以处理实际视频中的掩蔽现象和杂乱背景。基于预先训练好的人体部位或人体检测器也可以实现运动者定位。然而,它们的检测器受到隐藏在训练集中的人体部位外形先验信息的约束,导致检测器的可拓展性有限,难以处理各种运动不同程度的掩蔽现象和姿态。

为此,本文首先提出一种改进的人类动作识别和定位算法。基于UCF-Sports[8]和HighFive[9]数据集的实验结果表明,本文方法的性能要优于当前最新算法。

1 分层空间-时间段

1.1 视频帧分层分割

本文视频帧分段的主要思路是:使用颜色和运动信息来缩小背景和刚性对象内的边界,增强人体不同部位运动导致的人体内部运动边界。后续分层分段进一步降低背景和刚性对象的不相关分段,保持人体密集多分辨率分段。对每个视频帧,根据文献[10]中的方法,使用3种颜色信道和5种运动信道(包括光流,单元归一化光流和光流幅值)来计算边界图。然后根据文献[11]中的方法结合边界图来计算Ultrametic等高线图(UCM)。

UCM表示一个视频帧的分层分段,根是整个视频帧。我们遍历该分段树以去除冗余段或者无效段。接着删除分段树的根,获得一组分段树 T^t(t 是帧的索引)。每个 $T_j^t \in T^t$ 看成是人体分段树候选,且有 $T_j^t = \{s_{ij}^t\}$,其中每个 s_{ij}^t 表示一个分段,s_{0j}^t 是根段。图1最右侧给出了2个候选分段树示例,这两个分段树在后续修剪步骤后仍然保留。

1.2 修剪候选分段树

为了提取静态及非静态相关段,需在修剪过程中保留静态

收稿日期:2014-03-12;　**修回日期**:2014-04-14。
基金项目:国家自然科学基金(61172035);国防预研基金。
作者简介:周晓青(1978-),女,山西代县人,硕士研究生,讲师,主要从事视频检索技术,图像重构算法方向的研究。

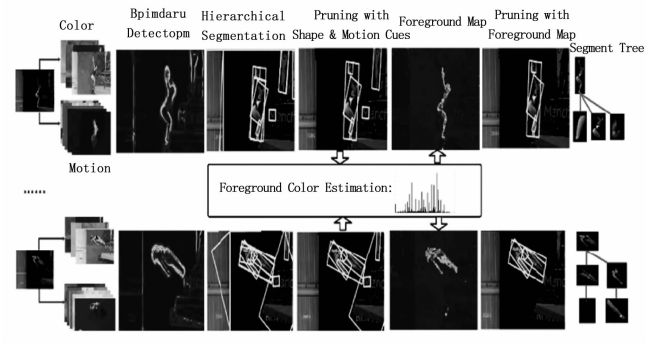

图 1　分层视频帧的段提取流程

但相关的段。要实现这一点，必须分析段间的分层关系，以决定某个段是被修剪还是保留，此外还要考虑段本身包含的局部信息，考虑相同候选分段树内所有段的信息。所有后续修剪均是在候选层面进行，一个候选分段树要么彻底删除，要么连同其分段保留。通过这种方法，即使很小的身体部位发生运动，我们也可以提取整个人体。

我们分析多个与运动相关的线索以修剪候选分段树，根据相关流程（图 1）的顺序描述如下。

（1）形状线索：带有直线边界的背景对象（比如楼宇）是常见的人造场景，但是人体边界包含的零曲率点较少。于是，对每个候选 $T_j^t \in T^t$，我们计算其根段 s_{0j}^t 所有边界点的曲率。如果曲率近似为 0 的点的比率较大（文中为大于 0.6），则我们去除 T_j^t。边界点 (x_a, y_a) 处的曲率 κ_a 计算如下：

$$\kappa_a = \left| \frac{x_a - x_{a+\delta}}{y_a - y_{a+\delta}} - \frac{x_{a-\delta} - x_a}{y_{a-\delta} - y_a} \right| \tag{1}$$

式中，$(x_{a-\delta}, y_{a-\delta})$ 和 $(x_{a+\delta}, y_{a+\delta})$ 为 (x_a, y_a) 在分段边界上的两个邻近点。

（2）运动线索：对每个段 $s_{ij}^t \in T_j^t$，我们计算 s_{ij}^t 内所有像素的平均运动幅度。为了计算真实的运动幅度，我们计算当前帧和相邻帧间的仿射变换矩阵，以近似相机运动并相应调整流场。如果有一个或一个以上的平均运动幅度大于阈值，则保留 T_j^t，否则将其修剪。我们分析两种一般假设，以估计前景图，并对候选做进一步修剪。首先，作为一种接合对象，非静态人体区域往往包含许多由不同身体部位运动而产生的内部运动边界。其次，前景人体分段呈现的一致性往往要高于人工边缘或错误分段操作而生成的分段。我们将这种现象称为整个视频序列的全局颜色线索。前景图构建方法如下：

设 $\hat{T}^t$ 表示基于形状和运动线索修剪后剩余的候选树集合，S 表示所有帧的所有剩余分段集合，即：$S = \{s_{ij}^t \mid \forall s_{ij}^t \in T_j^t, \forall T_j^t \in \hat{T}^t, \forall t\}$。为了避免标记法过于混乱，在下文中有 $S = \{s_k\}$。对每个 $s_k \in S$，计算 $L\infty$ 归一化颜色直方图 c_k。于是，前景色直方图 c 被所有分段的投票结果为：

$$c = \sum_{s_k \in S} 2^{h_k} \cdot c_k \tag{2}$$

其中：h_k 是分段树中段 s_k 的高度。对根段 s_{0j}^t，我们定义其高度 h_{0j}^t 为 1，对非根段 s_{ij}^t，其高度 h_{ij}^t 设为路径上从根到该段的边缘数量再加 1。然后，对颜色直方图 c 做 $L1$ 规一化。如式（2）所示，出现频率较高的分段的颜色和高度较大的分段的颜色，获得的投票也较多。

设 F^t 表示帧 t 的前景图。它在第 i 个像素的值设为 $F_i^t = c(c_i^t)$，其中 c_i^t 表示帧 t 第 i 个像素的颜色。对帧 t 候选分段树 $T_j^t \in \hat{T}^t$ 的每个分段 s_{ij}^t，我们将其前景概率计算为被 F^t 中 s_{ij}^t 覆盖的平均数值。如果 T_j^t 的所有分段的前景概率较低，则修剪 T_j^t，否则保留。

1.3　提取分层空间－时间段

在对候选分段树修剪后，我们提取了包含剩余候选分段树的一个集合 $\hat{T}^t$。为了刻画人体或人体部位的时间动态特征，对每个 $T_j^t \in \hat{T}^t$，我们跟踪每个段 $s_{ij}^t \in T_j^t$，以构建一个空间－时间段。

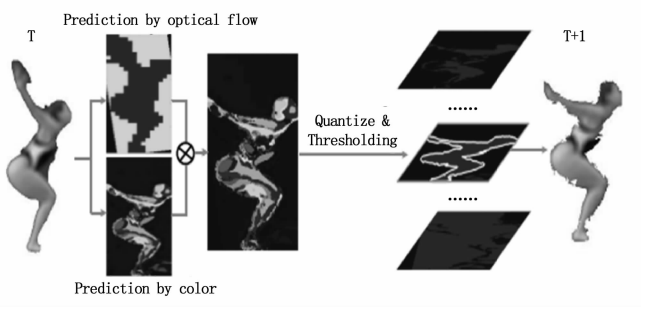

图 2　本文提出的区域跟踪方法

本文提出一种新的非刚性区域跟踪方法（图 2）。设 R 表示当前帧的被跟踪区域。设 $a = (x_a, y_a)$ 表示 R 内点的坐标，$(\Delta x_a, \Delta y_a)$ 表示其相应的中值滤波光流场的流向量。从流向量角度来看，下一帧的预测区域为 $R' = \{(x'_a, y'_a)\} = \{(x_a + \Delta x_a, y_a + \Delta y_a)\}$。设 B 表示边与水平和垂直轴平行的区域 R' 的边界框，设 $\hat{B}$ 表示较长边与 R' 最小惯性轴平行的 R' 的紧固边界矩形。设 h 表示被跟踪的初始段的颜色直方图。然后，我们针对 B 计算流预测图 M_f 和颜色预测图 M_c。假设下一帧的点 $b' \in B$ 的颜色为 $c_{b'}$，则我们设置 $M_c(b') = h(c_{b'})$，且对 $M_f(b')$ 有：

$$M_f(b') = \begin{cases} 2 & b' \in \boldsymbol{R'} \\ 1 & b' \in \hat{B} \wedge b' \notin R' \\ 0 & otherwise \end{cases} \tag{3}$$

我们对两个图进行综合，且 $M(b') = M_f(b') \cdot M_c(b')$。当计算 M_f 时，我们使用 $\hat{B}$ 上的一个网格，于是同一单元内的点将被设为该单元内的最大值。这可以缓解光流场噪声造成的孔洞现象。

2　运动识别和定位

我们使用简单学习方法来训练运动分类器，然后使用学习过的模型来进行运动定位。对每个空间－时间分段，使用空间－时间网格来分割与其轴平行的边界框，然后在每个空间－时间单元内计算特征，将所有单元内的特征连接起来以生成最终特征向量。

本文基于 K 均值聚类方法为根空间－时间段和部位空间－时间面分别生成码本。然后，利用基于其空间－时间段和代码字间的相似性数值的最大汇总方法，并基于词袋表示对每个测试视频进行编码。我们对所有训练视频的词袋表示均训练一对多线性支持向量机，以进行多类运动分类，测试视频的运动标签为：

$$y = \underset{y \in Y}{\operatorname{argmax}} \begin{pmatrix} w_y^r \\ w_y^p \end{pmatrix} (x^r x^p) + b_y \tag{4}$$

其中：x^r 和 x^p 分别是测试视频的根空间－时间段和部位空间－时间段的词袋表示。w_y^r 和 w_y^p 分别是根和部位训练后的分离超平面的元素，b_y 是偏差项，Y 是相关运动分类标签集合。

对运动定位，我们得到测试视频中有效的空间－时间段，并输出包含这些段的轨迹。具体来说，已知一个测试视频作为根空间－时间段 $S^P=\{s_a^r\}$ 的集合以及部位空间－时间段 $S^P=\{s_b^p\}$ 的集合，且 $C^r=\{c_k^r\}$ 和 $C^p=\{c_k^p\}$ 分别表示对应于 w_y^r 和 w_y^p 正值元素的代码字集合。我们将集合 U 计算为：

$$U=\{\hat{s}^r:\hat{s}^r=\underset{s_a^r\in S^r}{\operatorname{argmax}}h(s_a^r,c_k^r),\forall c_k^r\in C^r\}$$
$$\cup\{\hat{s}^p:\hat{s}^p=\underset{s_b^p\in S^p}{\operatorname{argmax}}h(s_b^p,c_k^p),\forall c_k^p\in C^p\} \quad (5)$$

其中：函数 h 衡量两个空间－时间段的相似性，且我们使用这些段的特征向量的直方图交集。然后，我们将包含集合 U 中至少一个空间－时间段的所有轨迹输出为运动定位结果。通过这种方法，虽然 U 中空间－时间段可能只覆盖一个视频帧稀疏集，但是我们的方法可以输出更为密集的定位结果。综上所述，本文提出的人类动作识别和定位算法的总体流程如图 3 所示。

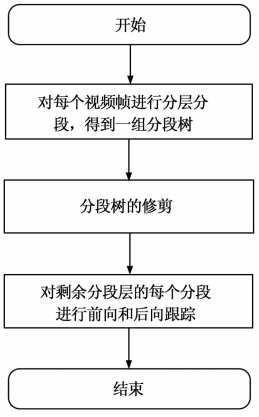

图 3 算法流程图

3 实验

3.1 实验设置

我们在 Matlab2012 上部署本文方法，并采用文献［8－9］中的 UCF－Sports 数据集和 HighFive 数据集做为实验对象。我们采用文献［9］中的训练/测试分配方法。我们的空间－时间网格设置与 UCF－Sports 数据集设置相同，但是为了和其他方法的结果做公平比较，我们对每个空间－时间单元只计算 MBH 特征。我们对根空间－时间段构建一个 1 800 个词的码本，为部位空间－时间段构建一个 3 600 个词的码本。同时，我们在训练时不使用人体边界框注释。

3.2 实验结果

运动识别：表 1 和表 2 分别给出了 UCF－Sports 和 High－Five 数据集的运动检测结果。对 UCF－Sports 数据集，本文方法性能略优于文献［12］中的算法（2.3%），远优于文献［13］中的方法（8.6%）。这是因为文献［12－13］使用的分类器比我们使用的线性支持向量机复杂，更为重要的是，文献［12－13］对每个帧均需要成本较高的人体边界框注释，而本文方法不需要。此外，虽然文献［12］算法的分类性能与本文方法比较接近，但是其算法无法实现有效的运动定位。文献［13］中的算法可以给出定位结果，但是定位性能远低于本文方法（见表 3）。

表 1 UCF－Sports 数据集各类分类精度均值

方法	监督	精度(%)
本文方法(根＋部位)	标签	81.7
本文方法(只有部位)	标签	71.3
文献[12]的方法	标签＋方框	79.4
文献[13]的方法	标签＋方框	73.1

表 2 HighFive 数据集平均精度(*MAP*)

方法	*MAP*(%)
本文方法(根＋部位)	53.3
本文方法(只有部位)	46.3
文献[14]的方法	55.6
文献[5]的方法	53.4
文献[4]的方法	36.9
文献[9]的方法	32.8

对 High Five 数据集，本文方法与文献［4－5，9，14］中的 4 种方法做比较。文献［14］中方法对密集轨迹聚类树使用非线性支持向量机，并可生成最优结果。文献［5］和［4］生成结果时使用的支持向量机，其直方图交集内核分别为密集轨迹包和 STIP 包。文献［9］中的方法使用分层支持向量机。与以上方法相比，本文方法虽然简洁，但是与文献［14］和［5］的性能相当，且远优于文献［4］和［9］中算法的性能。而文献［4－5，9，14］中算法的运动定位性能均不如本文算法。

运动定位：表 3 和表 4 给出了 UCF－Sports 和 HighFive 数据集的运动定位结果。定位分值计算为被测试的视频帧的平均 IOU 值（交集并集比，intersection－over－union）。

表 3 UCF Sports 数据集衡量为 IOU 均值(%)的运动定位结果

	部分帧				所有帧			
	[15]	[16]	[13]	本文	[15]	[16]	[13]	本文
跳水	16.4	36.5	43.4	46.7	22.6	37.0	—	44.3
高尔夫	—	—	37.1	51.3	—	—	—	50.5
踢	—	—	36.08	50.6	—	—	—	48.3
举	—	—	68.8	55.0	—	—	—	51.4
骑	62.2	68.1	21.9	29.5	63.1	64.0	—	30.6
跑	50.2	61.4	20.1	34.3	48.1	61.9	—	33.1
滑	—	—	13.0	40.0	—	—	—	38.5
滑动－b	—	—	32.7	54.8	—	—	—	54.3
滑动－s	—	—	16.4	19.3	—	—	—	20.6
走	—	—	28.3	39.5	—	—	—	39.0
平均	—	—	31.8	42.1	—	—	—	41.0

注：－表示结果无法获得。

对 UCF－Sports 数据集，文献［13］中的方法只能给出部分帧的定位结果，因此我们给出这部分帧的比较结果。对这部分帧，本文方法的平均性能比文献［13］要高出 10.3%。我们也给出了本文对所有帧的性能，性能结果与部分帧的结果类似。文献［15］和［16］中的方法只给出 UCF Sports 数据

（下转第 2177 页）

规范，使各个现存的图像处理功能模块得以最大限度的综合与重复利用。整个平台从逻辑需求出发，严格基于接口进行设计，符合面向对象设计的开闭原则和依赖反转原则，即使更改细节也不会影响到既有插件的运行。整个平台改变了以往系统提供底层格式的惯例，使得现存的图像处理库得到充分再利用。

参考文献：

[1] 孙　浩，陈　安，胡跃明. 基于 DSP 和 FPGA 的通用图像处理平台设计 [J]. 电子设计工程，2009，6 (17)：41 - 43.

[2] 黎松奇. 基于 .Ne t 平台的通用自动测试系统设计 [J]. 自动化与仪器仪表，2011，(5)：45 - 47.

[3] 李　允，罗　蕾，雷昊峰，等. 嵌入式 Java 虚拟机的性能优化技术 [J]. 计算机工程，2004，(18)：46 - 49.

[4] 严永斌，吴健平. 基于 .NET Compact Framework 的移动 GIS 软件开发 [J]. 测绘与空间地理信息，2008，31 (4)：37 - 41.

[5] 王鸣秋. 浅谈数字图形图像存储格式 [J]. 湖南广播电视大学学报，2004，(1)：31 - 36.

[6] 梁　艳. 基于模糊理论的图像分割算法研究 [D]. 西安：西安电子科技大学，2012.

[7] 张　伟. 指纹图像分割算法的研究 [J]. 计算机测量与控制，2011 (7)：1746 - 1748.

[8] 陈　果. 图像阈值分割的 Fisher 准则函数法 [J]. 仪器仪表学报，2003，24 (6)：564 - 567.

[9] 夏红霞，钟　珞. 分而治之算法的并行性 [J]. 计算机应用，1989，4：29 - 31.

(上接第 2150 页)

集 3 种类别的运动定位结果（跑步，俯冲和骑马）。本文方法对俯冲类别的性能优于文献 [15 - 16] 中的方法，但是对其余两种低于文献 [15 - 16] 中的方法。然而，与本文方法相比较的其他各种算法均使用成本很高的人体边界框注释，且它们的学习过程的复杂度远高于本文方法。

对 HighFive 数据集，对注释为包含交集的帧计算 IOU。我们的 IOU 为 30.8%，虽然仍然非常优异，但是低于 UCF Sports 数据集。为验证这一观点，我们也计算了检索率（检索率定义为交集区域与被注释的运动区域之比）。表 4 中检索率较高，证明被注释的运动区域主要由本文方法检测出。文献 [15] 和 [16] 分别给出了接吻类别的运行定位结果，分别为 18.5%和 39.5%。这些结果仍然没有直接可比性，因为文献 [15] 和 [16] 均需要人体边界框，而本文方法需要的监管量很小（只需标签）。

表 4　High Five 数据集衡量为 IOU(%)和检索率(%)均值的运动定位性能

类别	握手	致意	拥抱	接吻	平均
IOU	26.9	32.9	34.2	29.3	30.8
检索率	79.4	88.8	82.6	80.8	82.3

4　结论

本文提出了一种改进的人类动作识别与定位方法，可用于有效检测真实视频中事件发生的可靠性和地点。仿真实验结果表明，本文方法是有效的。我们的下一步工作是基于压缩感知理论，进行视频的异常检测与定位。

参考文献：

[1] 徐光祐，曹媛媛. 动作识别与行为理解综述 [J]. 中国图象图形学报，2009，14 (2)：189 - 195.

[2] 吴伟和，郝爱民，赵永涛，等. 一种人体运动骨骼提取和动画自动生成方法 [J]. 计算机研究与发展，2012，49 (7)：1408 - 1419.

[3] 谭论正，夏利民，黄金霞，等. 基于 pLSA 模型的人体动作识别 [J]. 国防科技大学学报，2013，35 (5)：102 - 108.

[4] Laptev I，Marszalek M，Schmid C，et al. Learning realistic human actions from movies [A]. Computer Vision and Pattern Recognition，2008. CVPR 2008. IEEE Conference on. IEEE [C]，2008：1 - 8.

[5] Wang H，Klaser A，Schmid C，et al. Action recognition by dense trajectories [A]. Computer Vision and Pattern Recognition (CVPR)，2011 IEEE Conference on. IEEE [C]，2011：3169 - 3176.

[6] Poppe R. A survey on vision－based human action recognition [J]. Image and vision computing，2010，28 (6)：976 - 990.

[7] Wang Y，Tran D，Liao Z，et al. Discriminative hierarchical part－based models for human parsing and action recognition [J]. The Journal of Machine Learning Research，2012，13 (1)：3075 - 3102.

[8] Rodriguez M D，Ahmed J，Shah M. Action mach a spatio－temporal maximum average correlation height filter for action recognition [A]. In CVPR [C]，2008：1123 - 1132.

[9] Patron－Perez A，Marszalek M，Zisserman A，et al. High five：Recognising human interactions in tv shows [J]. 2010.

[10] Leordeanu M，Sukthankar R，Sminchisescu C. Efficient closed－form solution to generalized boundary detection [M]. Computer Vision - ECCV 2012. Springer Berlin Heidelberg，2012：516 - 529.

[11] Arbelaez P，Maire M，Fowlkes C，et al. From contours to regions：An empirical evaluation [A]. Computer Vision and Pattern Recognition，CVPR 2009. IEEE Conference on. IEEE [C]，2009：2294 - 2301.

[12] Raptis M，Kokkinos I，Soatto S. Discovering discriminative action parts from mid－level video representations [A]. Computer Vision and Pattern Recognition (CVPR)，2012 IEEE Conference on. IEEE [C]，2012：1242 - 1249.

[13] Lan T，Wang Y，Mori G. Discriminative figure－centric models for joint action localization and recognition [A]. Computer Vision (ICCV)，2011 IEEE International Conference on. IEEE [C]，2011：2003 - 2010.

[14] Gaidon A，Harchaoui Z，Schmid C. Recognizing activities with cluster－trees of tracklets [A]. BMVC [C]，2012.

[15] Tran D，Yuan J. Optimal spatio－temporal path discovery for video event detection [A]. Computer Vision and Pattern Recognition (CVPR)，2011 IEEE Conference on. IEEE [C]，2011：3321 - 3328.

[16] Tran D，Yuan J. Max－margin structured output regression for spatio－temporal action localization [J]. 2012，21 (13)：2013 - 2020.