视频中旋转与尺度不变的人体分割方法

薄一航 [1] HAO Jiang [2]

摘 要 提出了一种旋转与尺度不变的人身体部位所在区域的视频分割方法. 方法中不仅考虑到躯干与四肢之间的关系, 还考虑到四肢之间的相互关系, 通过空间与时间的连续性约束对每帧中各个可能的身体部位进行优化组合, 并巧妙地用动态规划对非线性图模型进行优化, 且不受运动目标尺度变化与各种翻转运动的影响. 该方法首先用动态规划的优化方法得到每一帧中最优的 N 个身体部位组合, 将每一个组合作为图模型中的一个节点, 并用动态规划对所有帧中的各个组合所构成的网格状图结构进行优化, 最终得到每一帧中最优的身体部位组合. 实验结果表明, 该视频分割方法不仅适用于行人视频, 还适用于具有各种姿势的运动视频, 且具有较好的鲁棒性.

关键词 视频分割, 旋转不变, 尺度不变, 动态规划

引用格式 薄一航, HAO Jiang. 视频中旋转与尺度不变的人体分割方法. 自动化学报, 2017, **43**(10): 1799−1809

DOI 10.16383/j.aas.2017.c150841

A Rotation- and Scale-invariant Human Parts Segmentation in Videos

BO Yi-Hang[1] HAO Jiang[2]

Abstract This paper proposes a rotation- and scale-invariant method for human body parts segmentation. The proposed method considers not only the relationship between torso and limbs but also between limbs. The method optimizes the candidate assembly of body parts in each frame with spatial and temporal constraints, and uses dynamic programming to optimize a non-linear graph model smartly, which is rotation and scale invariant. First, it generates the best N combinations of human body parts using dynamic programming, each being a node in the graph. Then it optimizes the graph which is the grid made up of all the nodes in each frame, using dynamic programming to get the optimal human body part combination in each frame. Experiments show that this method can robustly get efficient and accurate results both on pedestrian videos and sports videos with any human poses.

Key words Video segmentation, rotation invariant, scale invariant, dynamic programming

Citation Bo Yi-Hang, Hao Jiang. A rotation- and scale-invariant human parts segmentation in videos. *Acta Automatica Sinica*, 2017, **43**(10): 1799−1809

视频分割问题是当前计算机视觉领域一个比较热门的话题. 与静态图像分割方法不同的是视频分割不仅要考虑到单视频帧内各个像素点或超像素块之间的关系, 还要保证相邻视频帧之间对应像素点或超像素块的连续性与光滑性. 视频分割的结果可以为更高一级的视频及视频中目标的分析工作提供较好的分析基础.

起初, 针对静止摄像机拍摄的视频, 即视频背景为静止不变的情况, 可以通过简单的去背景的方法得到整个运动的前景区域[1−4]. 从目前的视频分割方法来看, 包括基于像素点的分割、基于超像素块的分割和基于提议 (Proposals) 的分割等. 但是, 对于视频分割而言, 考虑到运算量和运算速度的问题, 基于像素点的分割方法很不现实, 也很少被采用. 当前比较流行的视频分割方法以基于超像素块的分割和基于提议的分割为主. 首先, 对基于超像素块的分割而言, 研究者们试图通过区域块跟踪的方法来处理[5−8] 得到不同的分割区域. 鉴于视频数据本身的特殊性, 还有一些视频分割方法将视频分割成底层特征随时间变化连续的超像素块[7−10]. 然而, 超像素块本身往往不具备完整的语义信息, 每个超像素块可能是一个完整的目标, 也可能是构成某个目标的一部分, 这样的分割结果并不利于进一步的目标分析工作. 并且, 分割结果的优劣很大程度上还依赖于所选择的分割阈值, 我们通常很难选择一个合适的阈值使得每一个分割区域都是一个完整且有意义的目标或目标的组成部分. 另外, 对于比较长的视

收稿日期 2015-12-14 录用日期 2016-10-26
Manuscript received December 14, 2015; accepted October 26, 2016
北京市教委科研计划一般项目 — 目标跟踪与分割算法在电影抠像中的应用与研究 (KM201710050001) 资助
Supported by Beijing Municipal Education Commission, General Plan of Scientific Research Plan-Application and Research on Object Tracking and Segmentation in Film Keying (KM201710050001)
本文责任编委 桑农
Recommended by Associate Editor SANG Nong
1. 北京电影学院美术学院 北京 100088 中国 2. 波士顿学院计算机科学系 波士顿 02467 美国
1. Fine Art Department, Beijing Film Academy, Beijing 100088, China 2. Computer Science Department, Boston College, Boston 02467, USA

频而言, 在整个视频分割的过程中, 会出现前后帧相对应的分割区域错位的情况. 近几年, 还有研究者提出针对视频中运动目标的分割方法[11−12], 比如文献 [13] 中用一种全自动的方法, 通过将 Grab-Cut 方法[14] 扩展到时空领域来得到视频中目标的闭合轮廓. 为了得到更有意义的分割结果[15−16], 基于提议 (Proposals) 的视频分割方法越来越受到研究者们的青睐[9, 17−20], 每一个提议都极有可能是一个有意义的目标或目标的某个组成部分. 其中, 文献 [21] 通过 SVM (Support vector machine) 分类器提取出每个视频帧中较优的一些提议, 再通过求解一个全连接的条件随机场的最大后验对前景和背景进行分类, 得到的前景区域往往是一个完整的、有意义的目标所在的区域. 文献 [22] 利用特征空间优化的方法将视频进行语义分割, 得到视频中各个语义目标所在的区域. 文献 [23] 借助目标检测以及目标跟踪的结果对视频中的目标进行分割.

然而, 这些视频分割方法得到的是整个前景目标所在的区域[24−25], 未能细化到构成目标的每一个组成部分. 如果要进一步对运动目标的姿势等进行识别与分析, 仅仅得到整个目标所在的区域是远远不够的, 因此, 与上述方法不同, 本文所提出的视频分割方法可以具体到构成运动目标的每个主要部位.

在各类运动目标中, 人是最普遍, 也是最复杂的一种. 与其他刚性物体不同, 由于人姿势变化的不确定性和无规律性, 其旋转、尺度以及外貌的变化都会给分割过程带来很大的困难. 目前, 已有不少关于人身体各部位的跟踪与检测方法, 将人的身体分成若干个运动部位, 如图 1 (a) 所示, 不同的部位由不同灰度的矩形框来标定, 而非具体的身体部位所在的区域. 此类方法通常是基于模板的匹配, 根据人姿势、尺度的变化, 分别与各个角度和尺度的模板进行匹配, 从而得到与测试图像最为接近的一个模板作为匹配结果, 称这种方法为 “图案结构 (Pictorial structure)”[26−27]. 该方法的模型为树形结构, 只考虑到四肢与躯干之间的关系, 而没有对四肢之间的关系加以约束, 往往会引起某一只胳膊或者某一只腿的漏检或错检. 另外, 该方法虽然已被广泛地应用到人的跟踪与姿势的估计中, 但是, 由于人运动姿势变化的随机性和不可预知性, 无法事先知道目标尺度和旋转角度的变化范围, 逐一模板匹配的过程会很大程度地影响运算速度.

针对以上问题, 本文提出一种旋转与尺度不变的运动视频中人身体部位所在区域的分割方法, 如图 1 (b) 所示为单帧的分割标注结果. 该方法不仅考虑到躯干与四肢之间的关系, 同时还考虑到四肢之间的相互关系. 其最大的优势就在于, 它不需要考虑不同尺度与旋转角度的模板匹配, 而是利用人体各个部位的相对面积及比例关系, 构建一个旋转与尺度不变的视频分割方法. 实验结果表明, 该方法比 “图案结构” 方法的鲁棒性更强, 尤其是对于目标旋转和尺度变化较大的视频, 并与现有的 “图案结构” 方法进行了定性和定量的比较. 这样的分割结果无论是在体育赛场、舞蹈演出, 还是在视频监控系统中都具有重要的应用潜质.

(a) “图案结构” 检测结果
(a) Detection result of “pictorial structure” method

(b) 该方法分割结果
(b) Segmentation result of proposed method

图 1 “图案结构” 检测结果与本方法分割结果图

Fig. 1 Detection result of “pictorial structure” method and the segmentation result of proposed method

本文最大的贡献就是提出了一种新的旋转与尺度不变的人身体各部位所在区域的视频分割方法. 如图 2 所示为整个方法的鸟瞰图, 首先, 找到每一帧 (Frame 1, Frame 2, $\cdots$, Frame n) 中可能的身体部位所在的区域块; 然后, 根据每帧内各个身体部位间的相对位置、大小、对称性等约束找到每一帧中可能的身体部位组合; 最后, 利用相邻帧之间运动的连续性、光滑性等约束条件, 采用动态规划的方法找到每一帧中最优的人身体部位的组合. 该方法不仅适用于行人视频, 同样也适用于复杂的运动视频.

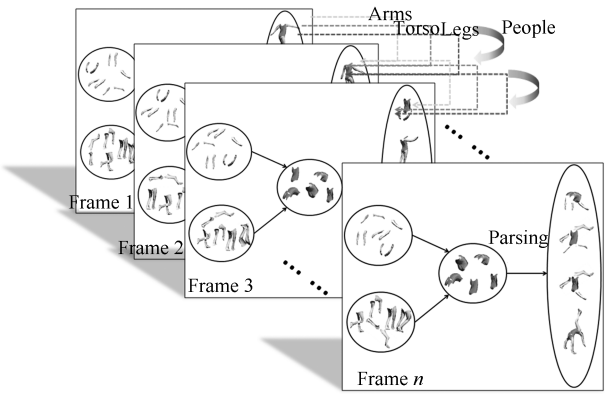


图 2 旋转与尺度不变的视频分割方法鸟瞰图

Fig. 2 The bird-view of rotation and scale invariant video segmentation method

1 方法

本文提出的视频分割方法旨在分割出视频中人身体各部位所在的区域. 该方法根据人体各部分组成结构之间空间与时间的连续性, 对可能的人体部

位组成结构进行优化选择. 为了使得分割结果不受目标运动过程中旋转以及尺度变化的影响, 人体部位组成结构的图模型应为一个环状结构, 也就是说, 不仅要考虑躯干与四肢之间的关系, 还要考虑四肢之间的关系. 如何有效地对该环状结构进行优化具有一定的挑战性. 本文提出一种生成最优的 N 个人体部位组合的方法, 每一帧中所有人体部位之间形成一个环状的图结构, 分别找到每一帧中最佳的 N 个人体部位组合, 根据帧与帧之间每个身体部位以及整个人运动的连续性和光滑性, 采用动态规划的优化方法找到每一帧中最优的一组人体部位组合, 从而巧妙地解决了该非树形结构的优化问题.

1.1　能量函数

本方法所采用的人体部位组成结构主要包括 5 个身体部位: 躯干 (Torso)、左右胳膊 (Arm1, Arm2) 和左右腿 (Leg1, Leg2), 由于头的位置可以简单地通过两只胳膊和躯干的位置检测到, 考虑到模型的简洁性, 该方法没有包括头部. 每帧内各个身体部位之间的结构关系以及相邻帧间相应身体部位之间位移、形状变化的关系, 如图 3 所示, 图中每个节点表示一个身体部位, 每条边表示它所连接的两个身体部位之间的关系. 其中, 虚线边代表单帧内身体各部位之间的关系, 实线边代表相邻帧之间各部位之间的关系, 每个点线方框代表一个视频帧. 这里, 不仅考虑到躯干–胳膊、躯干–腿、胳膊–胳膊、腿–腿之间的关系, 还考虑到胳膊–腿之间的关系. 并且对于相邻的前后帧之间, 身体各个部位以及整个身体的连续性和一致性也是必须要考虑的.

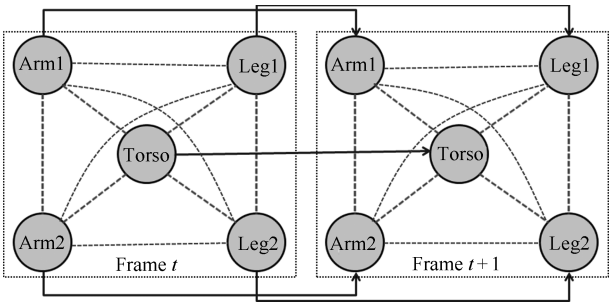


图 3　单帧内与相邻帧之间身体部位关系图

Fig. 3　Human body parts relationships in single frame and between adjacent frames

该方法把身体部位所在区域的视频分割转化成一个图模型的优化问题, 即把每一个身体部位分配给图模型中的一个节点, 通过优化过程使得分配的花费最小. 这里, 可能的身体部位所在的区域由文献 [28] 所提出的方法得到. 该方法可得到一系列与目标类无关的提议 (Proposals). 这些提议都具有较高的属于某个目标类的分值, 也就是说, 这些通过合并超像素块得到的提议很有可能是一个有意义的目标. 这也是提议比普通超像素块的优势所在. 另外, 通过分割算法得到的超像素块很容易将具有相同表观特征的不同目标划分为同一个区域, 而提议则可以在很大程度上避免这种错误的产生. 本方法将最有可能属于身体部位的提议集合起来构成可能的身体部位的组合.

$$E(f)=\underbrace{\min\left(\sum_{k=1}^{m}(\alpha P(f_k)+\beta G(f_k)+\gamma O(f_k)+\delta A(f_k))\right)}_{\text{Intra-frameenergy}}+\underbrace{\sum_{k=2}^{m}\left(\eta S(f_k,f_{k-1})+\phi L(f_k,f_{k-1})+\theta H(f_k,f_{k-1})\right)}_{\text{Inter-frameenergy}} \tag{1}$$

如式 (1) 所示, 同时考虑到帧内与帧间的连续性与一致性, 能量函数 $E(f)$ 包括帧内能量 (Intra-frame energy) 和帧间能量 (Inter-frame energy) 两大部分, 其中帧内能量主要包括身体部位的形状匹配花费 $P(f_k)$、身体部位之间的距离 $G(f_k)$、身体部位之间的重叠 $O(f_k)$、身体部位之间的面积比例 $A(f_k)$ 等, 身体部位的形状越接近真实形状, $P(f_k)$ 就越小; 身体部位之间的距离和重叠区域越小, $G(f_k)$ 和 $O(f_k)$ 就越小; 身体部位之间的面积比越接近真实比例, $A(f_k)$ 就会越小. 而帧间能量主要包括身体部位以及整个目标形状的连续性 $S(f_k,f_{k-1})$、位置的连续性 $L(f_k,f_{k-1})$ 以及颜色的连续性 $H(f_k,f_{k-1})$, 帧与帧之间身体各部位以及整个目标的形状变化越小、位移越小以及颜色的改变越小, $S(f_k,f_{k-1})$、$L(f_k,f_{k-1})$ 和 $H(f_k,f_{k-1})$ 就会越小. 系数 α、β、γ、δ、η、ϕ 和 θ 为控制各分项比重的常量系数.

1.1.1　身体部位形状匹配花费 (P)

首先通过文献 [28] 中所提出的方法得到各个候选区域块. 每一个候选区域块为一个可能的身体部位, 即一个提议. 每一个身体部位, 比如躯干、胳膊等, 均具有一组模板. 通过度量候选区域与模板之间所对应形状描述子[29] 的欧氏距离来衡量候选区域的形状与真实身体部位形状的相似性. 区域的形状描述子定义为区域内部任意点对之间的距离直方图. 当计算这个直方图时, 用区域内所有点对距离的最大值对其进行归一化处理. 该形状描述子是旋转与尺度不变的, 即不随区域旋转和尺度的变化而变化的. 具体的身体部位形状匹配花费 P 定义为

$$P(f_k)=\sum_i c(i,f_k(i)) \tag{2}$$

其中, i 表示各个身体部位的索引值, $f_k(i)$ 为身体

部位 i 的候选区域, $c(i, f_k(i))$ 为分配候选区域 $f_k(i)$ 给身体部位 i 的花费. c 为区域 $f_k(i)$ 的形状描述子与身体部位 i 的模板之间的最短距离. 为了减少候选区域的个数, 提高运算速度, 实验过程中用 RANSAC (Random sample consensus) 方法去掉背景部分. 即取先前若干帧和未来若干帧, 比较它们的 SIFT (Scale-invariant feature transform) 特征, 由于前景目标往往只占每一帧的一小部分区域, 因此, 前景目标上的 SIFT 特征点在 RANSAC 特征匹配中成为野点. 匹配过程中, 只匹配背景点, 将当前帧与其前后帧相减并求均值, 得到一个估计的背景, 从而可得到大致的前景区域. 当然, 由于受到光照变化、摄像机抖动等外界条件的影响, 视频的背景并非完全静止, 也就是说, 这种去背景的方法并不能保证去掉所有的背景部分. 需要说明的是去背景的过程是可选的, 并不会影响最终的分割结果.

1.1.2 身体部位间的距离 ($\boldsymbol{G}$)

除了保证每一个身体部位所在的区域有正确的形状之外, 还要确保躯干与四肢之间的距离足够小, 也就是说, 所有的躯干和四肢之间是连接的, 而不是离散的. 设 t 为躯干的索引值, j 为四肢的索引值. 计算四肢 j 与躯干之间的最小边界距离 $d(f_k(j), f_k(t))$, 那么身体部位之间的距离则表示为

$$G(f_k) = \sum_{j \in L} d(f_k(j), f_k(t)) \tag{3}$$

其中, L 为四肢的集合.

1.1.3 身体部位间的重叠 ($\boldsymbol{O}$)

将身体部位之间的重叠 O 作为惩罚项, 使得各个身体部位之间尽量的展开, 又不会排斥部位之间的重叠, 比如, 我们允许胳膊和躯干之间的重叠, 而当有展开的胳膊和躯干存在时, 会优先选择身体部位展开的情况:

$$O(f_k) = \sum_{\{i,j\} \in N} \frac{A(F_k(i) \cap F_k(j))}{A(F_k(i) \cup F_k(j))} \tag{4}$$

其中, $F_k(i)$ 为第 k 帧内部位 i 的估计区域, N 为身体部位对的集合, 包括胳膊–胳膊, 腿–腿, 胳膊–躯干, 腿–躯干, 胳膊–腿等部位对, 函数 A 给出了区域的面积.

1.1.4 身体部位间的面积比 ($\boldsymbol{A}$)

不同的身体部位, 比如胳膊和腿, 可能会具有相似的形状描述子. 因此, 仅通过形状描述子进行约束是不够的, 模型需要更有力的条件来对其进行约束. 进一步讲, 尽管不同的部位可能具有相似的形状, 但不同部位的面积比例往往不同且有一定的规律, 是服从高斯分布的, 高斯分布的参数可由训练样本得到:

$$A(f_k) = \sum_{i \in P} \sum_{j \in P} \frac{(r(f_k(i), f_k(j)) - \mu_{i,j})^2}{\sigma_{i,j}^2} \tag{5}$$

其中, $r(f_k(i), f_k(j))$ 为部位 i 的候选区域 $f_k(i)$ 与部位 j 的候选区域 $f_k(j)$ 的面积比, $\mu_{i,j}$ 和 $\sigma_{i,j}^2$ 分别为高斯分布的均值与方差. P 为身体部位的集合.

除了帧内身体部位的位置比例关系之外, 为了进一步保证运动的光滑性, 还需要进一步考虑相邻帧之间目标的连续性. 这里由以下特征来衡量目标在时间上的连续性.

1.1.5 相邻帧间形状连续性 ($\boldsymbol{S}$)

通常情况下, 相邻帧之间目标的形状变化往往不大, 而且不会发生快速的变化. 这样一来, 目标所在区域轮廓的变化也是光滑的. 模型通过衡量身体部位所在区域轮廓变化的光滑性来判断目标形状的连续性 S. 这里, 区域的形状用其边界的朝向直方图[30] 来表示. 需要说明的是, 这里用朝向直方图而没有用内部距离的原因是不需要保证帧与帧之间目标形状的旋转和尺度不变性, 朝向直方图更适合此种类型的形状匹配.

设 $s_{f_k(i)}$ 为第 k 帧内第 i 个身体部位候选区域 $f_k(i)$ 的形状描述子, s_{f_k} 表示第 k 帧内整个前景目标区域的形状描述子, 即其包括了所有的身体部位. 形状的连续性特征表示为

$$S(f_k, f_{k-1}) = \sum_{i \in P} \| s_{f_k(i)} - s_{f_{k-1}(i)} \| + \| s_{f_k} - s_{f_{k-1}} \| \tag{6}$$

注意, 边界朝向直方图没有进行归一化处理, 而且它还包含有区域的大小信息. 通过最小化 S, 可以保证多个视频帧之间所估计目标的形状和大小的连续性.

1.1.6 相邻帧间位置连续性 ($\boldsymbol{L}$)

与形状的连续性类似, 同样要求帧与帧之间身体部位的位置不会发生突然的变化. 相邻帧之间每个身体部位的位置变化用该部位所在区域中心点的位移来表示. 设 $l_{f_k(i)}$ 为第 k 帧内第 i 个身体部位的候选区域 $f_k(i)$ 的中心位置, 那么该部位位置变化则定义为

$$L(f_k, f_{k-1}) = \sum_{i \in P} \| l_{f_k(i)} - l_{f_{k-1}(i)} \| \tag{7}$$

1.1.7 相邻帧间颜色连续性 ($\boldsymbol{H}$)

假设目标的外貌在连续的相邻帧中不会发生突然的变化. 颜色的连续性可以保证身体部位的颜色在连续帧中的稳定性. 这里, 我们用 RGB 直方图来

量化人身体部位的颜色. 颜色选项定义为

$$H(f_k, f_{k-1}) = \sum_{i \in P} \| h_{f_k(i)} - h_{f_{k-1}(i)} \| \quad (8)$$

其中, $h_{f_k(i)}$ 为第 k 帧中第 i 个身体部位候选区域的颜色直方图.

通过整合这些特征选项, 可以得到一个完整的能量函数. 能量函数的最小化可以保证在每一帧内得到一组最优的身体部位组合. 这里所提出的模型是非树形的, 因此, 我们没办法用动态规划直接对能量函数进行优化. 另外, 由于无法估算候选区域的个数, 因此无法直接使用贪婪的搜索算法. 下一节将提出一种巧妙地将非树形结构转化为树形结构的方法, 从而能够直接用动态规划的方法进行能量函数的优化.

1.2　优化过程

1.2.1　单帧内最优 N 个身体部位组合优化过程

对于视频中的每一帧, 都会产生若干个可能的身体部位组合, 组合的数量是整个优化过程中必须要考虑的问题, 而且每帧中可能组合的数目也是无法事先预知和估算的. 如果不对可能的组合进行筛选, 优化运算的时间复杂度会成倍增加. 因此, 我们需要一种有效地提取每一帧中最优的 N 个身体部位组合的方法, 其中 N 是动态规划算法中所能驾驭的相对最小值.

本方法最大的创新之处就在于, 在处理人体各个部位的关系时, 不仅同文献 [31] 一样要考虑躯干与四肢之间的关系, 还要考虑到四肢之间的关系, 这就使原本的线性结构变成了非线性结构, 从而也增加了选取最优身体部位组合优化过程的难度. 下面来分析一下身体各个部位之间的关系. 如果我们把两个胳膊看作同一个节点, 两条腿看作同一个节点, 那么躯干、胳膊和腿之间的关系就如图 4 (a) 所示, 为一个环状结构. 对躯干进行复制并将其分开, 即有两个相同但不相连的躯干, 那么图 4 (a) 中的图模型就转变为图 4 (b) 中所示的链状结构, 如此一来, 便可以直接用动态规划来对其进行优化, 即如图 4 (c) 所示, 左右两个躯干为同一个躯干, 每次固定一个候选躯干, 然后用标准的动态规划优化算法选出对于每一个候选躯干最优的胳膊和腿的组合. 而对于所有可能的躯干, 把每个躯干得到的身体部位组合进行优劣排序, 最终保留最优的 N 个组合. 此时, 对于视频中的每一帧, 可以分别得到 N 个最优的身体部位组合.

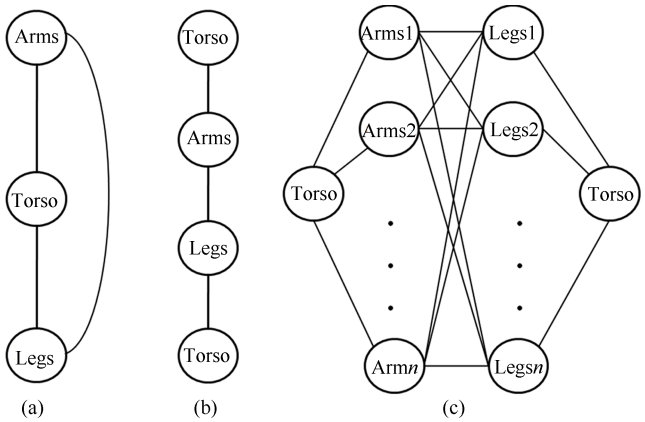


图 4　身体部位关系解析图

Fig. 4　The relationship of human body parts

1.2.2　相邻帧间最优身体部位组合优化过程

根据式 (1) 中的能量函数以及图 3 中所示的图模型可以看出, 除了要考虑单帧内每一对身体部位之间的相关性及位置关系, 还要考虑相邻帧之间对应身体部位之间的连续性与光滑性. 图 3 给出了该方法的图模型, 为一个非树形结构, 我们无法直接用线性的优化方法对其进行优化. 而在第 1.2.1 节中, 每一帧已经产生出了最优的 N 个身体部位组合, 这里, 把每帧中的每一个身体部位组合作为图中的一个节点, 即把图 3 中的每一个子图作为一个节点, 把相邻帧中的各个节点用边连接起来, 这些边和节点就会构成一个网格状的图结构, 每个节点的花费由帧内能量函数 (如式 (1) 中的 Intra-frame energy) 决定, 每条边上的花费由帧间的能量函数 (如式 (1) 中的 Inter-frame energy) 决定. 找到一条使得节点花费 (帧内能量) 和边缘花费 (帧间能量) 均最小的路径, 路径上所有的节点即为我们想要找的每一帧中最优的身体部位组合. 这条最优路径通过动态规划的优化方法得到. 假设每一帧中有 N 个可能的身体部位组合, 视频共有 M 帧, 那么该优化过程的时间复杂度为 $\mathrm{O}(M \times N)$.

2　实验

实验中, 我们把该方法应用到颇具挑战性的各种运动视频序列中, 其中包括复杂的人体姿势和各种翻转动作. 前四个视频 (Video 1, Video 2, $\cdots$, Video 4) 取自 Youtube 视频, 最后一个视频 (Video 5) 取自 HumanEVA 数据库[32]. 下面, 分别给出定性的和定量的实验结果与分析, 以及该模型应用在行人姿势估计上的结果. 实验中, 能量函数里控制各分项比重的系数根据不同视频的具体情况分别设定. 下面, 对能量函数中各个参数的设置做出具体解释和分析. 由于人各种姿势的不同特征, 在考虑各个身体部位之间的关系时应根据不同动作和姿势下各个

部位之间的不同关系和规律, 具体问题具体分析. 式 (1) 所示的能量函数中, Intra-frame energy 的各项在整个能量函数中所起的作用大小各不相同, 比如, 在 Video 1 ∼ Video 4 中, 运动目标均完成了翻转或者平转等动作, 此时胳膊和腿的形状会发生较大的变化, 因此, 这种情况下形状匹配花费 P 就会被设置较小的比重. 而在 Video 5 中, 包含了行人行走的各个朝向, 此时胳膊与躯干之间总会处于相互重叠的状态, 那么在这种情况下, 身体部位间的重叠项 O 就会被设置较小的比重. 而对于 Inter-frame energy 中的各项, 帧与帧之间目标形状、位置以及颜色的连续性均不会受到运动目标姿势的影响, 因此, 对于所有的测试视频, 这其中各项都会设置为相同的比重系数. 对于 N 的选择, 无论是在选取单帧中最优的 N 个身体部位组合时, 还是在选择每一帧中最优的那一组身体部位组合, 都使用的是动态规划的优化方法. 能量分值最小的未必是最优的那一个组合, 因此实验中会选择多个可能的身体部位及其组合参与优化过程. 然而每个阶段的节点数目过大会影响到优化速度, 但如果 N 值选的太小 (小于 10) 运算结果的准确性又会受到一定程度的影响. 经过反复实验, 我们选择了一个既不会对运算速度有太大影响, 又不会降低运算结果准确度的 N 值, 这里设置 N 为 100.

2.1 定性实验结果

我们用文献 [28] 提出的区域提取方法得到各个可能的候选身体部位所在的区域. 用第 1.1.1 节中提到的 RANSAC 方法进行去背景处理, 由于受到光照、摄像机抖动等因素的影响, 视频的背景并非完全静止不动, 因此, 这个方法不能去掉所有的背景区域, 而且, 目标的影子会随目标的运动而运动 (本方法中, 前景目标的影子也被视为背景) 也不能被去除, 换句话说, RANSAC 方法只能去掉完全静止不动的背景区域. 举两个比较典型的去背景后的例子, 如图 5 所示, 第一个例子中 (图 5 中第一行), 由于摄像机的抖动, 发生抖动的背景区域并不能被去掉, 而第二个例子中 (图 5 中第二行), 人的影子随人的运动而运动, 也被误认为是前景部分. 需要说明的是, 去背景与否并不会影响到我们最终的实验结果. 部分去背景虽然减少了大部分的背景噪音, 但是我们仍可以得到一个相对比较干净的前景区域, 这对于提高检测各个身体部位的运算速度有很大的帮助, 但是诸如影子等无法被去掉的背景噪音对我们的检测也是一个非常大的挑战. 图 6 给出了分别在 5 段视频上的分割结果, 包括了不同的运动姿势, 比如, 跳、翻转、倒立、平转以及正常行走等. 所给出的帧均等间距的采样于整个视频. 从分割结果中可以看出, 即使是在比较有挑战性的、姿势变化较大的运动视频上, 该模型也可以得到不错的分割结果.

图 5 去背景后效果图

Fig. 5 Results after background removed

当然, 从实验结果中我们也可以看出, 最终视频分割结果的好坏很大程度上还依赖于提议 (Proposals) 检测的准确与否. 比如, 图 6 中第 6 行第 3 列 Video 3 中的分割结果, 胳膊与躯干被同时检测为躯干, 此时头部则被误认为是胳膊, 同样, 图 6 中第 8 行第 3 列 Video 4 中的分割结果也是如此. 这也是接下来的工作中需要改进和增强之处.

我们也与目前较新的类似的视频分割方法做了定性的对比与分析. 大部分的视频分割方法[33] 基于视频帧图像的底层特征将视频分割成时间上连续的立体超像素块 (Supervoxel), 没有考虑视频中前景目标的语义信息以及上下文关系, 并且, 其分割结果在很大程度上依赖于分割阈值的大小, 阈值选的越大, 分割结果越细; 相反, 分割结果会越粗. 文献 [34] 所提出的基于时空特性的前景目标提议的检测方法把 2D 的目标提议检测方法扩展到具有时间连续性的视频数据中, 从而得到立体的超像素块, 可以正确地检测出视频中的前景目标. 该方法利用颜色[35]、光流[36] 等特征, 以及时间的连续性, 光流梯度和边缘在相邻帧间的位移等信息对视频进行分层分割, 如图 7 中第 2 行至第 6 行所示, 为不同分割阈值下的分割结果, 从上到下分割阈值依次增大. 对这些在不同阈值下得到的分割结果进行合并聚类, 进而得到较为理想的目标所在的区域, 如图 7 中第 7 行所示 (图 7 中所示为去背景后的结果). 由于测试视频背景为静止状态, 因此, 分割和检测结果不受是否进行去背景操作的影响. 然而, 该方法并未考虑前景目标本身各个组成部分的结构和比例关系, 如图 7 中第 7 行的结果所示, 无法解决影子对前景目标检测分割结果的影响, 图 7 第 8 行为本文的分割结果. 另外, 该方法并没有对目标的各个组成部分所在的区域进行语义标注, 因此, 实验中并未与本文的方法进行定量的比较.

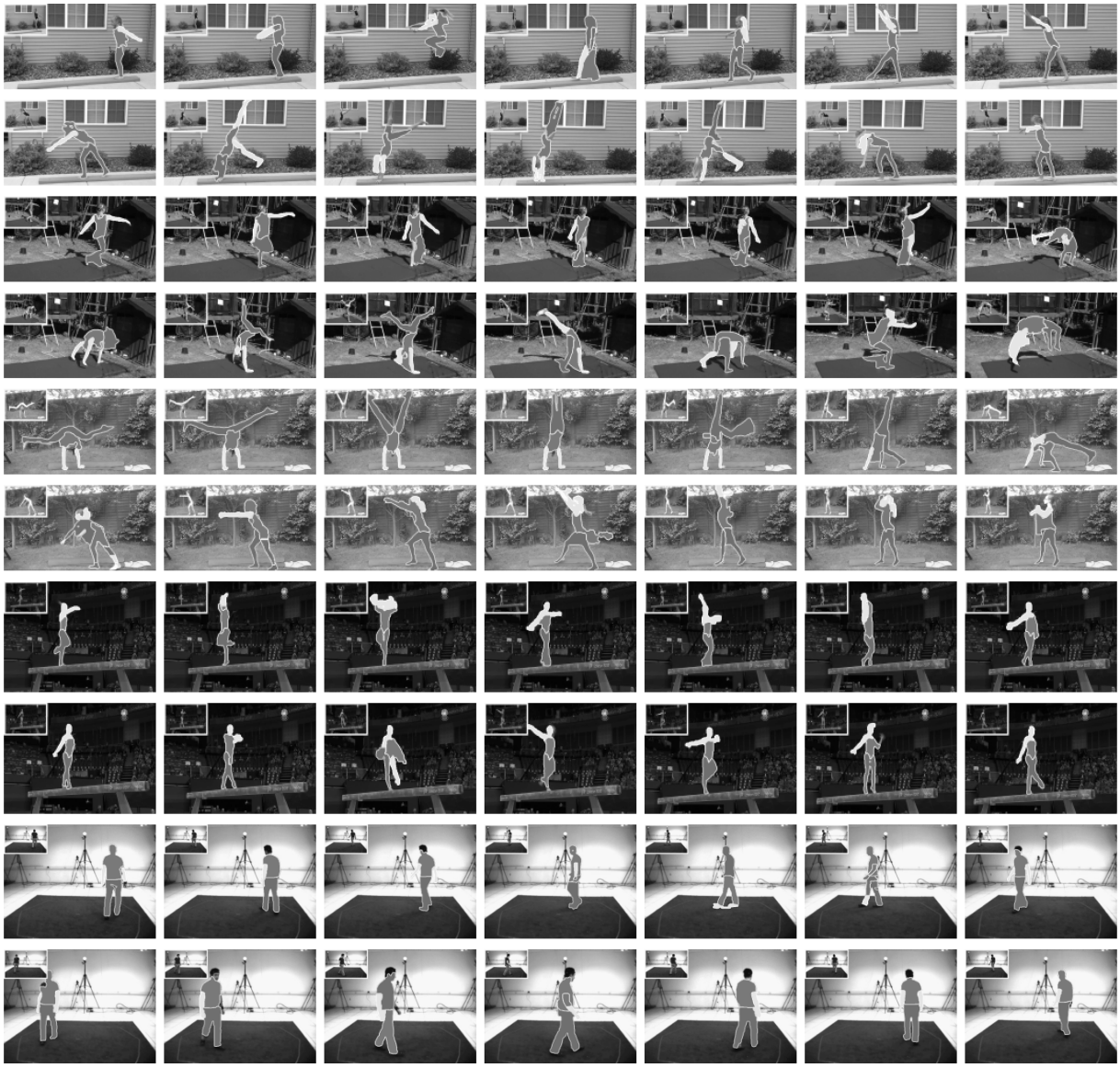


图 6　本方法在 5 段测试视频上的部分分割结果

Fig. 6　Sample results of proposed methods on five test videos

2.2　定量实验结果

该实验把本文所提出的方法与文献 [31] 中提出的 nbest 的方法进行定量的比较分析, 即分别把该方法得到的分割结果和 nbest 方法得到的结果与 Ground truth (GT), 也就是手工标注的真实的身体部位所在的区域相比较.

nbest[31] 方法利用构成人体各个部位之间的“图案结构”对于人体的各个组成部分进行检测, 该结构最大的问题就是只考虑到了躯干与四肢之间的位置关系, 而忽略了四肢之间的关系, 因此, 对于直立状态的人体而言, 该方法可以得到较好的检测结果, 而对于发生旋转的、非直立状态的人体而言, 该方法很难奏效. 如图 8 所示, 为 nbest 方法对非直立姿势的人体的检测结果, 图中第 1 列为原始视频帧, 第 2 列为 nbest 方法的检测结果, 不同颜色的矩形框表示不同的身体部位, 第 3 列为本文所提出的方法的检测结果.

为了公平起见, 实验中同样对 nbest 方法的输入数据也进行去背景操作. 另外, 我们的方法得到的是分割的区域, 而 nbest 方法得到的是每个身体部位区域所在的矩形绑定框, 因此, 我们按照一定的合适的比例扩张 nbest 方法得到的矩形区域的中轴线, 使矩形区域腐蚀为一定比例的圆柱形区域, 让这个圆柱形区域无限地接近身体部位所在的分割区域. 由于 nbest 方法[31] 不是尺度和旋转不变的, 它对于翻转幅度比较大的情况得到的实验结果会很差. 而

本文提出的方法恰恰克服了这一点, 不论目标发生如何旋转和尺度的变化, 均可以得到可靠的分割结果.

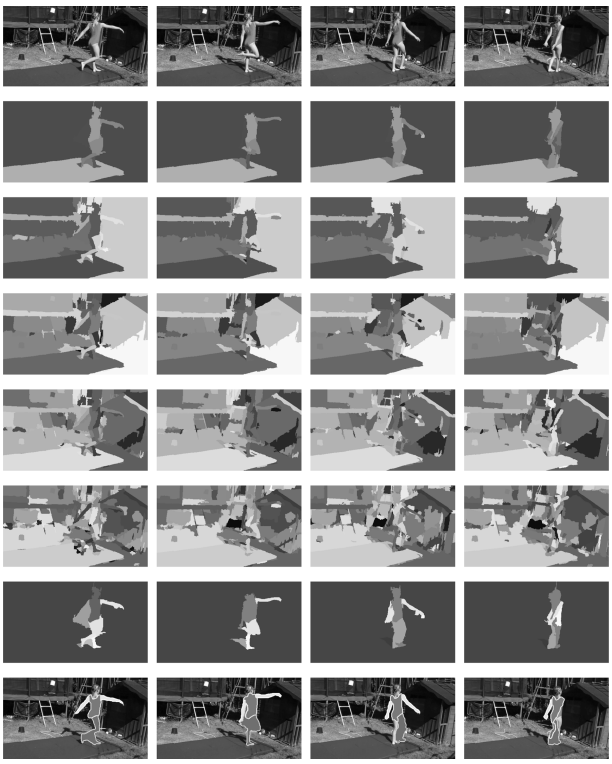


图 7 文献 [31] 的方法与本方法测试结果对比示例

Fig. 7 Example results of the method in [31] and proposed method

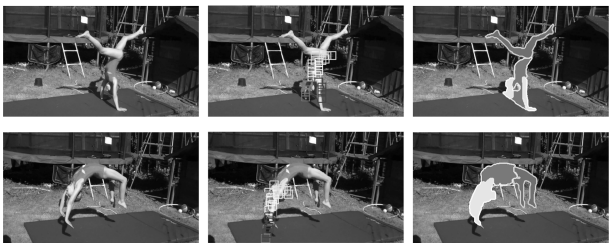


图 8 nbest 方法检测结果与本方法结果示例

Fig. 8 Example results of nbest method and proposed method

对于每一个身体部位所在的区域, 这里定义了一个匹配分值, $A(P \cap G) \setminus A(P \cup G)$, 其中, P 是分割得到的身体部位所在的区域, G 为对应的真实身体部位所在的区域, A 为区域的面积函数. 表 1 中给出了本方法与文献 [31] 所提出的 nbest 方法对相同视频检测结果的比较分值. 无论哪种运动情况, 该方法的结果均比 nbest 方法要改进和提升很多. 对于整体的平均检测和分割结果, 我们的方法依旧要优于所比较的方法.

图 9 给出了本文提出的方法与 nbest 方法实验结果的正确率曲线, 其中包括单个身体部位以及整个人体的正确率. 每条检测曲线都给出了所检测到的高于某一阈值的正确的身体部位占整个检测结果的比例. 比该阈值高的均认为是正确的检测结果. 并且, 当阈值为 1 时, 检测结果的正确率为 0, 而阈值为 0 时, 检测结果正确率为 1. 从图 9 的正确率曲线不难看出, 该方法得到结果的正确率明显高于 nbest 方法.

2.3 行人姿势估计的应用

由于该方法分割结果的特殊性, 以及行人正常行走姿势的规律性, 可将其应用到行人的姿势估计上. 分割结果可分为上身和下身两部分, 躯干与胳膊属于上身, 腿属于下身. 根据直立行走的行人身体各个部位的比例位置关系, 可以找到行人身体上可能的各个关节点, 比如, 肩膀、肘部、手腕、臀部、膝盖和脚踝等. 然后, 用扩展动态规划 (Extended dynamic programming) 的方法求得各个最优的关节点, 从而得到行人的姿势.

这里, 每一对相邻的关节点被看作是动态规划中的一个状态. 所用到的各种约束条件包括两相邻关节点之间距离与行人高度比、两相邻状态之间的内夹角, 以及两相邻状态连线与对应身体部位所在区域轮廓之间的平行性. 另外, 还需要考虑当前状态与先前状态的连续性和上身关节点与下身关节点的对齐, 进而估计出不同朝向行人的关节点, 用大小不同的原点表示关节点, 关节点越大表示其离摄像头

表 1 该方法和 nbest 方法分别与 GT 的比较结果

Table 1 Comparison of proposed method and GT, nbest method and GT

	nbest Arms	Ours Arms	nbest Legs	Ours Legs	nbest Torso	Ours Torso	nbest All	Ours All	nbest Mean	Ours Mean
Video 1	13.96 %	**25.90 %**	**45.30 %**	37.37 %	24.99 %	**40.31 %**	45.70 %	**62.45 %**	32.49 %	**41.51 %**
Video 2	12.15 %	**32.49 %**	24.71 %	**43.87 %**	42.61 %	**56.41 %**	38.47 %	**62.43 %**	29.49 %	**48.80 %**
Video 3	12.62 %	**25.00 %**	42.69 %	**42.99 %**	**45.41 %**	44.03 %	48.75 %	**67.98 %**	37.37 %	**45.00 %**
Video 4	22.54 %	**25.93 %**	44.76 %	**54.29 %**	51.20 %	**53.81 %**	50.21 %	**67.77 %**	42.18 %	**50.45 %**
Video 5	22.29 %	**56.10 %**	**65.32 %**	64.17 %	49.75 %	**63.18 %**	62.96 %	**84.58 %**	50.08 %	**67.01 %**
Mean	16.71 %	**33.08 %**	44.56 %	**48.54 %**	42.79 %	**51.55 %**	49.22 %	**69.04 %**	38.32 %	**50.55 %**

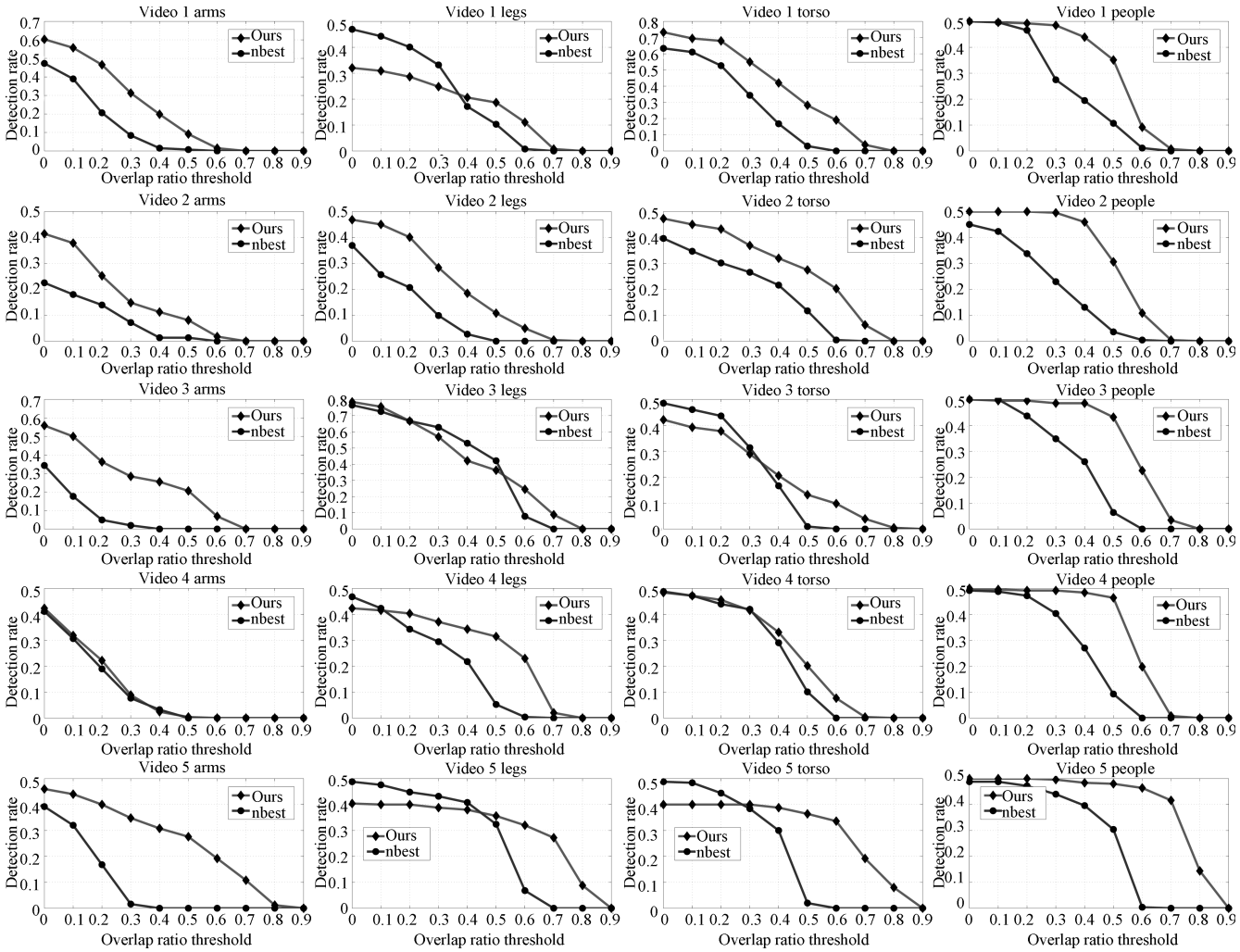


图 9　该方法与 nbest 方法实验结果的正确率曲线图

Fig. 9　Detection rate comparisons of nbest and proposed method

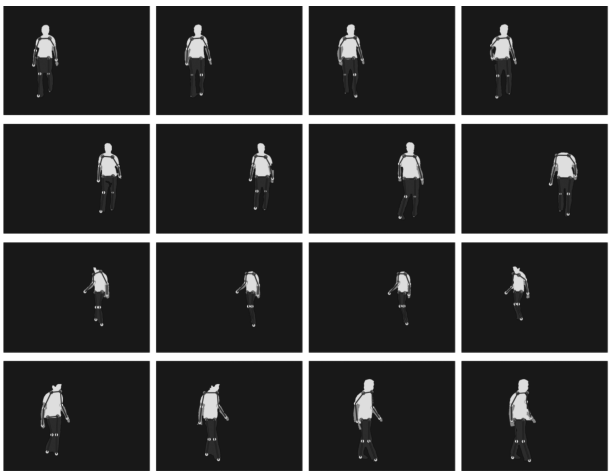


图 10　行人姿势估计结果

Fig. 10　Pedestrian pose estimation results

距离越近; 反之越远. 图 10 给出了在本方法分割结果的基础上, 4 个不同朝向的行人姿势估计结果, 图中第 1 行到第 4 行分别为正面、背面、左面和右面 4 个朝向.

3　总结与展望

本文提出了一种新的人身体部位所在区域的视频分割方法. 该方法不需要任何初始化, 对于各种旋转与尺度的变化都具有较好的鲁棒性. 实验中分别对该方法进行了定性和定量的分析比较, 实验结果表明, 与类似的方法相比, 该方法不仅适用于直立行走的行人, 对各种姿势的人也可以得到较好的实验结果. 另外, 还试将行人视频的分割结果应用到行人行走姿势的估计中, 为进一步行人异常行为的分析奠定了良好的基础. 当然, 针对实验中出现的不足, 比如如何提高提议 (Proposals) 的准确率等问题, 也是接下来的工作中需要解决的. 另外, 在接下来的工作中, 会在该工作的基础上继续进行体育、舞蹈等运动视频中目标姿势的估计与分析, 以及其在智能视

频监控与人机交互领域的应用.

References

1 Criminisi A, Cross G, Blake A, Kolmogorov V. Bilayer segmentation of live video. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2006. 53−60

2 Cheung S C S, Kamath C. Robust techniques for background subtraction in urban traffic video. In: Proceedings of SPIE 5308, Visual Communications and Image Processing. San Jose, USA: SPIE, 2004, **5308**: 881−892

3 Hayman E, Eklundh J. Statistical background subtraction for a mobile observer. In: Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE, 2003. 67−74

4 Ren Y, Chua C S, Ho Y K. Statistical background modeling for non-stationary camera. *Pattern Recognition Letters*, 2003, **24**(1−3): 183−196

5 Giordano D, Murabito F, Palazzo S, Spampinato C. Superpixel-based video object segmentation using perceptual organization and location prior. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 4814−4822

6 Brendel W, Todorovic S. Video object segmentation by tracking regions. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 833−840

7 Li F X, Kim T, Humayun A, Tsai D, Rehg J M. Video segmentation by tracking many figure-ground segments. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 2192−2199

8 Varas D, Marques F. Region-based particle filter for video object segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 3470−3477

9 Arbeláez P A, Pont-Tuset J, Barron J T, Marques F, Malik J. Multiscale combinatorial grouping. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 328−335

10 Tsai Y H, Yang M H, Black M J. Video segmentation via object flow. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016.

11 Ramakanth S A, Babu R V. Seamseg: video object segmentation using patch seams. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 376−383

12 Faktor A, Irani M. Video segmentation by non-local consensus voting. In: Proceedings British Machine Vision Conference 2014. Nottingham: BMVA Press, 2014.

13 Papazoglou A, Ferrari V. Fast object segmentation in unconstrained video. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 1777−1784

14 Rother C, Kolmogorov V, Blake A. "Grabcut": interactive foreground extraction using iterated graph cuts. *Acm Transactions on Graphics*, 2004, **23**(3): 309−314

15 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 580−587

16 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L. Microsoft COCO: common objects in context. In: Proceedings of the 13th European Conference. Zurich, Switzerland: Springer International Publishing, 2014. 740−755

17 Endres I, Hoiem D. Category-independent object proposals with diverse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **36**(2): 222−234

18 Krähenbühl P, Koltun V. Geodesic object proposals. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer International Publishing, 2014. 725−739

19 Zhang D, Javed O, Shah M. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, Oregon, USA: IEEE, 2013. 628−635

20 Fragkiadaki K, Arbelaez P, Felsen P, Malik J. Learning to segment moving objects in videos. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 4083−4090

21 Perazzi F, Wang O, Gross M, Sorkine-Hornung A. Fully connected object proposals for video segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 3227−3234

22 Kundu A, Vineet V, Koltun V. Feature space optimization for semantic video segmentation. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016.

23 Seguin G, Bojanowski P, Lajugie R, Laptev I. Instance-level video segmentation from object tracks. In: Proceeding of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016.

24 Lee Y J, Kim J, Grauman J. Key-Segments for video object segmentation. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spanish: IEEE, 2011. 1995−2002

25 Tsai D, Flagg M, Rehg J. Motion coherent tracking with multi-label MRF optimization. In: Proceedings of the British Machine Vision Conference 2010. Aberystwyth: BMVA Press, 2010. 190−202

26 Ramanan D, Forsyth D A, Zisserman A. Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(1): 65−81

27 Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA: IEEE, 2011. 1385−1392

28 Endres I, Hoiem D. Category independent object proposals. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010. 575−588

29 Ling H B, Jacobs D W. Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(2): 286−299

30 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 886−893

31 Park D, Ramanan D. N-best maximal decoders for part models. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 2627−2634

32 Sigal L, Black M J. HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion. Techniacl Report CS-06-08. Brown University, USA, 2006

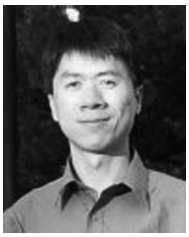
33 Grundmann M, Kwatra V, Han M, Essa I. Efficient hierarchical graph based video segmentation. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 2141−2148

34 Oneata D, Revaud J, Verbeek J, Schmid C. Spatio-temporal object detection proposals. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer International Publishing, 2014. 737−752

35 Pele O, Werman M. Fast and robust earth mover′s distance. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 460−467

36 Brox T, Malik J. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(3): 500−513

薄一航 北京电影学院美术学院讲师. 2011 年博士毕业于北京交通大学, 2011～2014 年分别在中国科学院自动化所与波士顿学院从事博士后研究工作. 主要研究方向为图像与视频分割, 人的行为和姿势估计, 目标跟踪, 交互设计. 本文通信作者.
E-mail: boyihang@sina.com
(**BO Yi-Hang** Assistant professor in Fine Art Department, Beijing Film Academy. She received her Ph. D. degree at the School of Computer and Information Technology, Beijing Jiaotong University in 2011, postdoctor at the Institute of Automation, Chinese Academy of Sciences, China and Boston College, USA from 2011 to 2014. Her research interest covers image and video segmentation, human pose and action recognition, object tracking and interactive design. Corresponding author of this paper.)



HAO Jiang 波士顿学院计算机科学系副教授. 主要研究方向为图像匹配, 目标检测, 目标跟踪, 姿势和行为估计.
E-mail: hjiang@cs.bc.edu
(**HAO Jiang** Associate professor in the Computer Science Department, Boston College, USA. His research interest covers image matching, object detection, tracking, pose and action recognition.)