JIG
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

2020
07
VOL.25
ISSN1006-8961
CN11-3758/TB

俯视深度头肩序列识别 P1393

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第7期（总第291期）
2020年7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
中文核心期刊

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 顾行发
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218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

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 Gu Xingfa
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
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

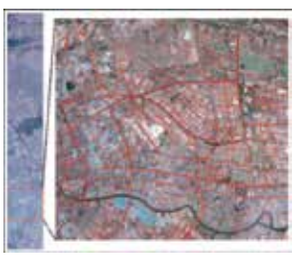
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年7月 第25卷 第7期（总第291期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

非真实感绘制技术研究现状与展望(第1283页)

图卷积网络下牙齿种子点自动选取(第1481页)



迁移学习下高分快视数据道路快速提取(第1501页)

Volume 25, Number 7
Published July 16, 2020

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Research status and progress of non-photorealistic rendering technology(P1283)

Automatic selection of tooth seed point by graph convolutional network(P1481)

Rapid road extraction from quick view imagery of high-resolution satellites combined with transfer learning(P1501)

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2020)07-1459-14

论文引用格式:Ma M, Li Y B, Wu X Q, Gao J F and Pan H P. 2020. Human pose tracking based on multi-feature fusion in videos. Journal of Image and Graphics,25(07):1459-1472(马淼,李贻斌,武宪青,高金凤,潘海鹏. 2020. 视频中多特征融合人体姿态跟踪. 中国图象图形学报,25(07):1459-1472)[DOI:10.11834/jig.190494]

视频中多特征融合人体姿态跟踪

马淼[1],李贻斌[2],武宪青[1],高金凤[1],潘海鹏[1]

1. 浙江理工大学,杭州　310018;2. 山东大学,济南　250100

摘　要:**目的**　目前已有的人体姿态跟踪算法的跟踪精度仍有待提高,特别是对灵活运动的手臂部位的跟踪。为提高人体姿态的跟踪精度,本文首次提出一种将视觉时空信息与深度学习网络相结合的人体姿态跟踪方法。**方法**　在人体姿态跟踪过程中,利用视频时间信息计算出人体目标区域的运动信息,使用运动信息对人体部位姿态模型在帧间传递;考虑到基于图像空间特征的方法对形态较为固定的人体部位如躯干和头部能够较好地检测,而对手臂的检测效果较差,构造并训练一种轻量级的深度学习网络,用于生成人体手臂部位的附加候选样本;利用深度学习网络生成手臂特征一致性概率图,与视频空间信息结合计算得到最优部位姿态,并将各部位重组为完整人体姿态跟踪结果。**结果**　使用两个具有挑战性的人体姿态跟踪数据集 VideoPose2.0 和 YouTubePose 对本文算法进行验证,得到的手臂关节点平均跟踪精度分别为 81.4% 和 84.5%,与现有方法相比有明显提高;此外,通过在 VideoPose2.0 数据集上的实验,验证了本文提出的对下臂附加采样的算法和手臂特征一致性计算的算法能够有效提高人体姿态关节点的跟踪精度。**结论**　提出的结合时空信息与深度学习网络的人体姿态跟踪方法能够有效提高人体姿态跟踪的精度,特别是对灵活运动的人体姿态下臂关节点的跟踪精度有显著提高。

关键词:人体姿态跟踪;视觉目标跟踪;人机交互;深度学习网络;关节点概率图

Human pose tracking based on multi-feature fusion in videos

Ma Miao[1], Li Yibin[2], Wu Xianqing[1], Gao Jinfeng[1], Pan Haipeng[1]

1. Zhejiang Sci-Tech University, Hangzhou　310018, China; 2. Shandong University, Jinan　250100, China

Abstract: **Objective** Human pose tracking in video sequences aims to estimate the pose of a certain person in each frame using image and video cues and consecutively track the human pose throughout the entire video. This field has been increasingly investigated because the development of artificial intelligence and the Internet of Things makes human-computer interaction frequent. Robots or intelligent agents would understand human action and intention by visually tracking human poses. At present, researchers frequently use pictorial structure model to express human poses and use inference methods for tracking. However, the tracking accuracy of current human pose tracking methods needs to be improved, especially for flexible moving arm parts. Although different types of features describe different types of information, the crucial point of human pose tracking depends on utilizing and combining appropriate features. We investigate the construction of effective features to accurately describe the poses of different body parts and propose a method that combines video spatial and tem-

收稿日期:2019-10-08;**修回日期**:2019-12-12;**预印本日期**:2019-12-19

基金项目:国家自然科学基金项目(61803339,61673245);浙江省自然科学基金项目(LQ19F030014,LQ18F030011);浙江理工大学青年创新专项(2019Q035)

Supported by: National Natural Science Foundation of China (61803339, 61673245); Natural Science Foundation of Zhejiang Province, China (LQ19F030014, LQ18F030011); Fundamental Research Funds of Zhejiang Sci-Tech University (2019Q035)

poral features and deep learning features to improve the accuracy of human pose tracking. This paper presents a novel human pose tracking method that effectively uses various video information to optimize human pose tracking in video sequences. **Method** An evaluation criterion should be used to track a visual target. Human pose is an articulated complex visual target, and evaluating it as a whole leads to ambiguity. In this case, this paper proposes a decomposable human pose expression model that can track each part of human body separately during the video and recombine parts into an entire body pose in each single image. Human pose is expressed as a principal component analysis model of trained contour shape similar to a puppet, and each human part pose contour can be calculated using key points and model parameters. As human pose unpredictably changes, tracking while detecting would improve the human pose tracking accuracy, which is different from traditional visual tracking tasks. During tracking, the video temporal information in the region of each human part target is used to calculate the motion information of each human part pose, and then the motion information is used to propagate the human part contour from each frame to the next. The propagated human parts are treated as human body part candidates in the current frame for subsequent calculation. During propagation, the background motion would disturb and pollute the foreground target motion information, resulting in the deviations of human part candidates obtained through propagation using motion information. To avoid the influence of propagated human part pose deviation, a pictorial structure-based method is adopted to generate additional human body pose candidates and are then decomposed into human body part poses for body part tracking and optimization. The pictorial structure-based method can detect relatively fixed body parts, such as trunk and head, whereas the detection effect of arms is poor because arms move flexibly and their shapes substantially and frequently change. In this circumstance, the problem of arm detection should be solved. A lightweight deep learning network is constructed and trained to generate probability graphs for the key points of human lower arms to solve this problem. Sampling from the generated probability graphs can obtain additional candidates of human lower arm poses. The propagated and generated human part pose candidates need to evaluated. The proposed evaluation method considers image spatial information and deep learning knowledge. Spatial information includes color and contour likelihoods, where the color likelihood function ensures the consistency of part color during tracking, and the contour likelihood function ensures the consistency of human part model contour with image contour feature. The proposed deep learning network can generate probability maps of lower arm feature consistency for each side to reveal the image feature consistency for each calculated lower arm candidates. The spatial and deep learning features work together to evaluate and optimize the part poses for each human part, and the optimized parts are recombined into integrated human pose, where the negative recombined human body poses are eliminated by the shape constraints of the proposed decomposable human model. The recombined optimized human entire pose is the human pose tracking result for the current video frame and is decomposed and propagated to the next frame for subsequent human pose tracking. **Result** Two publicly available challenging human pose tracking datasets, namely, VideoPose2.0 and YouTubePose datasets, are used to verify the proposed human pose tracking method. For the VideoPose2.0 dataset, the key point accuracy of human pose tracking for shoulders, elbows, and wrists are 90.5%, 82.6%, and 71.2%, respectively, and the average key point accuracy is 81.4%. The results are higher than state-of-the-art methods, such as the method based on conditional random field model (higher by 15.3%), the method based on tree structure reasoning model (higher by 3.9%), and the method based on max-margin Markova model (higher by 8.8%). For the YouTubePose dataset, the key point accuracy of human pose tracking for shoulders, elbows, and wrists are 86.2%, 84.8%, and 81.6%, respectively, and the average key point accuracy is 84.5%. The results are higher than state-of-the-art methods, such as the method based on flowing convent model (higher by 13.7%), the method based on dependent pairwise relation model (higher by 1.1%), and the method based on mixed part sequence reasoning model (higher by 15.9%). The proposed crucial algorithms of additional sampling and feature consistency of lower arm are verified on the VideoPose2.0 dataset, thereby effectively improving the tracking accuracy of lower arm joints by 5.2% and 31.2%, respectively. **Conclusion** Experimental results show that the proposed human pose tracking method that uses spatial-temporal cues coupled with deep learning probability maps can effectively improve the pose tracking accuracy, especially for the flexible moving lower arms.

Key words: human pose tracking; visual target tracking; human-computer interaction; deep learning network; probability map for joints

0 引 言

视频序列中的人体姿态跟踪任务是利用视频图像信息估计出每一帧图像中特定人体姿态，并能够在连续的视频中对人体姿态变化进行准确跟踪。随着人工智能技术的发展和人机交互需求的增加，视频中人体姿态跟踪问题的研究受到越来越多的关注（Bajcsy 等，2018；Duckworth 等，2019）。

提高人体姿态跟踪精度需要研究如何有效利用视频信息对人体姿态跟踪结果进行优化。目前常用的人体姿态优化方法可分为基于图形模型的方法（Cherian 等，2014；Zhao 等，2015）和基于深度学习的方法（Wei 等，2016；Cao 等，2017）两大类。

现有的基于图形模型的人体姿态表达方法大多以图形结构（pictorial structures）模型为基础（Fischler 和 Elschlager，1973），利用树结构算法对人体各个部位进行表达和推理，人体各部位之间定义旋转和平移变换，从而实现对复杂铰接人体姿态的描述。为了解决人体姿态跟踪问题，需要有效利用视频帧之间的连续性信息，因此有效提取视频中的运动信息和图像表观特征是基于图形模型优化人体姿态的关键。Fragkiadaki 等人（2013）使用图像特征计算出人体姿态候选样本后，利用帧间光流图解析和推理出相邻帧之间的动力学约束，在图像信息和运动学约束间迭代优化，进而实现对视频中人体姿态的最优估计。Cherian 等人（2014）利用视频运动信息使相邻图像中的人体姿态在帧间构成环路，然后利用图像信息的一致性对环路进行优化，剔除误估计的人体姿态样本，从而获得最优人体姿态。然而，上述方法虽然将图像特征与运动特征相结合，却需要视频图像之间的往复运算和推理，无法实现复杂铰接人体姿态的在线连续跟踪。

人体姿态跟踪问题需要将人体姿态估计问题与视觉跟踪问题相结合，实现边跟踪边检测。由于人体肢体活动的灵活性，使得这类问题较难实现。Zhao 等人（2015）将复杂铰接人体姿态跟踪视为特殊的视觉目标跟踪问题，利用时间解析的马尔可夫链实现跟踪，利用空间解析的马尔可夫网络实现单帧内的人体姿态优化，两个马尔可夫模型联合训练，从而实现视频中的人体姿态跟踪。Ma 等人（2016）提出一种基于可变结构的全局局部人体姿态表达模型，局部层侧重基于运动特征的视觉目标跟踪问题，全局层侧重基于图像特征的人体姿态优化问题，从而达到人体姿态在线跟踪的目的。

上述基于图模型的方法将图像特征与帧间运动特征相结合，从而对人体姿态进行优化。此类方法的人体姿态跟踪精度受所选用的图像特征有效性影响，优质的图像特征或多种图像特征的有效融合能够提升姿态跟踪任务的精度。

深度学习网络是一种挖掘深层次图像特征的新兴方法。基于深度学习解决视觉问题的方法层出不穷（Krizhevsky 等，2012；Szegedy 等，2015），卷积神经网络（convolutional neural network，CNN）作为用于图像处理的一种深度学习算法，在图像处理中的应用可以看做是一种特殊的特征提取过程，每一个卷积核提取一种图像特征，卷积层的叠加使用可以挖掘图像中更深层次的信息，这些信息可看做融合图像上下文信息的语义特征。因此，在充分有效训练的前提下，利用卷积神经网络提取的图像特征更具有表达性。通过精巧的网络架构和大量数据的训练，卷积神经网络能够较好地处理图像中的目标识别问题（Girshick 等，2014；He 等，2016）。

图像中的人体姿态亦可看做一种多视觉目标铰接组合的形式，因此在精细设计网络结构，使用大量卷积层和参数，以及海量训练集充分训练的前提下，可以将人体的每个肢体作为图像目标，用深度学习网络识别出来。Newell 等人（2016）提出一种形状类似于堆叠的沙漏的复杂深度学习网络，将多层卷积产生的不同分辨率特征单独处理后再进行穿插连接，数据汇总后再集中使用卷积处理。这样的复杂网络着重训练了人体各部位的图形特征，从而输出人体关键关节位置图。Chen 等人（2018）提出的深度学习网络同样使用了不同卷积层之间的跳跃连接，同时又加入了不同卷积层的分组汇总，通过各组之间有规律的融合、分叉等复杂网络结构的级联操作，加强各级神经元之间的信息传递，从而结合更多特征对图形中的人体各部位进行识别，得到人体肢体的位置信息。从上述方法可看出，为了对图像中的人体姿态进行更好的识别，学者们采用越来越复杂的网络结构，且需要海量数据集对网络参数进行训练。此外，由于卷积神经网络的输入只能是单帧图像，因此无法有效利用视频中的运动信息和

时间一致性信息，在视觉跟踪中的应用受到了限制。

针对上述人体姿态跟踪中存在的问题，并考虑到在人机交互中人们更关注人体上半身特别是手臂的姿态和动作（Zheng 等，2019；López-Quintero 等，2017），本文提出一种将运动信息、图像空间信息以及深度学习网络相结合的人体姿态跟踪方法。利用运动信息加强视频帧之间的关联性，利用丰富的图像空间信息有针对性地优化人体姿态，并构造轻量级的深度学习网络强化对活动最为灵活的手臂的识别，从而实现人体姿态的高精度跟踪。

本文的主要贡献有：

1）在人体姿态跟踪过程中将人体姿态进行解耦，加强并细化人体各个部位在视频帧间的运动特征和表观特征，为提高人体姿态跟踪精度提出新的解决思路；

2）针对人体姿态中运动最灵活的手臂部位，有针对性地构造并训练轻量级的深度学习网络，生成表达手臂关节位置的概率图和手臂连续性判别概率图；

3）首次采用可解耦人体姿态表达模型，结合时空信息与深度学习信息，实现人体姿态的高精度跟踪。

1 人体姿态表达与跟踪策略

1.1 可解耦人体姿态表达模型

为了更好地利用视频中人体区域的图像和运动特征，需要对人体轮廓进行表达。人体轮廓是一种复杂铰接的图形形状，因此无法用单一的模型对人体姿态进行表达。本文将木偶模型（Zuffi 等，2012）进行解耦，木偶模型是一种人体姿态轮廓表达模型，通过对轮廓主成分进行分析和训练，能够利用人体姿态的关节点计算出相应肢体对应的图像轮廓，从而更有针对性地计算图像特征。人体上半身姿态可分解为 6 个部分，如图 1 所示，图中人体部位轮廓两两相连处为铰接关节点。

人体每一个部位 i（i 表示 6 种人体部位之一，$i \in [1,2,\cdots,6]$）的轮廓 $\boldsymbol{C}_i$ 以及人体关节点位置 $\boldsymbol{k}_i$ 与预先训练得到的人体形状模型之间的关系为

图 1 可解耦人体姿态表达模型

Fig. 1 Decoupled human pose expression model

$$\begin{bmatrix} \boldsymbol{C}_i \\ \boldsymbol{k}_i \end{bmatrix} = \boldsymbol{B}_i z_i + \boldsymbol{m}_i \tag{1}$$

式中，$\boldsymbol{k}_i$ 表示人体部位 i 中关节点位置；矩阵 $\boldsymbol{B}_i$ 和向量 $\boldsymbol{m}_i$ 表示人体形状模型参数，是由 3 维模型 SCAPE（shape completion and animation of people）（Anguelov 等，2005）的 3D 轮廓基于主成分分析的方法训练得到的（Zuffi 等，2012），向量 $\boldsymbol{m}_i$ 表示部位 i 的轮廓及关节点位置的均值，矩阵 $\boldsymbol{B}_i$ 包含训练数据的特征向量主特征值；z_i 是形变系数，解耦时通过继承人体上半身各关节之间的相对位置形变计算得到。解耦后的人体部位轮廓可使用变化的关节点位置与轮廓模型参数计算，即

$$\boldsymbol{C}_i = f(\boldsymbol{k}_i \mid \boldsymbol{B}_i, \boldsymbol{m}_i, z_i) \tag{2}$$

使用可解耦的人体姿态轮廓能够有针对性地从图像中提取出特定人体姿态所包含的图像特征、运动特征等信息，在视觉目标跟踪中能够更有效地对比相邻帧中同一人体部位的特征，获取视频中的一致性信息，从而使人体姿态跟踪更具有针对性。

1.2 人体姿态跟踪策略

本文提出的结合时空信息与深度学习网络的人体姿态跟踪策略如图 2 所示。要解决复杂场景中的人体姿态跟踪问题，需要考虑场景中的各种干扰，例如背景杂乱或场景中有多人出现的问题，如图 2（a）所示。因此算法第 1 帧需要确定被跟踪人体目标及其初始姿态，计算并记录图像特征识别信息，如各部位区域内颜色分布信息，如图 2（b）所示。人体姿态的跟踪过程需要用到视频时间信息、图像空间信息和深度学习信息，如图 2（c）所示。

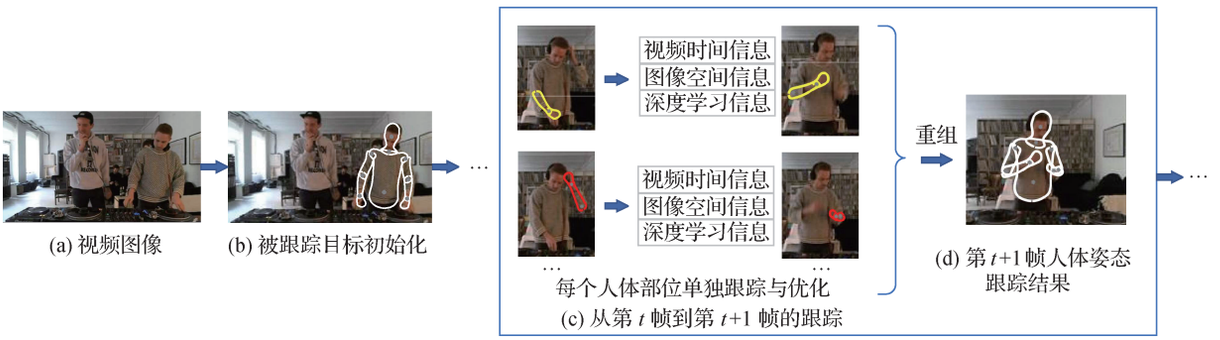


图 2　人体姿态跟踪策略图

Fig. 2　Schematic of the human posture tracking strategy ((a) video frame; (b) initialization of tracked target; (c) tracking from frame t to $t+1$; (d) human pose tracking result of frame $t+1$)

在跟踪过程中，利用视频时间信息计算出人体各部位目标区域的运动信息，并使用运动信息将人体各部位姿态轮廓从第 t 帧传递到第 $t+1$ 帧，作为第 $t+1$ 帧中的人体部位姿态候选样本。为避免由于背景运动对前景运动信息的干扰而造成由运动信息传递得到的人体姿态产生偏差，采用 Yang 和 Ramanan(2012)提出的人体姿态样本生成的方法在第 $t+1$ 帧中利用单帧图像空间信息生成附加的人体姿态候选样本。考虑到基于单帧图像空间信息的方法对形态较为固定的人体部位如躯干和头部能够较好地检测，而对手臂的检测效果较差，因此需要着重解决手臂检测问题。本文构造并训练一种轻量级的深度学习网络，用于提取人体手臂部位的附加候选样本。然后，在第 $t+1$ 帧图像中针对人体每个部位，利用图像信息及深度学习特征，计算得到最优部位姿态，并将各部位重组为第 $t+1$ 帧中完整人体姿态跟踪结果，如图 2(d)所示。

2　人体姿态跟踪算法

2.1　利用时空信息对人体姿态进行跟踪

2.1.1　利用运动信息实现目标的帧间传递

视频的时间信息中包含了人体目标在帧与帧之间的相对运动信息。为了从视频时间序列中有效提取出人体目标的运动，在跟踪过程中需要计算相邻帧之间的光流图 $\boldsymbol{F}$（Liu，2009），光流图 $\boldsymbol{F}$ 为与视频图像尺寸相同的两通道数据，记为 $\boldsymbol{F}_x$ 与 $\boldsymbol{F}_y$，分别代表第 t 帧图像中各像素传递到第 $t+1$ 帧图像时对应的 x 方向和 y 方向运动。为求取人体姿态中每个关节点的运动，在光流图中计算每个关节点对应邻域 $\boldsymbol{\Omega}$ 内的 x 方向与 y 方向运动均值作为帧间关节点相对运动的像素距离

$$\begin{cases} d_x^k = \dfrac{1}{N_{\boldsymbol{\Omega}_k}} \displaystyle\sum_{(i,j)\in\boldsymbol{\Omega}_k} F_x(i,j) \\ d_y^k = \dfrac{1}{N_{\boldsymbol{\Omega}_k}} \displaystyle\sum_{(i,j)\in\boldsymbol{\Omega}_k} F_y(i,j) \end{cases} \tag{3}$$

式中，d_x^k 和 d_y^k 分别表示关节点 k 在帧间发生的 x 方向与 y 方向的运动，$\boldsymbol{\Omega}_k$ 表示在第 t 帧中关节点 k 位置附近的邻域，(i,j) 表示邻域 $\boldsymbol{\Omega}_k$ 内的点，$N_{\boldsymbol{\Omega}_k}$ 表示 $\boldsymbol{\Omega}_k$ 邻域内的点的个数。图 3 中给出了利用视频运动信息实现人体姿态帧间传递的示意图，视频中每对相邻图像如图 3(a)所示，计算得到的相邻图像的光流图如图 3(b)所示，左边为沿 x 轴方向的运动信息，图中颜色色调越暖表示沿 x 方向正向运动幅度越大，色调越冷表示沿 x 方向反向运动幅度越大，右边为沿 y 轴方向的运动信息。由第 t 帧中的人体姿态可使用式(3)计算出每个关节点处对应的 x 方向和 y 方向运动，并根据计算出的运动将第 t 帧中的人体

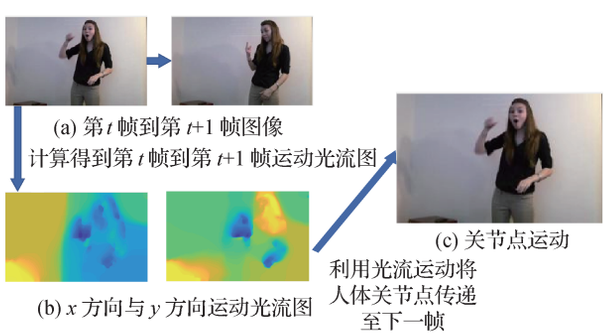


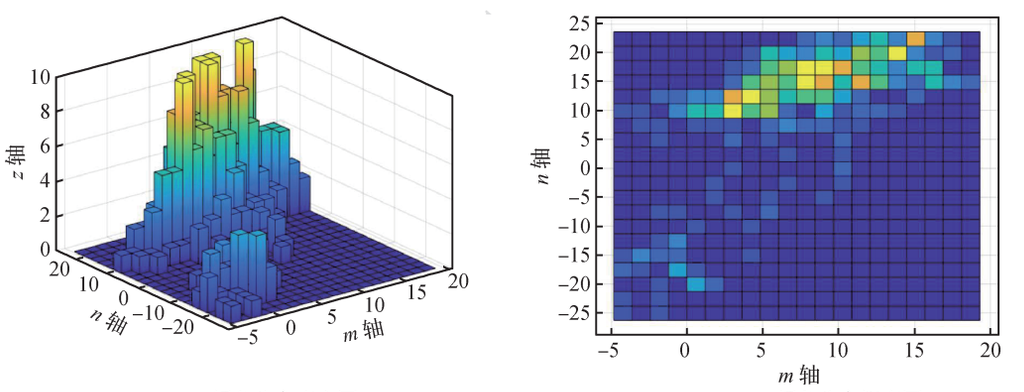
图 3　利用视频运动信息实现人体姿态帧间传递

Fig. 3　Propagate human pose between frames utilizing video motion information ((a) frames from t to $t+1$; (b) optical flow images of x direction and y direction; (c) motion of key points)

姿态关节点位置传递到第 $t+1$ 帧中，如图 3(c)所示。

2.1.2 颜色似然函数

在各类图像信息中，颜色信息能够直观地区分不同视觉目标（许允喜和陈方，2015），通过计算颜色似然函数能够强化视频序列中目标的一致性。在人体姿态目标初始化时，首先将人体各部位轮廓内图像区域的颜色信息由 RGB 3 通道转换为 CIE *lab* 三通道颜色空间，其中 l 通道表示亮度，a 和 b 两通道表示颜色。转换的目的是通过对 l 通道施加阈值限制，排除光照引起的颜色畸变对颜色似然函数的影响。计算人体姿态各部位轮廓内的颜色直方图，考虑由关节点计算得到的人体姿态部位轮廓与图像中实际轮廓难以完全吻合，因此将计算得到的轮廓进行膨胀后取轮廓内图像区域，计算其 a 和 b 两通道的颜色直方图，记为 $\boldsymbol{h}_{cl}^{i}$，如图 4(a)所示。其中 m 轴和 n 轴表示将 CIE *lab* 颜色空间按颜色进行分格，这里采用 20×20 像素的分格方式，z 轴表示落入每个格子颜色范围内的像素个数。



(a) CIE *lab*颜色分布直方图　　(b) CIE *lab*颜色分布概率图

图 4　CIE *lab* 颜色空间直方图和概率图

Fig. 4　Color histogram and probability graph of CIE *lab* color space

((a) color distribution histogram in CIE *lab* color space; (b) color distribution probability graph in CIE *lab* color space)

分别将每个部位记录的颜色直方图归一化为颜色概率图

$$V^{i}(m,n)=\frac{h_{cl}^{i}(m,n)}{\sum_{(m,n)\in bins}h_{cl}^{i}(m,n)} \tag{4}$$

式中，$V^{i}(m,n)$ 表示部位 i 区域内颜色空间对应的概率，$(m,n)\in bins$ 表示 20×20 分格中的所有格子。归一化的颜色概率图如图 4(b)所示。在人体姿态部位跟踪过程中，第 $t+1$ 帧中部位 i 的第 q 个姿态候选样本颜色一致性似然函数为

$$p_{\mathrm{cl}}(\boldsymbol{C}_{i}^{(q)}\mid \boldsymbol{I}_{t},\boldsymbol{I}_{t+1})=\sum_{(x,y)\subset \boldsymbol{C}_{i}^{(q)}}V^{i}(CIE_{ab}^{bins}(x,y)) \tag{5}$$

式中，下标 cl 代表颜色（color），$\boldsymbol{I}_{t}$ 和 $\boldsymbol{I}_{t+1}$ 表示第 t 帧和第 $t+1$ 帧图像，$\boldsymbol{C}_{i}^{(q)}$ 表示部位 i 的第 q 个姿态候选样本的轮廓；点 (x,y) 表示轮廓 $\boldsymbol{C}_{i}^{(q)}$ 膨胀操作后内部的像素点；$CIE_{ab}^{bins}(x,y)$ 表示将点 (x,y) 处像素的颜色值转换为 CIE *lab* 3 通道颜色空间，并将其 a 和 b 两通道的颜色对应到颜色空间概率图的分格 (m,n) 中；$V^{i}(\cdot)$ 函数表示用式(4)计算归一化颜色概率值。

2.1.3 轮廓似然函数

轮廓信息反映了视觉目标与背景图像之间的界限，在人体姿态关节点跟踪效果较好的情况下，用式(2)计算得到的人体部位轮廓能够较好地拟合图像中的人体部位，因此可通过计算轮廓的拟合情况来判断人体姿态跟踪结果与图像轮廓信息的拟合情况。由于人体姿态表达模型由不相关数据集训练得到，计算得到的轮廓与图像真实轮廓之间存在偏差在所难免，为了对轮廓进行更好的拟合，在计算得到的轮廓基础上向内向外各增加一层轮廓。对 3 层轮廓对应的图像位置计算梯度直方图特征向量 $\boldsymbol{h}_{\mathrm{ct}}^{j}$，其中 $j\in[1,2,3]$ 表示 3 层轮廓单独计算梯度直方图（下标 ct 代表轮廓（contour））。使用支持向量机计算每个梯度直方图的得分来表征该梯度直方图属于真实图像轮廓的概率，记为 $\zeta_{i}(\boldsymbol{h}_{\mathrm{ct}}^{j})$。将支持向量机的输出结果正则化（Zuffi 等，2013），并取 3 层轮廓直方图中与图像轮廓拟合最优的作为部位 i 的第 q 个姿态候选样本的轮廓似然函数

$$p_{\mathrm{ct}}(\boldsymbol{C}_{i}^{(q)}\mid \boldsymbol{I}_{t},\boldsymbol{I}_{t+1})=$$

$$\max_{j\in[1,2,3]}\left(\frac{1}{1+\exp(a_i\zeta_i(\boldsymbol{h}_{\mathrm{ct}}^{j})+b_i)}\right) \tag{6}$$

式中，$\max_{j\in[1,2,3]}(\cdot)$ 表示对 3 层轮廓的正则化轮廓似然函数取最大值；a_i 和 b_i 为对应每个人体姿态部位支持向量机 ζ_i 的正则化系数。

2.2 利用深度学习特征对人体姿态进行跟踪

人体姿态中手臂的姿态及动作变化最灵活，是姿态跟踪中的难点，采用边跟踪边检测的方法可弥补手臂跟踪算法的不足。本文针对图像中的手臂位置构建并训练一种轻量级的深度学习网络（如图 5 所示），生成手臂关节点概率图和手臂特征一致性概率图，优化手臂姿态的跟踪效果。

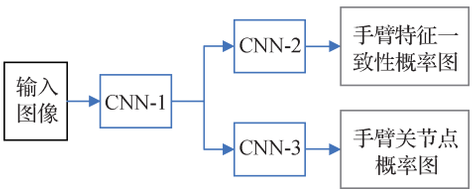


图 5 轻量级的深度学习网络

Fig. 5 Simple deep learning networks

提出的轻量级深度学习网络包含两个分支，CNN-1 为预处理网络，CNN-2 和 CNN-3 为分支中分别生成手臂关节点概率图和手臂特征一致性概率图的子网络。

2.2.1 深度学习网络结构的选择和架构思路

在设计轻量级深度学习网络时，借鉴了 VGG（visual geometry group）网络的设计思路。VGG 网络是由牛津大学视觉几何课题组提出的（Simonyan 和 Zisserman，2015），用来完成图像识别任务，网络参数由 ILSVRC（large scale visual recognition challenge）数据集（Russakovsky 等，2015）训练得到。该网络采用 3 ×3 的感受野叠加的卷积形式，其优点是小卷积核的叠加使用比直接使用大卷积核涉及的参数指数级减少，而且多个卷积操作叠加使用能够加深非线性激活层的交替作用，从而挖掘出图像中更深层次的表观特征。因此本文提出的网络结构中预处理网络 CNN-1 采用了 3 ×3 的卷积层叠加形式。

考虑到图 5 中的 CNN-1 网络需要尽量多地挖掘出视频图像中的图形端点信息和图形连接情况，因此采用 10 层 3 ×3 的卷积结构，如图 6 所示。图 5 中的 CNN-2 网络和 CNN-3 网络的作用分别是提取手臂一致性概率图和手臂关节点概率图，CNN-2 和 CNN-3 的网络结构相同，为更好地训练和感知手臂的特征，采用具有较大感受野的 7 ×7 的卷积结构，如图 7 所示。

C1	C2	P2	C3	C4	P4	C5	C6	C7	C8	P8	C9	C10	C11	C12
64	64	64	128	128	128	256	256	256	256	256	512	512	512	128

C1–C10: 卷积层
kernel: 3×3;
stride = 1;
pad = 1.

P2、P4、P8: 池化层
kernel: 2×2;
stride = 2;
pad = 0.

C11、C12: 降维卷积层
kernel: 1×1;
stride = 1;
pad = 0.

图 6 CNN-1 的网络结构

Fig. 6 Network structure of CNN-1

C1	C2	C3	C4	C5	C6	C7
128	128	128	128	128	64	4

C1–C5: 卷积层
kernel: 7×7;
stride = 1;
pad = 3.

C6、C7: 降维卷积层
kernel: 1×1;
stride = 1;
pad = 0.

图 7 CNN-2 与 CNN-3 的网络结构

Fig. 7 Network structure of CNN-2 and CNN-3

图 6 中 CNN-1 的前 10 层卷积层均使用 3 ×3 的卷积核，卷积步长为 1，填充为 1；每个池化层采用 2 ×2 的卷积核，卷积步长为 2；后两个卷积层使用 1 ×1 的卷积核，加入两层使用 1 ×1 的卷积核的目的是快速降维。图 6 中标注了每个卷积层使用的卷积核个数。

图 7 中 CNN 结构的前 5 层卷积层均使用 7 ×7 的卷积核，卷积步长为 1，填充为 3；后两个卷积层为快速降维卷积层，使用 1 ×1 的卷积核，卷积步长为 1，填充为 0。

本文构造的轻量级的深度学习网络的输出为关节位置概率图，而现有的常用 CNN 网络的输出为特征向量，因此与其他常用 CNN 网络相比本文使用的网络结构及连接方式较为简单，参数数量较少且计算量较小，从而计算速度得以加快，因此称之为轻量级深度学习网络。该网络的输入图像最小尺寸为 368 ×368 像素，当输入图像小于该尺寸时可做填充处理或对像素进行插值。

CNN-1 网络前 10 层的参数与 VGG 网络前 10 层参数一致，即由牛津大学视觉几何课题组在 ILSVRC 数据集（Russakovsky 等，2015）上训练得到，

训练过程设定初始学习率为 α_0，当精度趋于稳定时迭代更新学习率 α，更新方法使用学习率离散下降法，即

$$\alpha = \gamma \times \alpha_0 \tag{7}$$

式中，初始学习率 $\alpha_0 = 10^{-2}$，迭代因子 $\gamma = 0.1$，迭代次数上限为 $K = 3.7 \times 10^5$。

CNN-1 网络的后两层及 CNN-2 和 CNN-3 网络的参数采用 Wei 等人(2016)提出的网络训练方法，在 MPII(Max Planck Institute for Informatics)数据集(Andriluka 等,2014)上注释并训练得到，训练过程使用学习率指数减缓法，即

$$\alpha = \gamma^{fl(\iota/\iota_0)} \times \alpha_0 \tag{8}$$

式中，初始学习率 $\alpha_0 = 8 \times 10^{-5}$，迭代因子 $\gamma = 0.333$，函数 $fl(\cdot)$ 表示向下取整，t 为当前迭代次数，步长尺度 $t_0 = 1 \times 10^5$，迭代次数上限为 $K = 6 \times 10^5$。

由于视频图像中人体姿态大小尺寸不同，因此在使用图 5 所示的深度学习网络挖掘并提取特征时，借鉴尺度不变特征变换(scale-invariant feature transform，SIFT)算子(Lowe，2004)中使用图像金字塔解决尺度问题的思路，使用不同尺度的 4 层金字塔图像作为网络的输入。

2.2.2 手臂特征一致性似然函数

为了得到手臂位置的特征一致性表达，在训练 CNN-2 网络时需要注释手臂上各点的一致性标记，选择沿手臂的方向(Cao 等,2017)进行一致性标记；在使用时，CNN-2 网络输出各像素位置对应的手臂一致性概率值，计算其下臂姿态候选样本的下臂方向一致性作为手臂特征一致性似然函数。具体地，图 8(a)中向量 $\boldsymbol{AB}$ 表示肘关节点指向腕关节点的向量，x 正方向和 y 正方向如图 8(a)左上角所示。计算向量 $\boldsymbol{AB}$ 对应的单位向量记为 $\boldsymbol{e}_{AB}$，对于图形中任意一点 C，记 $\theta_C = \angle(\boldsymbol{AB}, \boldsymbol{AC})$，用式(9)中两条原则判断点 C 是否属于下臂

$$\begin{cases} 0 \leqslant |\boldsymbol{AC}| \cdot \cos\theta_C \leqslant |\boldsymbol{AB}| \\ |\boldsymbol{AC}| \cdot \sin\theta_C \leqslant \delta_l \end{cases} \tag{9}$$

式中，δ_l 为垂直于手臂方向的距离阈值。当点 C 满足式(9)中两条原则时，将训练图像的点 C 处标记为 $\boldsymbol{e}_{AB}$，不满足式(9)时则标记为零向量。这样，训练图像实际含有两通道，分别对应下臂方向向量 $\boldsymbol{e}_{AB}$ 的 x 分量和 y 分量。

在人体姿态跟踪过程中，深度学习网络为每一帧图像中的左右下臂分别计算手臂特征一致性概率图，即网络 CNN-2 的输出是 4 通道的手臂特征一致性概率图，4 个通道分别对应左下臂 x 方向和 y 方向的图像特征一致性以及右下臂 x 方向和 y 方向的图像特征一致性，如图 8(b)(c)所示。下面以右下臂为例计算手臂特征一致性似然函数，右下臂特征一致性 x 方向概率图记为 $\boldsymbol{M}_x$，y 方向记为 $\boldsymbol{M}_y$。由余弦定理知，$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}| \cdot |\boldsymbol{b}| \cdot \cos\theta$，当 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 均为单位向量时，$\boldsymbol{a} \cdot \boldsymbol{b} \in [-1, 1]$ 表达了向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的方向一致性程度。因此将手臂特征一致性似然函数定义为

(a) 手臂特征一致性的表达方式

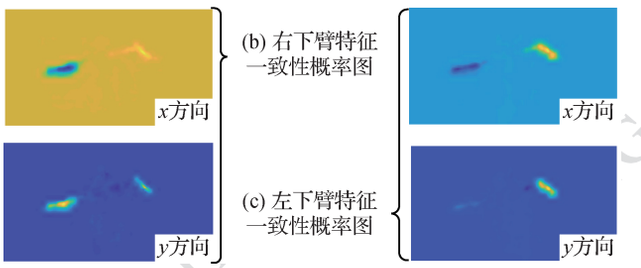


图 8 手臂特征一致概率图

Fig. 8 Probability map of arm features consistency

((a) expression of features consistency for lower arms; (b) probability graph of features consistency of right lower arm; (c) probability graph of features consistency of left lower arm)

$$p_{\mathrm{cs}}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1}) = \frac{1}{n-1} \sum_{j=1,2,\cdots,n-1} \boldsymbol{e}_{k_e k_w} \cdot \boldsymbol{M}_x(k_{c_j}) \boldsymbol{M}_y(k_{c_j}) \tag{10}$$

式中，下标 cs 代表一致性(consistency)，$\boldsymbol{e}_{k_e k_w}$ 表示手臂姿态候选样本的肘关节点 k_e 指向腕关节点 k_w 的单位方向向量

$$\boldsymbol{e}_{k_e k_w} = \frac{\boldsymbol{k}_e \boldsymbol{k}_w}{|\boldsymbol{k}_e \boldsymbol{k}_w|} \tag{11}$$

式(10)中的点 $k_{c_j}(j = 1, 2, \cdots, n-1)$ 表示向量 $\boldsymbol{k}_e \boldsymbol{k}_w$ 上的 n 等分点，如图 9 所示；$\boldsymbol{M}_x(k_{c_j})$ 和 $\boldsymbol{M}_y(k_{c_j})$ 分别表示 $\boldsymbol{M}_x$ 和 $\boldsymbol{M}_y$ 中点 k_{c_j} 位置的值，即 k_{c_j} 位置对应人体姿态图像中利用深度学习方法计算得到的所属手臂方向。似然函数 $p_{\mathrm{cs}}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1})$ 的值越大表示手臂特征一致性越好；反之越差。

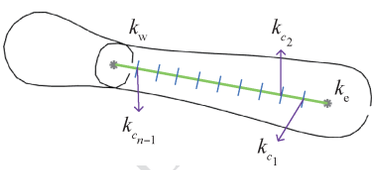


图 9　肘关节点和腕关节点连线的等分点

Fig. 9　Equal diversion point of the segment between elbow and wrist

2.2.3　生成附加下臂姿态

网络 CNN-3 的输出是 4 通道的手臂关节点概率图，4 个通道分别对应左右肘关节和左右腕关节的位置概率图，如图 10 所示。通过在对应关节位置概率图上采样，可得到左右手臂附加的姿态候选样本。新生成的手臂姿态候选样本与已有的样本一同由颜色似然函数、轮廓似然函数和手臂特征一致性似然函数联合评价。

图 10　由深度学习网络得到的手臂关节点概率图

Fig. 10　Probability graph of arm nodes obtained by deep learning network

2.3　生成人体姿态跟踪结果

通过人体部位姿态在帧间进行传播以及附加候选样本的生成，在第 $t+1$ 帧图像中会产生多个人体姿态候选样本，这些姿态样本的优劣可由颜色似然函数、轮廓似然函数以及手臂特征一致性似然函数联合评价，即

$$\begin{aligned} p_{\mathrm{ps}}(\boldsymbol{C}_i^{(q)}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1}) = {} & \lambda_{\mathrm{cl}}\cdot p_{\mathrm{cl}}(\boldsymbol{C}_i^{(q)}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1}) + \\ & \lambda_{\mathrm{ct}}\cdot p_{\mathrm{ct}}(\boldsymbol{C}_i^{(q)}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1}) + \\ & \lambda_{\mathrm{cs}}\cdot p_{\mathrm{cs}}(\boldsymbol{C}_i^{(q)}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1}) \end{aligned} \tag{12}$$

式中，$p_{\mathrm{cl}}(\boldsymbol{C}_i^{(q)}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1})$ 为式(5)定义的颜色似然函数，λ_{cl} 为其权重常量；$p_{\mathrm{ct}}(\boldsymbol{C}_i^{(q)}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1})$ 为式(6)定义的轮廓似然函数，λ_{ct} 为其权重常量；$p_{\mathrm{cs}}(\boldsymbol{C}_i^{(q)}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1})$ 为式(10)定义的手臂特征一致性似然函数，λ_{cs} 为其权重常量。

得到每个人体姿态部位最优样本后，将人体各部位连接为完整的人体姿态，连接顺序为先将上臂与下臂连接起来，头部与躯干连接起来，再将手臂连接到躯干。由于在跟踪过程中每个人体部位会产生多个姿态样本，在连接过程中会出现属于不同人体部位的同一关节位置不一致的情况。因此，定义连接上下臂的肘关节为上下臂延长线的交点；连接各部位与躯干是选择相应关节的中点。连接过程中出现的异性人体姿态可由木偶模型的形状约束将其剔除，最终获得的最优姿态样本由所连接的各部位姿态代价函数之和 $p_{\mathrm{ps}}(\boldsymbol{C}^{(q)}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1})$ 作为判别依据，即

$$p_{\mathrm{ps}}(\boldsymbol{C}^{j}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1}) = \sum_{i=1}^{6} p_{\mathrm{ps}}(\boldsymbol{C}_i^{j}\mid \boldsymbol{I}_t,\boldsymbol{I}_{t+1}) \tag{13}$$

最后，计算得到的最优人体姿态样本作为该帧人体姿态跟踪结果。

3　实验结果与分析

3.1　验证数据集及验证实验

采用两个公开的具有挑战的人体姿态数据集对提出的人体姿态跟踪算法进行验证，分别为美国宾夕法尼亚大学发布的 VideoPose2.0 和英国牛津大学发布的 YouTubePose 数据集。

VideoPose2.0 数据集是由宾夕法尼亚大学计算机与信息科学学院的学者从两个热播美剧《Friends》和《Lost》中提取的(Sapp 等，2011)。该数据集有 26 段人体姿态视频，包含手臂快速移动、身体旋转、与他人互动等具有挑战性的姿态和动作。YouTubePose 数据集是牛津大学与利兹大学的学者共同从 YouTube 网站中选取的具有挑战性的人体姿态视频(Charles 等，2016)，共有 50 段视频，涵盖了多种人体姿态及动作，如舞蹈、演讲、体育运动等。

所使用的两个数据集均注释了每帧中人体姿态的关节点位置，记为 k_i^{*t}。在数据集上对本文人体姿态跟踪算法进行验证时，将每帧中通过算法计算得到的人体姿态关节点坐标记为 k_i^t。计算跟踪得到的关节点与真实关节点之间的偏差，当偏差小于阈值 δ_a 时，将此结果视为正确，否则视为错误

$$\begin{cases} \text{正确} & \|k_i^t - k_i^{*t}\|_2 < \delta_a \\ \text{错误} & \text{其他} \end{cases} \tag{14}$$

式中,阈值 δ_a 取 20 像素。记录视频中每个关节点跟踪正确的个数,记为 n_r^k,其占视频图像总帧数 n_c^k 的百分比即为针对该关节的人体姿态跟踪精度,即

$$s_k = \frac{n_r^k}{n_c^k} \times 100\% \tag{15}$$

使用本文算法在数据集 VideoPose2.0 与 YouTubePose 上分别进行验证,实验结果分别如表 1 和表 2 所示。由于手臂部位的跟踪是人体姿态跟踪中最困难的部分,因此,表 1、表 2 中着重列出了与手臂相关的肩关节、肘关节以及腕关节的跟踪精度。此外,为验证本文算法的有效性,将 VideoPose2.0 数据集的实验结果与 3 种先进的人体姿态跟踪和估计算法的效果进行对照,这 3 种算法分别为:基于条件随机场模型的人体姿态跟踪方法(Kumar 和 Batra,2016);基于树结构图模型推理的人体姿态跟踪方法(Samanta 和 Chanda,2016);使用最大边界马尔可夫模型的人体姿态跟踪方法(Zhao 等,2015)。同样地,在 YouTubePose 数据集上进行验证的实验结果与另外 3 种先进的人体姿态跟踪和估计算法的效果进行对照,这 3 种算法分别为:基于流动卷积模型的人体姿态跟踪方法(Pfister 等,2015);基于人体部位模型成对推理的人体姿态跟踪方法(Chen 等,2018);基于人体模型序列混合推理的人体姿态跟踪方法(Cherian 等,2014)。

表 1 VideoPose2.0 数据集中人体姿态跟踪的关节点精度
Table 1 Key points accuracy of human pose tracking on VideoPose2.0 dataset

/%

算法	肩关节	肘关节	腕关节	平均值
Kumar 和 Batra(2016)	88.0	56.1	54.2	66.1
Samanta 和 Chanda(2016)	92.1	80.2	60.3	77.5
Zhao 等人(2015)	**95.0**	72.3	50.5	72.6
本文	90.5	**82.6**	**71.2**	**81.4**

注:加粗字体表示最优结果。

从表 1 中可以看到,使用本文算法在 VideoPose2.0 数据集中能够有效提高下臂关节点的跟踪精度,肘关节的跟踪精度可达到 82.6%,腕关节的跟踪精度可达到 71.2%。

与 Kumar 和 Batra(2016)的基于条件随机场的方法相比,本文算法能够更好地跟踪各关节,如表 1 所示,关节点跟踪精度平均提高了15.3%。原因是

表 2 YouTubePose 数据集中人体姿态跟踪的关节点精度
Table 2 Key points accuracy of human pose tracking on YouTubePose dataset

/%

算法	肩关节	肘关节	腕关节	平均值
Pfister 等人(2015)	82.7	70.7	59.0	70.8
Chen 等人(2018)	**90.7**	83.1	76.5	83.4
Cherian 等人(2014)	84.7	66.9	54.3	68.6
本文	86.2	**84.8**	**81.6**	**84.5**

注:加粗字体表示最优结果。

Kumar 和 Batra(2016)的方法过多依赖于候选样本的优劣,不能生成额外的候选样本,只能使用条件随机场对预先产生的姿态样本进行推理并生成轨迹跟踪。而本文算法在跟踪过程中边跟踪边检测,利用深度学习网络不断生成新的部位姿态候选样本,因此跟踪效果更好。

与 Samanta 和 Chanda(2016)的基于树结构图模型推理的算法相比,本文算法对肘关节的跟踪精度提高了 2.4%,腕关节提高了约 11%,这是由于 Samanta 和 Chanda(2016)的方法虽然也将人体姿态细化为人体关节点进行跟踪,但在特征选择方面仅使用了梯度直方图,不能针对性地描述人体姿态部位。而本文算法构造了颜色似然函数和轮廓似然函数,并使用深度学习网络挖掘深层次手臂特征,因此对下臂定位关节的定位精度更高。

Zhao 等人(2015)的基于马尔可夫模型的算法虽然也考虑了时间信息与空间信息的结合,但其方法使用的是全局时空信息,不如本文对人体部位进行计算的方法更具有针对性。该方法对活动幅度较小的人体躯干部位能够较好地跟踪预测,但依赖马尔可夫模型对活动灵活无规律的下臂姿态无法充分预测。而本文算法能够更有效地结合时空信息实现一边跟踪一边检测,弥补了马尔可夫模型预测算法的不足,从表 1 可知,本文算法得到的肘关节与腕关节定位精度分别提高了 10.3% 与 20.7%。

值得一提的是,与表 1 中 Samanta 和 Chanda(2016)和 Zhao 等人(2015)以及表 2 中 Chen 等人(2018)的方法相比,本文算法虽然在肘关节与腕关节的跟踪精度上占优势,但是肩关节的跟踪精度略有不足,原因是本文算法的目的是强化灵活运动的下臂的跟踪效果,而没有针对肩关节生成附加的上

臂姿态样本。本文算法对 VideoPose2.0 数据集的人体头部姿态跟踪精度为 90.8%，躯干姿态跟踪精度为 94.2%。

表 2 中给出了在 YouTubePose 数据集中使用本文算法得到的人体姿态精度，可以看到肘关节的跟踪精度为 84.8%，腕关节的跟踪精度可达到 81.6%，能够有效跟踪 YouTubePose 数据集中的人体姿态。

Pfister 等人(2015)的方法也用到了深度学习网络，该方法针对人体部位各个关节点构造并训练深度学习网络，从而在每幅图像中计算人体关节点的位置概率图，由于仅用深度学习网络计算得到的特征会由于欠学习等原因导致失败，因此用光流运动将前后多帧的位置概率图加权平均，得到统计的最优关节点位置。与之相比，本文使用专门构造的颜色似然函数和轮廓似然函数与深度学习特征构成互补，使提取特征的种类更加丰富，因此得到的人体姿态跟踪精度更高。从表 2 中可见，关节点的平均定位精度提高了 13.7%。

Chen 等人(2018)的方法与本文算法相比，肘关节定位精度相差 1.7%，腕关节相差 5.1%。该方法也使用了深度学习网络，但将深度学习网络用作分类器，用来将相邻两部位之间的成对连接关系分为 36 种，然后根据各关节之间的成对关系进而推理出完整的人体姿态。图像中形状较为固定的躯干、头部及上臂之间的连接可近似总结为简单的几种情况，然而下臂运动过于灵活，用分类器难以概括其运动。而本文使用深度学习网络生成的是概率图而不是分类器分类结果，因此能够涵盖并推理更多的下臂姿态。

本文算法与 Cherian 等人(2014)的方法相比，人体姿态关节点的平均跟踪精度提高了 15.7%，原因是该方法将预先生成的海量人体姿态拆分为部位姿态并在整个视频内混合推理，得到每帧中重组的新的人体姿态，因此只能从预生成的姿态中挑选出较优的候选姿态样本。而本文算法能够有针对性地额外生成新样本，且本文使用的代价函数充分利用了视频中的时空信息及深度学习信息，因此获得的人体姿态跟踪效果更好。

本文算法对 YouTubePose 数据集的人体头部姿态跟踪精度为 88.4%，躯干姿态跟踪精度为 91.3%。

表 3 中给出了本文算法与 3 种先进人体姿态跟踪算法的计算速度比较。由于算法实现的耗时与计算机主频有关，因此表 3 中亦列出了每种算法实现所用的计算机主频。

表 3　算法计算速度比较

Table 3　Algorithm calculation speed comparison

算法	Cherian 等人(2014)	Zhao 等人(2015)	Pfister 等人(2015)	本文
速度/(s/帧)	3.2	2.5	2	**1.8**
主频/Hz	3.6	3.47	2.4	2.9

注：加粗字体表示最优结果。

从表 3 中可以看到，Cherian 等人(2014)的方法用时较长，每帧人体姿态计算需用时 3.2 s，原因是该方法需要对大量候选样本进行计算，并在视频帧之间往复推理，因此算法较为复杂，耗时较长。Zhao 等人(2015)的方法每帧总用时约 2.5 s，原因是该算法对时间信息和空间信息分别构建马尔可夫随机场并进行计算和推理，计算量较大。Pfister 等人(2015)的方法中计算深度学习特征的耗时为 0.06 s，然而该方法每一帧的推理需要用到前后多帧的信息及光流图，因此总计算速度约在 2 s 左右。而本文算法在对每帧进行推理时只用到当前光流信息和相邻两帧图像，计算量较小，且所构造的深度学习网络连接简单、参数较少，因此总计算速度约为 1.8 s，优于其他算法。

3.2　算法有效性分析

本文提出的人体姿态跟踪算法主要有两个关键点：第 1 个关键点是本文提出的手臂一致性概率图的构造及使用方法；第 2 个关键点是利用深度学习算法得到的下臂附加样本概率图的使用。下面将对这两个算法关键点的有效性进行验证和分析。

3.2.1　手臂一致性概率图的有效性分析

为验证本文提出的手臂一致性概率图的有效性，在评价人体姿态候选样本时剔除手臂特征一致性似然函数，即仅使用颜色似然函数和轮廓似然函数，记为 $p_{\mathrm{ts}}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1})$

$$p_{\mathrm{ts}}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1}) = \lambda_{\mathrm{cl}} \cdot p_{\mathrm{cl}}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1}) + \lambda_{\mathrm{ct}} \cdot p_{\mathrm{ct}}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1}) \tag{16}$$

式中，$p_{\mathrm{cl}}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1})$ 为式(5)定义的颜色似然函数；$p_{\mathrm{ct}}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1})$ 为式(6)定义的轮廓似然

函数。

使用 $p_{ts}(\boldsymbol{C}_i^{(q)} \mid \boldsymbol{I}_t, \boldsymbol{I}_{t+1})$ 作为评价人体姿态候选样本的代价函数，并使用 VideoPose2.0 数据集进行对照实验，实验结果对比如表 4 所示。

表 4　算法有效性(精度)对照实验

Table 4　Control experiments of algorithm effectiveness (accuracy)

/%

算法	肘关节	腕关节
不使用手臂一致性概率图	76	66
不进行下臂附加采样	52	40
本文	**82.6**	**71.2**

注：加粗字体表示最优结果。

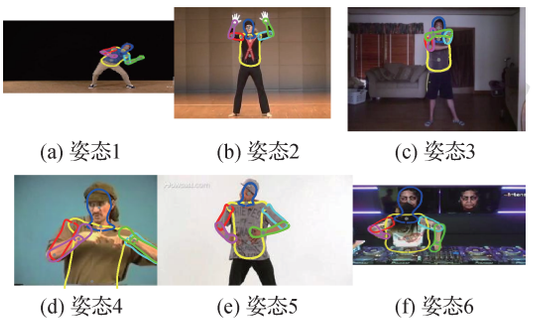
表 4 中第 3 行实验为使用本文算法对人体姿态进行跟踪的结果；表 4 中第 1 行实验方法为评价下臂姿态候选样本时不使用手臂一致性概率图计算手臂特征一致性似然函数，即使用式(16)替换式(12)得到的结果。从表 4 中可见，在使用手臂一致性概率图时得到的肘关节点跟踪精度比不使用时提高了 6.6%，腕关节的跟踪精度提高了 5.2%。其原因是在不使用手臂一致性概率图时，仅用轮廓似然函数和颜色似然函数进行评价，很多情况下会出现歧义，例如手臂与躯干服装颜色一致且带纹理时，在躯干区域误估计的手臂位置与真实手臂位置的优劣无法有效评价。在这种情况下通过深度学习算法得到的手臂特征一致性似然函数就能够将其区分，从而提高人体姿态的跟踪精度。

3.2.2　生成附加下臂姿态样本的有效性分析

针对下臂姿态运动灵活难以跟踪的问题，算法使用深度学习网络从图像中提取出下臂的肘关节点与腕关节点的位置概率图，并从中采样得到新的附加下臂姿态。为验证这一深度学习算法产生的概率图的有效性，对照实验中不使用手臂关节点位置概率图进行下臂附加采样，所得结果如表 4 第 2 行所示。对比表 4 第 2 行与第 3 行的实验结果可见，利用深度学习算法得到的下臂附加样本概率图的使用能够使肘关节的跟踪精度提高 30.6%，腕关节的跟踪精度提高 31.2%。这是由于人体下臂运动灵活且运动幅度大，在跟踪过程中易跟丢，而附加下臂姿态样本的生成使算法能够实现边跟踪边检测，有效弥补下臂跟丢时产生的误差。对照实验表明本文提出的利用深度学习算法采样下臂附加样本的关键算法能够有效提高人体姿态的跟踪精度。

3.3　实验结果展示

图 11 中给出了使用本文算法对人体姿态进行估计得到的结果实例。从图 11(a)(b)(d)(e)可看到，在人体姿态左右摇摆、向上举手以及手臂交叉在身体前侧的情况下均能正确识别出人体姿态，且效果较好。图 11(c)(f)给出了本文算法发生人体姿态误估计的例子，其中图 11(c)人体的手臂及手掌在胸前摆出了特殊的“口”字形，左右手臂与手掌相互连接且相互呈 90°，本文算法虽然能够准确定位腕关节及肘关节的位置，却在逻辑连接时由于先验知识不足而发生错误。图 11(f)中发生人体姿态误估计的原因是前景人体头部被遮挡而背景的头部恰巧出现在被遮挡的头部位置，因此算法误将背景头部与前景肢体相连接。



(a) 姿态1　(b) 姿态2　(c) 姿态3

(d) 姿态4　(e) 姿态5　(f) 姿态6

图 11　实验结果展示

Fig. 11　Demonstration of experimental results

((a) pose 1; (b) pose 2; (c) pose 3; (d) pose 4; (e) pose 5; (f) pose 6)

4　结　论

针对复杂铰接人体姿态目标在视频中难以用传统视觉跟踪方法进行跟踪，特别是活动灵活的手臂跟踪精度较低的问题，本文提出一种基于可解耦人体姿态模型的边跟踪边检测的人体姿态跟踪方法。该方法首次将视频运动信息、空间信息与深度学习信息相结合，利用运动信息实现视频中的人体姿态传递，保证了帧间人体姿态的连续性；构造人体姿态各部位的颜色似然函数，对跟踪过程中人体姿态目标的颜色一致性进行评价；构造人体姿态各部位的轮廓似然函数，对跟踪得到的人体部位轮廓与图像中图形轮廓的拟合程度进行评价；针对人体姿态中

灵活运动的手臂位置难以检测的问题，构造并训练轻量级的深度学习网络，挖掘图像中深层次的手臂一致性信息，从而优化手臂姿态的跟踪效果。通过使用两个公开的具有挑战性的人体姿态跟踪数据集对该方法进行实验验证和算法有效性分析，证明了本文提出的人体姿态跟踪方法能够有效提高人体姿态关节点的跟踪精度，特别是对灵活运动的手臂关节点的跟踪精度有显著提高。

未来本文的研究成果可扩展到智能家居场景中人体姿态的跟踪以及对人体危险姿态的识别，提高助老助残机器人及家居陪伴机器人的人机交互能力。

参考文献(References)

Andriluka M, Pishchulin L, Gehler P and Schiele B. 2014. 2D human pose estimation: new benchmark and state of the art analysis//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 3686-3693 [DOI: 10.1109/CVPR.2014.471]

Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J and Davis J. 2005. SCAPE: shape completion and animation of people. ACM Transactions on Graphics, 24(3): 408-416 [DOI: 10.1145/1073204.1073207]

Bajcsy R, Aloimonos Y and Tsotsos J K. 2018. Revisiting active perception. Autonomous Robots, 42(2): 177-196 [DOI: 10.1007/s10514-017-9615-3]

Cao Z, Simon T, Wei S E and Sheikh Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 7291-7299 [DOI: 10.1109/CVPR.2017.143]

Charles J, Pfister T, Magee D, Hogg D and Zisserman A. 2016. Personalizing human video pose estimation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 3063-3072 [DOI: 10.1109/CVPR.2016.334]

Chen Y L, Wang Z C, Peng Y X, Zhang Z Q, Yu G and Sun J. 2018. Cascaded pyramid network for multi-person pose estimation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 7103-7112 [DOI: 10.1109/CVPR.2018.00742]

Cherian A, Mairal J, Alahari K and Schmid C. 2014. Mixing body-part sequences for human pose estimation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 2353-2360 [DOI: 10.1109/CVPR.2014.302]

Duckworth P, Hogg D C and Cohn A G. 2019. Unsupervised human activity analysis for intelligent mobile robots. Artificial Intelligence, 270: 67-92 [DOI: 10.1016/j.artint.2018.12.005]

Fischier M A and Elschlager R A. 1973. The representation and matching of pictorial structures. IEEE Transactions on Computers, C-22(1): 67-92 [DOI: 10.1109/T-C.1973.223602]

Fragkiadaki K, Hu H and Shi J B. 2013. Pose from flow and flow from pose//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE: 2059-2066 [DOI: 10.1109/CVPR.2013.268]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Krizhevsky A, Sutskever I and Hinton G E. 2012. Imagenet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing System. Red Hook, NY: ACM: 1097-1105

Kumar R and Batra D. 2016. Pose tracking by efficiently exploiting global features//Proceedings of 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Placid, New York: IEEE: 1-9 [DOI: 10.1109/WACV.2016.7477563]

Liu C. 2009. Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. Cambridge, MA: Massachusetts Institute of Technology

López-Quintero M I, Marín-Jiménez M J, Muñoz-Salinas R and Medina-Carnicer R. 2017. Mixing body-parts model for 2D human pose estimation in stereo videos. IET Computer Vision, 11(6): 426-433 [DOI: 10.1049/iet-cvi.2016.0249]

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]

Ma M, Marturi N, Li Y B, Stolkin R and Leonardis A. 2016. A local-global coupled-layer puppet model for robust online human pose tracking. Computer Vision and Image Understanding, 153: 163-178 [DOI: 10.1016/j.cviu.2016.08.010]

Newell A, Yang K Y and Deng J. 2016. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer: 483-499 [DOI: 10.1007/978-3-319-46484-8_29]

Pfister T, Charles J and Zisserman A. 2015. Flowing ConvNets for human pose estimation in videos//Proceedings of 2015 International Conference on Computer Vision. Santiago: IEEE: 1913-1921 [DOI: 10.1109/ICCV.2015.222]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F.

2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Samanta S and Chanda B. 2016. A data-driven approach for human pose tracking based on spatio-temporal pictorial structure [EB/OL]. [2019-08-22]. https://arxiv.org/pdf/1608.00199.pdf

Sapp B, Weiss D and Taskar B. 2011. Parsing human motion with stretchable models//Proceedings of CVPR 2011. Providence: IEEE: 1281-1288 [DOI: 10.1109/CVPR.2011.5995607]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2019-08-22]. https://arxiv.org/pdf/1409.1556.pdf

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deepen with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Wei S E, Ramakrishna V, Kanade T and Sheikh Y. 2016. Convolutional pose machines//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 4724-4732 [DOI: 10.1109/CVPR.2016.511]

Xu Y X and Chen F. 2015. Recent advances in local image descriptor. Journal of Image and Graphics, 20(9): 1133-1150 (许允喜, 陈方. 2015. 局部图像描述符最新研究进展. 中国图象图形学报, 20(9): 1133-1150) [DOI: 10.11834/jig.20150901]

Yang Y and Ramanan D. 2012. Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12): 2878-2890 [DOI: 10.1109/TPAMI.2012.261]

Zhao L, Gao X B, Tao D C and Li X L. 2015. Tracking human pose using max-margin markov models. IEEE Transactions on Image Processing, 24(12): 5274-5287 [DOI: 10.1109/TIP.2015.2473662]

Zheng L, Huang Y J, Lu H C and Yang Y. 2019. Pose-invariant embedding for deep person re-identification. IEEE Transactions on Image Processing, 28(9): 4500-4509 [DOI: 10.1109/TIP.2019.2910414]

Zuffi S, Freifeld O and Black M J. 2012. From pictorial structures to deformable structures//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE: 3546-3553 [DOI: 10.1109/CVPR.2012.6248098]

Zuffi S, Romero J, Schmid C and Black M J. 2013. Estimating human pose with flowing puppets//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney: IEEE: 3312-3319 [DOI: 10.1109/ICCV.2013.411]

作者简介

马淼,1989 年生,女,讲师,主要研究方向为模式识别、图像处理、计算机视觉。E-mail: mamiao@zstu.edu.cn

李贻斌,男,教授,主要研究方向为智能机器人、特种机器人、人机交互。E-mail: liyb@sdu.edu.cn

武宪青,男,讲师,主要研究方向为非线性控制、智能控制、信号处理。E-mail: wxq@zstu.edu.cn

高金凤,女,教授,主要研究方向为多智能体系统、多机器人协作、人机交互。E-mail: jfgao@zstu.edu.cn

潘海鹏,男,教授,主要研究方向为智能检测、控制、图像信息处理。E-mail: pan@zstu.edu.cn