[文章编号] 1003－4684(2020)04-0006-05

基于改进自顶向下的行人运动预测方法

王淑青，刘逸凡

（湖北工业大学电气与电子工程学院，湖北 武汉 430068）

[摘 要] 针对传统的在复杂环境中多人姿态检测的自顶向下算法存在依赖人体框检测而导致的单人姿态检测错误的现象，提出一种基于改进的自顶向下的多人姿态检测并应用于行人运动预测的方法。该算法通过处理空间变换网络、单人姿势检测、反空间变换网络处理后的图像提取人体的骨点和姿势，经光流处理和长短期记忆神经网络的训练，预测行人接下来的动作。所述算法与 6 种经典多人姿态检测算法对比分析，实验结果表明：该方法得到的多人姿态检测图像准确，无冗杂人体框，无冗杂骨点，无骨骼交叉情况，行人运动预测效果良好。

[关键词] 自顶向下法；空间变换网络；单人姿势检测；光流处理

[中图分类号] TP242.2 [文献标识码] A

行人检测技术广泛应用于先进辅助驾驶领域，在该领域中，行人检测往往在环境条件好，行人之间没有重合的情况下检测效果明显，对复数情况下的行人检测一直以来是重要的研究内容[1]，本文提出了一种基于改进后的自顶向下法的多人姿势检测的算法并应用于一种实时的人体运动的预测。

识别图片中的单人姿态检测要比多人姿态检测容易得多[2]，而且识别图片中的行人姿态要比识别复杂环境中的人体姿态容易[3-4]。目前有两种主流的多人姿态检测方法：自顶向下法（Two-step framework）[5-7]和自底向上法（Part-based framework）[7-8]，自顶向下法是先检测环境中的每一个人体检测框，然后独立地去检测每一个人体边界的姿态，但这种方法极度依赖于姿态检测准确度，并且由于冗余的检测框也可能重复估计单人的边界框；自顶向上法则是首先检测出环境中的所有肢体节点，然后进行拼接得到多人的骨架，但由于这种方法取决于人的肢体节点，在两人离得非常近时容易出现错误连接。

为了克服这些错误，Toshev 等人提出的 DeepPose[9]、W. Ouyang 等人提出的基于深度神经网络[10]以及 A. Jain 等提出的基于卷积神经网络的人体识别算法可以简单地检测人体姿势[2-11]，J. Dong 等人还同时考虑了人体解析与姿势检测[12]，但上述算法极易出现错误的边界框，而且冗余的边界框会产生冗余的姿态，但在定位和认知方面的小误差是不可避免的，这些误差会导致姿态检查的错误，尤其是仅依赖于人体检测结果的方法，不能满足当下辅助驾驶系统的要求。X. Chen 等人针对行人被遮挡的情况提出一种将行人看作不同身体部位组合的模型[7]。G. Gkioxari 等人使用 K-poselets 检测行人并预测行人姿势的位置，计算所有姿势的加权平均值来预测最终姿势定位[5]。L. Pishchulin 等人提出先检测行人所有身体部件，再利用深度学习依次分割各个部件，通过积分线性编程对这些部件进行标记和组装[11-13]。E. Insafutdinov 等人提出一种基于 ResNet 的更强行人身体部位检测方法和更好的增量优化策略[6-8]，但只适用于较小的局部范围，应用性不强。

本文研究一种基于改进自顶向下法的多人姿势检测并应用于行人姿势预测方法。首先，将输入的视频提取一帧图片做人体边界检测处理得到不准确的人体边界框；然后，利用改进后的自顶向下法提取精确的骨点、人体边界框以及人体姿势；最后采用光流处理后经长短期记忆神经网络训练出 0.5 s 之后的行人运动姿势。

1 自顶向下法

Zhang 等[14]提出的自顶向下法又称两步法（Two-step framework），第一步，对目标人体的多个关节点进行检测，利用一个新的梯形仿射不变量，推导出相关的平行四边形具有这种仿射不变量。第

[收稿日期] 2019－07－28

[基金项目] 国家自然科学基金青年基金项目(61603127)

[第一作者] 王淑青(1969－)，女，河北衡水人，理学博士，湖北工业大学教授，研究方向为智能检测与控制，系统分析与集成

二步,将第一步得到的结果作为迭代过程的初始值,利用高斯—牛顿法对初始值进行修正,通过深度估计产生的 7 个误差函数和 4 个特征点的共平面性,建立了误差矩阵。

两步法改进了线性法对噪声的敏感程度、给出了第二步迭代的一个较好的初值并且该算法对每一帧图像分别进行处理,从而使图像的精度达到了预期的水平,消除了计算误差的相关性。

1.1　姿势识别

基于自顶向下法的行人姿势识别质量取决于人体边界框判定的精确程度,为保证姿势识别在计算高精确的人体边界判定框的同时具有高质量的姿势定位,本文采用一种基于双空间网络变换并单人姿势识别的改进自顶向下法来提高人体边界判定框的精度与姿势定位的质量。

1.1.1　人体边界判定框预处理　人体边界判定框是姿势识别的基石,是决定姿势识别质量的一个重要因素。在一定范围内,人体边界判定框越小,人体定位就越精确,但是,现有的算法为了避免人体关节的丢失,对判定框的计算都比较保守。为了减小人体边界判定框大小的同时也能保留易丢失的人体关节,本文采用 Liu 等[15]提出的基于回归法的判定物体的类别和位置,其计算过程为:

$$b=\begin{pmatrix}b^{cx}\\b^{cy}\\b^{w}\\b^{h}\end{pmatrix}=\begin{pmatrix}d^{w}l^{cx}+d^{cx}\\d^{y}l^{cy}+d^{cy}\\d^{w}\exp(l^{w})\\d^{h}\exp(l^{h})\end{pmatrix}$$

其中:$d=(d^{cx},d^{cy},d^{w},d^{h})$ 表示先验框的位置,其对应边界框的位置用 b 表示;(cx,cy) 表示边界框的中心坐标;w, h 分别表示边界框的宽与高;l 表示边界框的预测值,其中:

$$l^{cx}=\frac{(b^{cx}-d^{cx})}{d^{w}},\ l^{cy}=\frac{(b^{cy}-d^{cy})}{d^{h}}$$

$$l^{w}=\log\frac{b^{w}}{d^{w}},\ l^{h}=\log\frac{b^{h}}{d^{h}}$$

1.1.2　双空间网络变换并单人姿势识别　前文中使用的回归法以及现有的人体边界判定框算法中,为了提高判定框的准确率,过多地增加了判定框的数量,造成了大量冗杂的判定框的产生。空间网络变换(STN)可以通过变换输入的图片,降低受到数据在空间上多样性的影响,来提高卷积网络模型的分类准确率,而不是通过改变网络结构。为了避免人体边界判定框普遍存在的冗杂问题,本文采用双空间网络变换并单人姿势识别(SPPE)的自顶向下法来过滤冗杂问题,其计算过程为:

$$\begin{pmatrix}x\\y\end{pmatrix}=(\theta_1\theta_2\theta_3)\begin{pmatrix}b^{cx}\\b^{cy}\\1\end{pmatrix}$$

其中:$\begin{pmatrix}b^{cx}\\b^{cy}\\1\end{pmatrix}$ 为空间变换网络前的人体边界判定框的坐标;$\begin{pmatrix}x\\y\end{pmatrix}$ 为空间变换网络后的人体边界判定框的坐标;$(\theta_1,\theta_2,\theta_3)$ 均为二维空间的向量。

单人姿势检测的算法有很多,常用的有 CNN、DeepPose、OpenPose。本文采用精确度较高的 CNN 单人姿势检测[16],对每一个高质量人体区域框 $\begin{pmatrix}x_i\\y_i\end{pmatrix}$ 采用 CNN 单人姿势检测得到有冗杂的骨点置信度 E,置信度越高,则越有可能是正确的人体骨点,其计算过程为:

$$E=\int_{u=0}^{u=1}L_c(p(u))\cdot\frac{d_{j2}-d_{j1}}{\|d_{j2}-d_{j1}\|_2}\mathrm{d}u$$

其中:d_{j1}、d_{j2}分别为两个骨点的位置;L_c 为两骨点组成的线段,$p(u)=(1-u)d_{j1}+ud_{j2}$ 为两个骨点 d_{j1}、d_{j2}之间的插值。

为了避免精确度较高的 CNN 单人姿势检测算法以及现有的单人检测算法存在的冗杂骨点数量过多的问题,本文对冗杂骨点进行第二步置信处理:选取最大置信度的骨点 E_{max} 作为参考,定义 η 为标准的阈值,对其消除离得比较近且相似的的骨点(d_{j1}、d_{j2}),则:

$$E(d_i,d_j)=\begin{cases}1,\text{若 } d(d_i,d_j)>\eta\\0,\text{若 } d(d_i,d_j)\leqslant\eta\end{cases}$$

若 $E(d_i,d_j)$ 输出为 1,则表示骨点 d_i 是冗杂的,应被消除;若 $E(d_i,d_j)$ 输出为 0,则表示骨点 d_j 是冗杂的,应被消除。

在单人姿势识别之后,得到的姿势被映射到空间网络变换之后的图像 $\begin{pmatrix}x\\y\end{pmatrix}$,自然的,单人姿势识别估计的人体姿态应重新映射回原始图像坐标,即进行第二次空间网络变换,同时第二次空间网络变换解除变换 γ_i,并基于 γ_i 生成网络,计算过程为:

$$\begin{pmatrix}x_i\\y_i\end{pmatrix}=(\gamma_1\gamma_2\gamma_3)\begin{pmatrix}x\\y\end{pmatrix}$$

其中,$\begin{pmatrix}x_i\\y_i\end{pmatrix}$ 为原图坐标中的高质量人体边界判定框,并且

$$[\gamma_1\gamma_2]=[\theta_1\theta_2]^{-1}$$

$$\gamma_3=-1\times[\gamma_1\gamma_2]\theta_3$$

其中,$[\theta_1\theta_2]$ 的计算方法为:

$$\frac{\partial J(W,b)}{\partial[\theta_1\theta_2]}=\frac{\partial J(W,b)}{\partial[\theta_1\theta_2]}=\frac{\partial J(W,b)}{\partial[\gamma_1\gamma_2]}\times\frac{\partial[\gamma_1\gamma_2]}{\partial[\theta_1\theta_2]}+\frac{\partial J(W,b)}{\partial\gamma_3}\times\frac{\partial\gamma_3}{\partial[\gamma_1\gamma_2]}\times\frac{\partial[\gamma_1\gamma_2]}{\partial[\theta_1\theta_2]}$$

θ_3 的计算方法为：

$$\frac{\partial J(W,b)}{\partial \theta_3}=\frac{\partial J(W,b)}{\partial \gamma_3}\times\frac{\partial \gamma_3}{\partial \theta_3}$$

1.2 光流处理

光流检测用来描述相对于观察者的运动所造成的观测目标、表面或边缘的运动[17]。用本文所述方法对原视频连续 5 帧(当前帧与前 4 帧)分别进行处理，得到的高精度的骨点 d_i 以及高质量的人体边界判定框 $\begin{pmatrix}x_i\\y_i\end{pmatrix}$，$i=1,2,3,4,5$。

本文对这 5 张图片进行光流处理，得到各个骨点的位移向量 $\varepsilon(d)$，计算过程为：

$$\varepsilon(d)=\varepsilon(d_x,d_y)=\sum_{x=u_x-w_x}^{u_x+w_x}\sum_{y=u_y-w_y}^{u_y+w_y}\cdot(I(x,y)-J(x+d_x,y+d_y))^2$$

$$v=u+d=[u_x+d_x u_y+d_y]^{\mathrm{T}}$$

其中 v 为该骨点在下一帧的新位置，u 表示该骨点所在的位置。

由于 LSTM 神经网络适合于处理和预测时间序列中间隔和延迟非常长的重要事件[18]，本文将上述五个连续帧 $E(d_i)$ 以及光流处理后的骨点位移偏移量 $\varepsilon(d)$ 传输到长短期记忆神经网络(LSTM)进行训练模型，每进行 6 次训练即每 30 帧中提前 0.5 s生成一次骨骼框图 E_f，所述 E_f 即为实时的行人运动预测框架(图 1)。

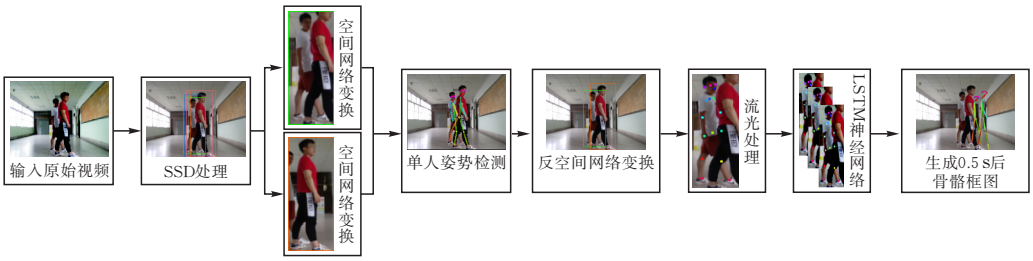

图 1 算法流程图

2 实验结果与分析

2.1 仿真平台

所有对比方法以及本文方法均基于 Ubuntu 16.04 LTS 操作系统，依赖环境 Python 3.6.5、Pytorch 0.4.1 以及 OpenCV 3.4.4，所用计算机处理器为 Intel(R) Core i7-7700HQ (2.80GHz, 64bit)，内存为 8GB。将本文算法与 Iqbal 等[19]、DeeperCut[8]、Levinkov 等[20]、Insafutdinov 等[6]、Cao 等[21]、Newell 等[22]提出的方法基于 MPII 多人数据集[23]进行对比分析。本文算法取值 $\eta=90\%$。

2.2 结果分析

MPII 多人数据集由 3844 个训练组和 1758 个测试组组成，这些测试组中既有遮挡的人，也有重叠的人。此外，它还包含超过 28 000 个单人姿势估计训练样本。本文使用单人数据集中的所有训练数据、多人数据集中的 90%用来确定单人姿势检测的调整，剩下 10%用于验证。

采用 MPII 多人数据集[23]所提供的精度计算公式，计算方法为：

$$\mathrm{mAP}=\frac{1}{|Q_R|}\sum_{q\in QR}AP(q)$$

其中：AP(Average Precision)为平均精确度；mAP 值(Mean Average Precision)为平均 AP 值，是判断行人识别精度质量情况，该评价模型的图像质量值区间为[0,100]，值越大，表示人体姿势检测质量越好。

在 MPII 多人数据集上测试了本文方法。表 1 给出了完整数据集的定量结果。本文所述方法在识别手腕、肘部、脚踝和膝盖等不同关节方面的平均准确度达到 72 mAP，比之前的最新结果高出 3.3 mAP。手腕的最终精度为 76.3 mAP，膝盖的最终精度为 79.8 mAP，本文所述方法在整体人体检测中进一步实现 82.0 mAP，比目前最佳结果高出 4.5 mAP。结果表明，该方法能准确预测多人图像中的姿态(表 1)，其中加黑数值表示当前最大值，结果可以发现，在大多数多人姿势检测中，本文方法优于其他方法，即本文方法能获得更精确的人体姿势。本测试部分结果如图 2 所示。

为了验证双空间网络变换并单人姿势识别的重要性，进行了两个对照实验。第一个实验中，从方法中移除了双空间网络变换并单人姿势识别。第二个实验中，只移除了单人姿势识别步骤，保持了双空间网络变换。这两个结果对照见表 2。在移除了单人姿势识别步骤的实验中，可以观察到多人姿势检测性能下降，证明了带有单人姿势识别的双空间网络变换很大程度使空间网络变换提取高质量的单人姿势，以最大限度地减少总冗杂。

如图 3 所示，笔者对步行动作进行了0.5 s的预测，可以证明本方法预测行走动作效果较好。

表 1　7 种方法的质量评价

	Iqbal	DeeperCut	Levinkov	Insafutdinov	Cao	Newell	Proposed
Head	58.4	78.4	89.8	88.8	91.2	**92.1**	91.3
Shoulder	53.9	72.5	85.2	87.0	87.6	89.3	**90.4**
Elbow	44.5	60.2	71.8	75.9	77.7	78.9	**83.9**
Wrist	35.0	51.0	59.6	64.9	66.8	69.8	**76.3**
Hip	42.2	57.2	71.1	74.2	75.4	76.2	**80.2**
Knee	36.7	52.0	63.0	68.8	68.9	71.6	**79.8**
Ankle	31.1	45.4	53.5	60.5	61.7	64.7	**72.3**
Total	43.1	59.5	70.6	74.3	75.6	77.5	**82.0**

图 2　部分测试结果

表 2　对照实验结果

	Full Proposed	Without SPEE	Without Double STN and SPEE		Full Proposed	Without SPEE	Without Double STN and SPEE
Head	91.3	89.0	89.0	Hip	80.2	77.8	77.1
Shoulder	90.4	88.0	86.9	Knee	79.8	74.0	73.3
Elbow	83.9	83.4	82.8	Ankle	72.3	65.8	65.0
Wrist	76.3	74.7	73.5	Total	82.0	79.1	78.2

图 3　0.5 s 后动作预测结果

3　结论

提出了一种基于改进自顶向下法多人姿态检测的行人运动预测方法,将对输入的视频提取一帧图片做人体边界检测处理得到不准确的人体边界框,利用改进后的自顶向下法提取精确的骨点、人体边界框以及人体姿势,采用光流处理后经长短期记忆神经网络训练出 0.5 s 之后的行人运动姿势。实验验证结果表明本文方法对多人姿态检测的精度准确,并且实现了对行走动作的预测。

[参 考 文 献]

[1] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016: 4724-4732.

[2] Fan X, Zheng K, Lin Y, et al. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 1347-1355.

[3] Sun M, Kohli P, Shotton J. Conditional regression forests for human pose estimation[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3394-3401.

[4] Ladicky L, Torr P H S, Zisserman A. Human pose estimation using a joint pixel-wise and part-wise formulation[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3578-3585.

[5] Gkioxari G, Hariharan B, Girshick R, et al. Using k-poselets for detecting people and localizing their keypoints[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 3582-3589.

[6] Insafutdinov E, Andriluka M, Pishchulin L, et al. Arttrack: Articulated multi-person tracking in the wild [C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 6457-6465.

[7] Chen X, Yuille A L. Parsing occluded people by flexible compositions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3945-3954.

[8] Insafutdinov E, Pishchulin L, Andres B, et al. Deepercut: A deeper, stronger, and faster multi-person pose estimation model[C]// European Conference on Computer Vision. Springer, Cham, 2016: 34-50.

[9] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 1653-1660.

[10] Yang W, Li S, Ouyang W, et al. Learning feature pyramids for human pose estimation[C]// proceedings of the IEEE international conference on computer vision, 2017: 1281-1290.

[11] Pishchulin L, Jain A, Andriluka M, et al. Articulated people detection and pose estimation: Reshaping the future[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3178-3185.

[12] Dong G, Li J. Efficient mining of emerging patterns: Discovering trends and differences[C]// Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. 1999: 43-52.

[13] Pishchulin L, Insafutdinov E, Tang S, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 4929-4937.

[14] Zhang S, Ding Y, Hao K, et al. An efficient two-step solution for vision-based pose determination of a parallel manipulator[J]. Robotics and Computer-Integrated Manufacturing, 2012, 28(2): 182-189.

[15] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]// European conference on computer vision. Springer, Cham, 2016: 21-37.

[16] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016: 4724-4732.

[17] Horn BKP, Brian G. Schunck. Determining optical flow [J]. Artificial Intelligence, 1980, 17(1-3): 185-203.

[18] Greff K, Srivastava R K, Koutník J, et al. LSTM: A search space odyssey[J]. IEEE transactions on neural networks and learning systems, 2016, 28(10): 2222-2232.

[19] Iqbal U, Gall J. Multi-person pose estimation with local joint-to-person associations[C]// European Conference on Computer Vision. Springer, Cham, 2016: 627-642.

[20] Levinkov E, Uhrig J, Tang S, et al. Joint graph decomposition & node labeling: Problem, algorithms, applications[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6012-6020.

[21] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 7291-7299.

[22] Newell A, Huang Z, Deng J. Associative embedding: End-to-end learning for joint detection and grouping [C]// Advances in neural information processing systems, 2017: 2277-2287.

[23] Andriluka M, Pishchulin L, Gehler P, et al, 2d human pose estimation: New benchmark and state of the art analysis[C]// Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014: 3686-3693.

A Pedestrian Motion Prediction Method Based on Improved Two-step Framework

WANG Shuqing, LIU Yifan

(*School of Electrical and Electronic Engin., Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: Aiming at the problem that traditional two-step framework algorithm of multi-person posture detection in complex environment relies on human frame detection, this paper proposes an improved two-step framework algorithm of multi-person posture detection method and applies it to pedestrian motion prediction. The algorithm extracts human bones and postures from images processed by spatial transformation network, single-person posture detection and anti-spatial transformation network, and predicts pedestrian's next actions through optical flow processing and training of long short-term memory network. Compared with six classical multi-person posture detection algorithms, the experimental results show that the multi-person posture detection image obtained by this method is accurate, without redundant human frame, no redundant bone points, no skeleton crossover, and the pedestrian motion prediction effect behaves well.

Keywords: two-step framework; Spatial transformation network; single person posture detection; optical flow

[责任编校: 张岩芳]