计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(4):254−259 [doi: 10.15888/j.cnki.csa.007376]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

级联层叠金字塔网络模型的服装关键点检测①

李维乾[1,2,3], 张紫云[1,2,3], 王　海[4], 张　艺[1,2,3]

[1](西安工程大学 计算机科学学院, 西安 710048)
[2](陕西省服装设计智能化重点实验室, 西安 710048)
[3](新型网络智能信息服务国家地方联合工程研究中心, 西安 710048)
[4](西北大学 信息科学与技术学院, 西安 710127)
通讯作者: 张紫云, E-mail: 365975031@qq.com

摘　要: 服装关键点的检测对服饰分类、推荐和检索效果具有重要的作用, 然而实际服装数据库中存在大量形变及背景复杂的服饰图片, 导致现有服装分类模型的识别率和服装推荐、检索的效果较差. 为此, 本文提出了一种级联层叠金字塔网络模型 CSPN (Cascaded Stacked Pyramid Network), 将目标检测方法与回归方法相结合, 首先采用 Faster R-CNN 结构对服装目标区域进行识别, 然后基于 ResNet-101 结构生成的多层级特征图, 构建级联金字塔网络, 融合服饰图像的多尺度高低层信息, 解决图片形变及复杂背景下服装关键点识别准确度不高等问题. 实验结果表明, CSPN 模型在 DeepFashion 数据集上较其他三种模型对服装关键点具有较高识别度.

关键词: 服装关键点检测; 层叠金字塔模型; Faster R-CNN; ResNet-101

引用格式: 李维乾,张紫云,王海,张艺.级联层叠金字塔网络模型的服装关键点检测.计算机系统应用,2020,29(4):254–259. http://www.c-s-a.org.cn/1003-3254/7376.html

Cascaded Stacked Pyramid Network Model for Key Point Detection of Clothing

LI Wei-Qian[1,2,3], ZHANG Zi-Yun[1,2,3], WANG Hai[4], ZHANG Yi[1,2,3]

[1](School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)
[2](Shaanxi Key Laboratory of Clothing Intelligence, Xi'an 710048, China)
[3](State and Local Joint Engineering Research Center for Advanced Networking and Intelligent Information Services, Xi'an 710048, China)
[4](School of Information Science and Technology, Northwest University, Xi'an 710127, China)

Abstract: The detection of key points of clothing plays an important role in the classification, recommendation, and retrieval of clothing. However, there are a large number of clothing pictures with deformation and complex background in the clothing database, which leads to the poor recognition rate of the existing clothing classification model and the effect of clothing recommendation and retrieval. For this reason, this study proposes a model called Cascaded Stacked Pyramid Network (CSPN) which combines the target detection method with the regression method. First, the costume target area is identified by the Faster R-CNN, and then the Cascaded Pyramid Network (CPN) is constructed based on the multi-level feature map generated by ResNet-101 structure. This model integrates the multi-scale and different-layer clothing image feature, and solves low image recognition accuracy about clothing key points of the deformation and complex background image. Experimental results show that the CSPN model has higher recognition rate on the key points of clothing than the other three models in the DeepFashion dataset.

Key words: key points detection of clothing; stacked pyramid model; Faster R-CNN; ResNet-101

① 基金项目: 国家自然科学基金 (61572401, 61672426, 61701400); 西安工程大学博士科研启动基金 (BS1330)
Foundation item: National Natural Science Foundation of China (61572401, 61672426, 61701400); Start-up Fund for Doctoral Research of Xi'an Polytechnic University (BS1330)
收稿时间: 2019-09-07; 修改时间: 2019-10-08; 采用时间: 2019-10-31; csa 在线出版时间: 2020-04-05

在电子商务应用领域, 常采用服装关键点从视觉方面来描述服装的形态信息, 从而将感兴趣的服装推荐给消费者, 进而满足客户的个性化需求. 而现有数据集中, 因服饰图片存在背景复杂、服饰变形等问题, 导致服装关键点的检测精度不高. 为此, Liu 等人[1]基于深度学习 Caffe 框架, 构建了三层级联神经网络模型, 利用深度时尚对齐技术 (Deep Fashion Alignment, DFA) 完成了服装关键点的检测研究. 但该模型未对复杂背景下的图片进行处理, 尤其是当服饰部分关键点被遮挡或服饰发生变形时, 该模型对关键点检测的准确率会降低. Yan 等[2]采用 VGG16 (Visual Geometry Group) 模型构建了 DLAN (Deep LAndmark Network) 网络, 该网络对因人体姿态不同产生变形的服装关键点进行了检测. Wang 等人[3]提出了一种双向卷积递归神经网络 (Bidirectional Convolutional Recurrent Neural Networks, BCRNNs), 根据运动学和服装对称关系来检测服装关键点. 然而 DLAN 和 BCRNNs 在对图像特征进行提取时, 仅使用了图像的深层信息, 未能融合多层网络语义信息, 导致关键点检测的准确率较低.

针对以上问题, 本文提出了一种自上而下的级联层叠金字塔网络模型 (Cascaded Stacked Pyramid Network, CSPN) 用于服装关键点的检测. 首先, 采用 Faster R-CNN[4]对服装目标区域进行定位, 以确定服装的位置, 消除图像中与服饰无关的图像特征. 然后, 利用 ResNet-101[5]提取服装区域的图像特征. 以此为基础, 将多个不同深度的特征图叠加在一起, 构成级联层叠特征金字塔[6]网络结构, 有效融合服饰图像的原始特征、全局特征和局部特征, 逐步对关键点进行预测与修正, 从而提高服装关键点检测的准确度.

1 服装关键点检测

在传统的关键点检测中, 采用全连接层直接回归坐标点[7]. 该类做法的优点是输出即为坐标点, 训练和前向速度很快, 且是端到端的全微分训练, 但缺点是缺乏空间泛化能力, 丢失了特征图上的空间信息. 此方法不利于检测被遮挡的服装关键点. 而本文采用预测高斯热图的方式, 输出特征图大, 空间泛化能力较强.

如图 1 所示, 原始图片经过金字塔网络提取特征之后, 融合高低层特征信息, 生成预测关键点热图, 计算 Loss 并反向传播.

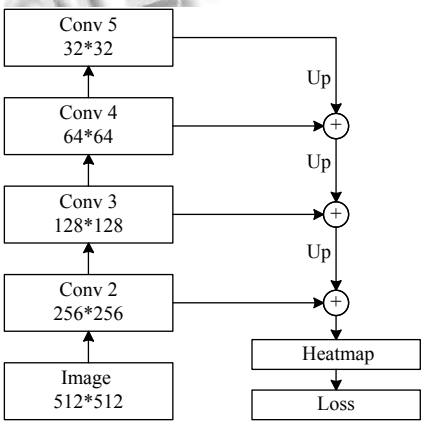

图 1 预测关键点热图并回传 Loss

2 层叠金字塔网络模型

CSPN 模型将基于目标检测和基于回归检测的方法相结合, 通过构建级联层叠特征金字塔网络模型来提高服装关键点的检测准确率. 该模型包含两个阶段, 第一阶段 Detection Proposal 是对服装区域进行识别, 第二阶段 Detection KeyPoint 是利用识别后的特征图对服装关键点检测. 图 2 为该模型对服装关键点进行检测的过程.

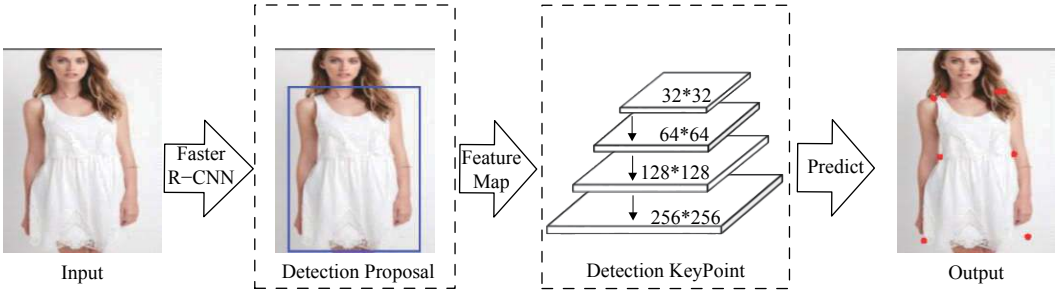

图 2 服装关键点检测过程

2.1 服装目标区域识别网络 (Detection Proposal)

针对复杂背景下关键点识别率低的问题, CSPN 模型利用 Faster R-CNN 对服装目标区域进行识别, 其流程如图 3 所示. 首先, 利用 VGG16 卷积层对服装图片数据进行特征提取, 将生成的 feature map1 特征图一方面用作服装边界框 proposals 的生成, 另一方面将该 feature map1 特征图和生成的 proposals 边界框作为 RoI (Region of Interest) pooling 层的输入, 得到固定大小的输出特征图 (feature map2). 然后, 利用分类器 (Classifier) 对固定大小的 feature map2 进行全连接操作. 最后, 使用 L1 Loss 边框回归操作, 获得服装区域的精确位置.

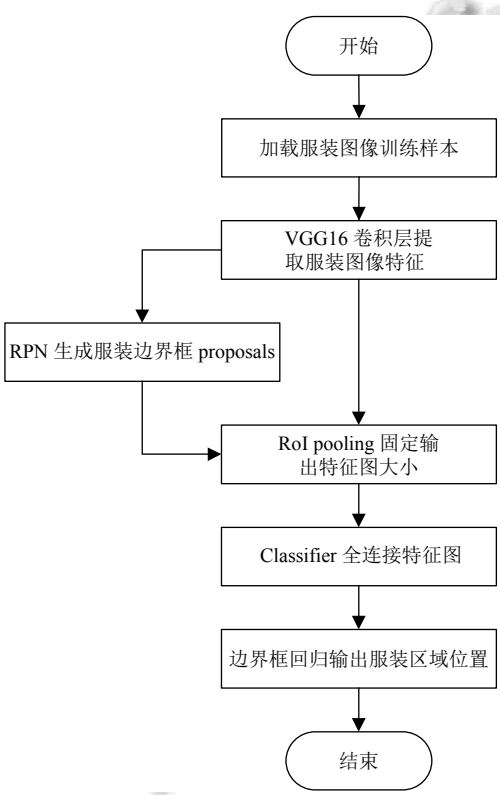

图 3 服装目标区域识别网络流程图

2.2 服装关键点检测网络 (Detection KeyPoint)

对于图像特征图, 较低尺度图像往往含有丰富的语义信息, 对图像上半部分的关键点检测效果较好; 而更高尺度的图像含有丰富的空间纹理信息, 对图像下半部分关键点的检测效果较好[8]. 因此, CSPN 模型采用级联金字塔网络自上而下的对图像特征进行提取, 从语义丰富的深层图像特征图中, 通过上采样方式重新构建高分辨率的浅层, 达到对关键点进行定位的目的. 由于经过上采样之后的图像, 在关键点检测时存在偏差, 所以需要在上采样重建时将相应的高分辨率特征图叠加在一起, 经过关键点信息的重复利用及级联方式, 不断修正关键点位置, 从而提高关键点检测的准确率.

服装关键点检测网络的模型结构分为 Stage1、Stage2、Stage3 和 Stage4 共 4 个阶段, 如图 4 所示. Stage1 阶段是对 ResNet-101 中的特征图进行上采样后叠加操作, 输出 3 个特征图作为下一阶段的输入记做 L1. L1 中 3 个特征图像素值的大小分别为 64*64、128*128 和 256*256. L0 表示残差网络 ResNet-101 中的特征图, 本文选取 conv2~5 最后残差块的特征图; Up 为上采样操作 Up sampling; +为将上采样扩大后的图像与其同像素大小的图像叠加. Stage2 是对 L1 特征图重复进行上采样和叠加操作, 其输出特征图和 L1 中的 conv5 记做 L2, 作为 Stage3 的输入. Stage3 和 Stage4 复制 Stage1 和 Stage2 过程, 通过多级级联模式生成最终的 L4 特征图 (大小为 256*256), 即为服装关键点的预测结果. 在 4 个阶段中, 每个阶段均以前一个阶段的预测结果作为输入, 通过级联方式提高关键点的检测精度.

对于可见关键点, Stage1 阶段可直接预测得到. 服装图片经过网络模型输出 8 个关键点的 heatmap(数据集定义服装关键点为 8 个), 把每个 heatmap 计算 Loss 之后再回传, Loss 表达式为式 (1), 记为$L2$.

$$L2 = \frac{\sum_{i=1}^{n}\left[(kg_i - kp_i)^2\right]}{n} \tag{1}$$

其中, n表示关键点的个数, kg_i,kp_i分别表示标注的关键点坐标和预测的关键点坐标.

对于不可见关键点, 可通过增大感受野, 以级联层叠的方式来获得关键点. Stage2, Stage3 和 Stage4 阶段, 关键点检测难度依赖于关键点训练时的损失, 因此, 在训练过程中取前 K (K=4) 个损失值较大的关键点计算 Loss 并回传, Loss 表达式记为$L2^*$. 与 Stage1 阶段 Loss 不同的是, 关键点的个数$n = K$.

$$L2^* = \frac{\sum_{i=1}^{K}\left[(kg_i - kp_i)^2\right]}{K} \tag{2}$$

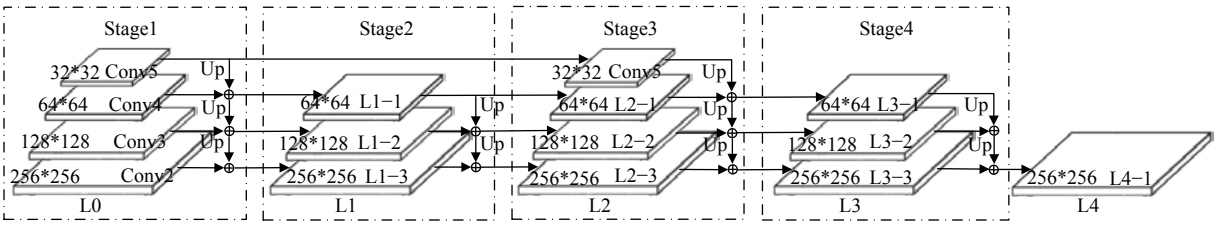

图 4 级联层叠金字塔服装关键点检测网络模型

3 实验

3.1 数据集

实验选用 DeepFashion 数据库进行训练和测试, 其 Landmark Detection Benchmark 是该数据库中一个公开的大型服装数据集, 包括 12.3 万张服装图片, 有上衣 (6 个关键点)、下装 (4 个关键点)、全身 (8 个关键点)3 种服装类型, 且存在正常、中等、严重等不同程度的变形图片. 图 5 为全身类服装 8 个关键点的标注. 本文从该数据集中选取 8.6 万张服装图片作为训练数据集, 3.7 万张作为测试数据集.

图 5 全身类服装关键点

3.2 评价指标

CSPN 模型采用 NE (Normalized Error) 值作为评价指标, 该值只考虑关键点可见情况下, 预测关键点与实际关键点之间的平均归一化距离. 数值越小代表预测的精确越高.

$$NE = \frac{\sum_k \left\{ \frac{d_k}{s_k} \delta(v_k = 1) \right\}}{\sum_k \{\delta(v_k = 1)\}} \times 100\% \tag{3}$$

其中, k为一张图片里关键点的编号, d_k为预测关键点和标注关键点间的距离, s_k为距离归一化参数 (上装及全身为左右袖口的欧式距离, 下装为左右腰侧的欧式距离), v_k表示该关键点是否可见. 在数据集标注中, 1 表示关键点可见, 0 表示关键点不可见, –1 表示关键点不存在. $v_k = 1$表示k关键点可见, $\delta(v_k = 1)$表示编号为k关键点参与计算. $\sum_k \delta(v_k = 1)$表示计算所有图片中编号为k的关键点的可见个数.

3.3 实验结果分析

由于关键点检测误差除与模型训练时的迭代次数有关, 还与级联层叠特征金字塔网络结构相关. 当级联层叠特征金字塔中层叠数和级联数大于 4 时, 继续增加特征金字塔结构中层叠数和级联数, 不仅不能有效提升检测效果, 而且会导致训练参数增加, 从而降低训练效率. 因而, 本实验将特征金字塔层叠数设置为 1～4, 并统计 4 个阶段中全身类服饰 8 个关键点的平均 *NE* 值, 如表 1 所示.

表 1 在 Landmark Detection Benchmark 数据集关键点检测的 *NE* 值

Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waistline	R.Waistline	L.Hem	R.Hem	Avg
FashionNet	0.0878	0.0910	0.0968	0.0925	0.0965	0.0845	0.0856	0.0832	0.0798
DFA	0.0615	0.0642	0.0635	0.0621	0.0746	0.0754	0.0642	0.0636	0.0661
DLAN	0.0576	0.0602	0.0668	0.0652	0.0703	0.0695	0.0635	0.0612	0.0642
CSPN	**0.0515**	**0.0523**	**0.0502**	**0.0556**	**0.0621**	**0.0648**	**0.0540**	**0.0575**	**0.0560**

表 1 列举了 CSPN 模型与 FashionNet[9]、DFA 和 DLAN 模型实验结果的对比. FashionNet 模型基于 VGG16 结构, 采用回归方法检测关键点; DFA 模型则基于 VGG16 结构, 通过 3 层级联形式使用回归方式预

测关键点; DLAN 模型是基于服装关键点原始数据的训练. 表 1 中, L.Collar (LC), R.Collar (RC), L.Sleeve(LS), R.Sleeve (RS), L.Waistline (LW), R.Waistline (RW), L.Hem(LH), R.Hem (RH) 分别表示左衣领、右衣领、左袖口、右袖口、左腰侧、右腰侧、左下摆、右下摆.

从表 1 可以看出, 由于 CSPN 模型引入了服装区域检测网络, 并采用级联层叠特征金字塔网络, 有效的提高了关键点检测的准确率.

表 2 为采用区域检测网络和未采用区域检测网络的服装关键点 *NE* 值对比. 其中 CSPN 为本文模型, NCSPN 为去掉服装区域检测网络 Detection Proposal 的模型.

表 2　是否采用检测网络关键点检测的 *NE* 值

Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waistline	R.Waistline	L.Hem	R.Hem	Avg
NCSPN	0.0706	0.0695	0.0718	0.0686	0.0642	0.0627	0.0715	0.0602	0.0673
CSPN	**0.0515**	**0.0523**	**0.0502**	**0.0556**	**0.0621**	**0.0648**	**0.0540**	**0.0575**	**0.0560**

从表 2 中可以看出, 相较于去掉服装区域检测网络, CSPN 模型具有更高的检测准确率. 在实验过程中, 缺少区域检测的步骤会导致左右袖口和左右衣摆易变形位置的关键点检测准确率较低. 而 CSPN 模型可以有效地解决这一问题.

另外, 图 6 列举了级联层叠金字塔结构中采用不同级联阶段时 8 个关键点的平均 NE 值及表 3 为不同 Stage 的算法性能对比.

实验中最终将 NE 降低到 5.56%, 一张图片在模型上的检测时间为 1 s. 随着层叠特征金字塔级联数量的增加, 关键点的预测值与真实值之间的距离逐渐缩小, 说明关键点的检测准确性得到了有效提高, 验证了 CSPN 模型的有效性.

图 7 为 CSPN 模型在人体不同姿态下, 对全身类服饰 8 个关键点的检测效果.

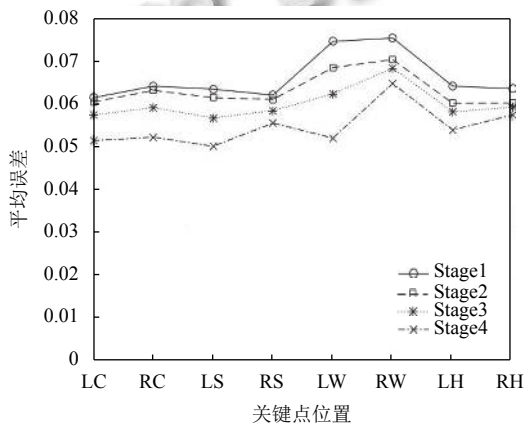

图 6　各关键点不同 Stage 平均误差

表 3　不同 Stage 的性能对比

阶段	Flops ($\times 10^9$)	参数量 (MB)	NE
Stage1	40	102	0.0661
Stage2	48	121	0.0632
Stage3	54	129	0.0600
Stage4	61	136	0.0560

图 7　不同人体姿态下全身类服饰关键点的检测效果

4　结束语

针对现有服装数据库中, 因服饰图片存在形变及背景复杂等因素导致服装分类识别率较低的问题, 本文结合服装目标区域识别和关键点回归方法, 提出了级联层叠金字塔网络 CSPN 模型. 该模型在 DeepFashion 数据集上与其他 3 种网络模型进行了对比, 结果显示: CSPN 模型能够有效提升服装关键点检测的准确率. 由于实验中采用的图像特征提取网络 ResNet-101 结构存在参数较多、效率较低等缺陷, 因而后续计划通过改变或替换该网络结构, 进一步提高关键点检测精度.

参考文献

1 Liu ZW, Yan SJ, Luo P, *et al*. Fashion landmark detection in the wild. Proceedings of the 14th European Conference on

Computer Vision. Amsterdam, The Netherlands. 2016. 229–245.

2 Yan SJ, Liu ZW, Luo P, *et al*. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, CA, USA. 2017. 172–180.

3 Wang WG, Xu YL, Shen JB, *et al*. Attentive fashion grammar network for fashion landmark detection and clothing category classification. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 4271–4280.

4 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

5 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

6 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 936–944.

7 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 779–788.

8 Chen YL, Wang ZC, Peng YX, *et al*. Cascaded pyramid network for multi-person pose estimation. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 7103–7112.

9 Liu ZW, Luo P, Shi Q, *et al*. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 1096–1140.