摘要

随着时尚产业和交互平台的发展，计算机视觉和模式识别领域中对人体目标的分析和理解技术的研究越来越受到关注。其中，人体解析正是人体目标分析和理解技术的核心方向之一，其目标是预测出图像中人体目标的各个像素所属的部件（包括肢体和衣服）类别。此研究能够帮助机器自动理解人类肢体语言，在消费电子、人机交互和智能安防等领域中提供有利的技术支撑，因此具有重要的研究价值与应用前景。

由于人体目标的姿态和外观多变、所处背景复杂，且人体解析是像素级预测任务，需要精细地获得每个像素的预测结果，这些因素均给人体解析技术带来了巨大挑战。因此，发掘图像中每个像素和其上下文之间的语义相关性成为人体解析中的一个重要研究方向，其中语义指的是图像的内容。本文从三个层次的图像语义相关性（部件之间的相关性、多任务之间的相关性和像素之间的相关性）对人体解析任务展开研究，从而获得精确的部件类别。本文的主要创新点如下：

第一，提出一种面向人体解析的部件相关性学习方法，用于学习人体的结构关系，从而联合当前部件与相关部件共同预测人体目标中每个像素所属的类别。传统基于卷积神经网络的特征学习方法使用多层卷积堆叠来学习不同人体区域的关系，模型复杂度高且较难收敛。为了解决这个问题，该方法引入边缘信息作为辅助，帮助人体解析模型感知并准确定位部件之间的边缘，从而获得更精确的部件特征。之后，基于图神经网络，通过将每个部件当做独立的图节点，学习部件与部件之间的相关性。在多个公开数据集上的实验结果表明，该方法取得了优越的性能，以ATR数据集上的实验结果为例，该方法比国际前沿方法TGPNet的解析性能提升了2.47%。

第二，提出一种面向复杂场景下人体解析的多任务相关性学习方法，用于从多个任务中获得辅助信息，从而在复杂场景下获得更合理的身体结构和准确的边缘。传统多任务联合的特征学习方法将不同的任务提取的特征简单地组合在一起，忽略了它们之间的相关性。为了解决上述问题，该方法同时学习三个密切相关的任务：人体解析、姿态估计和边缘检测，使用三个特征编码器分别提取部位、姿态和边缘特征，并使用多任务非局部关联模型学习部位特征与所有特征之间的相关性，从而指导网络从多个任务中获得重要的补充信息。之后，使用多个任务的损失函数训练模型，从而对特征编码器和多任务非局部关联模型进行优化。在多个公开数据集上的实验结果验证了探索多任务相关性方法的有效性，以LIP数据集上的实验结果为例，该方法比国际前沿方法CE2P的解析性能提升了3.08%。

第三，提出一种融合像素相关性的精细人体解析方法，用于学习像素之间的相关性并进一步和部件、多任务的相关性学习方法融合，从而从三个层次对人体目标进行感知，得到更精细的人体解析结果。传统基于像素相关性的特征学习方法并未直接指导像素相关性的学习过程，导致同类像素的差异仍然较大，不同类像素的差异仍然较小。为了解决上述问题，像素相关性学习方法首先学习所有像素之间的相似度，然后使用相似度保持损失函数对相似度施加约束，使得同类像素相似度尽可能大，不同类像素的相似度尽可能小。随后，基于施加约束后的类内和类间相似度，将同类像素以较大的权重进行聚合。此外，验证了部件相关性和多任务相关性也可以通过施加约束对相关性学习过程进行指导。在多个公开数据集上的实验结果表明，像素相关性学习方法取得了优越的性能，以LIP数据集上的实验结果为例，该方法比国际前沿方法SNT的解析性能提升了2.29%；当将像素相关性与部件、多任务的相关性进行融合后，性能的提升达到了3.37%。

综上所述，本文研究了基于多层次语义相关性的人体解析方法，通过学习人体部件、多任务以及像素之间的相关性，有效地提升了人体解析任务的性能。

关键词：人体解析，多层次语义相关性，部件相关性，多任务相关性，像素相关性

Multi-Level Semantic Correlation for Human Parsing

Ziwei Zhang (Computer Application Technology)
Directed by Prof. Xiaodong Xie

ABSTRACT

With the development of the fashion industry and interactive platform, the research of human analysis and understanding in computer vision and pattern recognition is gaining more and more attention. Human parsing is one of the core research directions of human target analysis and understanding techniques, which aims to partition a human image into semantic regions, including body parts and clothes. This research helps the machine to automatically understand human body language and give supports to the consumer electronics, human-computer interaction, and intelligent security. Therefore, it has significant research value and a wide range of application scenarios.

A number of factors pose great challenges for human parsing which is the changeable human posture and appearance, the complex background, and the necessity of obtaining the accurate result for each pixel to perform dense prediction. Therefore, exploring the semantic correlation between the pixel and its context has become an important research direction in human parsing, where semantic means the content in an image. This thesis studies how to exploit the semantic correlation from three perspectives to achieve precise human parsing. The main innovations of this thesis are as follows:

First, this thesis proposes a part correlation learning method for human parsing, exploring the structure relationship between body parts to jointly predict the category of each pixel by combining the features of the current part and its correlated parts.The traditional CNN-based feature learning method uses multiple convolutional layers to learn the relationship between different body regions, so that the model complexity is high and it is difficult to converge. In order to solve above problem, this thesis introduces edge information as an auxiliary to help the model to perceive and locate the edges so as to get accurate part feature. Then based on graph neural network, each part is treated as an independent node and their correlations are learned. Experimental results on several public datasets show that the proposed method

achieves superior performance. For example, the experimental results on the ATR dataset show a 2.47% performance improvement over TGPNet.

Second, this thesis proposes a multi-task correlation learning method for human parsing in complex scenarios, exploiting the correlations among multiple tasks and obtain reasonable body structure and accurate edge locations in complex scenes. The traditional feature learning methods based on multi-task joint learning simply combine the features learned from different tasks and ignore to explore the correlation between them. To solve the above problem, this thesis learns three tasks simultaneously: human parsing, pose estimation and edge detection, and designs three feature encoders to extract parsing, pose and edge features respectively. Then it uses a multi-task non-local correlation network to learn the correlation between parsing feature and all features, so as to obtain supplementary information from different tasks. Afterwards, multiple loss functions of different tasks are used to optimize the feature encoder and multi-task non-local correlation network. The experimental results on multiple public datasets verify the effectiveness of learning multi-task correlations. For example, the proposed method outperforms CE2P by 3.08% mIoU on the LIP dataset.

Third, this thesis proposes a fine human parsing method fusing pixel correlation with part correlation and multi-task correlation to perceive human from three levels and obtain finer results. Traditional feature learning methods based on pixel correlation do not explicitly constrain the correlation, resulting in large distances between pixels of the same category and small distances between pixels of different categories. To solve the above problem, pixel correlation learning method firstly learns the similarities of all the pixels. Then the similarity response is constrained by the similarity-preserving loss function to enlarge the similarity of within-class pixels and reduce the similarity of between-class pixels. Subsequently, pixels which are in the same category with high similarity are aggregated based on the learned relationships. Furthermore, correlation between parts and multiple tasks can also be guided by imposing a constraint to the correlation learning. The experimental results on multiple public datasets show that the proposed method has achieved significant performance improvement. For example, pixel correlation learning method outperforms SNT by 2.29% mIoU on LIP; when combining pixel correlation, part correlation and multi-task correlation, the performance is improved by 3.37%.

In summary, this thesis studies human parsing methods based on multi-level semantic correlation. From perspectives of learning the part correlation, multi-task correlation and pixel correlation, it effectively improves the performance of human parsing task.

KEYWORDS: Human Parsing, Multi-Level Semantic Correlation, Part Correlation, Multi-Task Correlation, Pixel Correlation