Image-based Learning of Human Body Information for Natural Human-Computer Interaction

Zhao, Ying

(Degree)
博士（工学）

(Date of Degree)
2020-03-25

(Date of Publication)
2021-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第7786号

(URL)
https://hdl.handle.net/20.500.14094/D1007786

（様式３）

論文内容の要旨

氏　名　ZHAO YING

専　攻　計算科学

論文題目（外国語の場合は，その和訳を併記すること。）

Image-based Learning of Human Body Information for Natural Human-Computer Interaction

自然なヒューマンコンピュータインタラクションにおけるイメージベースの人間身体情報学習

指導教員　羅　志偉

This thesis focuses on the image-based learning methods that contribute to the natural Human-Computer Interaction (HCI). Compare to physical equipment, human body is a much more natural medium for many HCI applications in both egocentric (first-person) and conventional (third-person) viewpoint, such as hand segmentation for virtual keyboard interaction of smart wearable glasses and human pose estimation for automatic stage lights control system. Therefore, this thesis targets to analyze images of human body and unconsciously recognize their diversity activities in the wild. This task is challenging due to unpredictable environmental condition in the natural application scene, such as rapid changes in illuminations, background clutter, multiple persons with variant articulation of body limbs, diverse appearance, self-occlusion, different scales, significant overlap between neighbor persons and crowd. To overcome the problems and meet the target, we develop image-based learning methods for hand segmentation and multi-person pose estimation.

In this thesis, we firstly review the basic knowledge enlightening our research. Image segmentation results are crucial cues for locating objects and boundaries in images. Superpixel generation and saliency detection are important processing related to image segmentation. We briefly review this basic knowledge that inspire our proposed approach for hand segmentation. With development of deep learning, the Convolutional Neural Network (CNN) becomes one of the mainstream tool to solve the problem of human pose estimation. Unlike handcrafted features used in classical approaches, CNN-based methods perform well with regard to both good feature extraction and strong discrimination ability. To design an effective CNN achieving enhanced performance, it is crucial to understand the principle of CNN components, the typical approaches representing the evolution of human pose estimation task and the evaluation metrics. Therefore, we also give a brief introduction of CNN-based methods with regard to human pose estimation.

We present an unsupervised on-the-fly hand segmentation approach which consists of top-down classification and bottom-up optimization. From the point of view of egocentric interaction loop, an unsupervised frame-level hand detector is proposed for the purpose of reducing the false positive caused by hand absence. We implement the frame-level detection by setting a non-interactive border based on an assumption that the hand is hardly to enter the view field from the top side for egocentric interaction. Based on the frame-level detection result, the superpixel-level and pixel-level classifiers are trained on-the-fly sequentially aimed at improving reliability of hand segmentation. To get stable samples for superpixel-level training, we select the candidates based on steps of confidence score calculation and energy optimization. In order to be robust to vary environmental conditions, the classifiers are updated from the bottom up based on the proposed performance evaluation method. This online-learning strategy makes our approach robust to varying illu-

mination conditions and hand appearances.

The recent popular method for pose estimation is extracting the local maximum response from each heatmap that trained for a specific type of keypoint. Learning each keypoint individually makes the approach difficult to parse from occlusion and variant pose. To target this problem, our cluster-wise feature aggregation method simultaneously learns complementary semantic information to encourage the detected keypoints subject to a certain contextual constraint. Specifically, we adopt stacked hourglass networks to generate paired multi-peak heatmaps for clusters of keypoints and the cluster configuration varies from stack to stack. The network encodes global and local consistency of entire body and parts by dense and sparse cluster branches. To enhance feature passing from shallow stack to deep stack, we aggregate information from different branches. The in-branch aggregation enriches the detection features in each branch by absorbing the holistic human region attention. The cross-branch aggregation further strengthens the detection features by fusing global and local context information between dense and sparse branches. Meanwhile, the intra-cluster and inter-cluster relationships are embedded with tag learning to guide the instance grouping and individual keypoint identification. Thus, the network optimizes the features at a specific keypoint by the received information from multi-level context.

From the structure point of view, the previous proposed method relies on sequential downsampling and upsampling procedures to capture multi-scale features and stacking basic modules to reassess local and global contexts. However, the network parameters become huge and difficult to be trained under limited computational resource. Motived by this observation, we design a lite version of conventional module that uses hybrid convolution blocks to reduce the number of parameters while maintaining performance. Moreover, due to the limitation of heatmap representation, the networks need extra and non-differentiable post-processing to convert heatmaps to keypoint coordinates. Therefore, we also introduce a novel bottom-up integral regression operation to reduce the quantization error of converting heatmaps to keypoint coordinates specifically for bottom-up pipeline multi-person pose estimation methods.

Multi-person pose estimation is mostly challenged by occlusion and variant postures. Existing bottom-up and top-down methods have their own advantages and limitations. The bottom-up approaches detect body keypoints without advance knowledge of person locations which are firstly explored in the top-down pipelines. However, the benefit of unifying the two pipeline is lack of exploration. Motived by this observation, we further propose a joint bottom-up and top-down learning method to better predict multi-person join locations. Our network not only learns the body keypoints but also the centers which indica-

te different person instances. Meanwhile, the offsets from body keypoints to the centers are encoded for both retrieving and grouping the joints. Based on shared features of keypoints-to-instances and instances-to-keypoints branches, our method can efficiently perform multi-person pose estimation. Moreover, we improve the bottom-up integral regression by adding a local restriction to complement the over-segmentation.

The thesis characterizes the benefits of image-based learning methods for natural HCI that unconsciously recognizes diversity human activities in the wild. This thesis demonstrates a hand segmentation approach that exploits the traits of egocentric viewpoint. Meanwhile, the thesis presents three approaches of multi-human pose estimation using unconstrained vision which offers extensive and diverse clues for not only the human activity but also the scene understanding. Experiments carried on public datasets validate the generality of the proposed approaches. In the future, we believe that these methods can be extended to or integrated with other computer vision tasks for improving human-computer interaction, such as object detection by exploring keypoints, sparse-to-dense tasks like depth recovering, virtual dressing by simultaneously detecting body joints and fashion landmarks, etc.

In a summary, we introduce the significance of our research topic as well as our contribution separately in 7 chapters. In Chapter 1, we introduce our research background, related previous research, research objective and overview of our research. In Chapter 2, we review the basic knowledge related to our research. In Chapter 3, we present an online learning approach that fully exploits the traits of egocentric vision. From Chapter4 to Chapter6, we discuss three approaches of multi-person pose estimation from perspectives network head and backbone structure. In Chapter 4, a deep convolution neural network is proposed for estimating multi-human poses in the wild. We use cluster-wise learning and feature aggregation to achieve this goal. In Chapter 5, we present a lite hourglass network for learning human body keypoints with less parameters and computational cost. In Chapter 6, we introduce a joint learning method which explores more geometric information from the top-down pipeline to complement the appearance features learnt from the bottom-up pipeline. Finally, in Chapter 7, we give a conclusion of our research topic and discuss the remaining issues.

（別紙１）　　論文審査の結果の要旨

氏名	Zhao Ying		
論文題目	**Image-based Learning of Human Body Information for Natural Human-Computer Interaction** 和訳：自然なヒューマンコンピュータインタラクションにおける イメージベースの人間身体情報学習		
審査委員	区　分	職　名	氏　　名
	主　査	教授	羅　志偉
	副　査	教授	上原　邦昭
	副　査	教授	陰山　聡
	副　査		印
	副　査		印

要　　旨

本研究は、人間と計算機や人工システムとのより自然な相互作用を目指して、異なる視点から撮影される画像情報を用いて、人工知能による学習処理を施し、人間の手の形状や身体各部位の関連情報の抽出を試みている。また、提案される学習方式と従来の提案手法との比較研究を行い、各種自然条件下での性能評価を行っている。人間の身体情報抽出は、数多くの応用課題に役に立つと考えられる。例えば、人間と計算機との相互作用における仮想キーボードの実現や各種舞台操作、画像合成に応用することが想定される。

本研究では、まずメガネにカメラが装着され、そのカメラ画像に撮影されているメガネ装着者自身の両手の領域を抽出することについて考えている。この場合、両手は基本的にカメラ画像の下部から画面に入り、画面上向きに手の姿勢を取ることが一般的であり、この特徴を利用して、画像から両手の領域を抽出するにあたり、自然環境における光や影などの環境要因による影響を削減し、よりロバストに両手の領域を抽出できるようにしている。この結果を活用することで、カメラメガネの装着者が目の前に提示される仮想キーボードを操作するための手の位置情報を取得することができ、様々な計算機入力操作を高い利便性で実現することができるようになる。この部分の研究成果は、査読付きの国際論文誌 EURASIP Journal on Image and Video Processing に掲載されている。

次に、世界座標のカメラから撮影された画像により、撮影された人物の頭部、肩、上半身、上肢と下肢などの各身体部分の抽出を行っている。具体的には、画像における一人物に対応して計 17 点で対応させ、大量なテスト画像とそれに対応する人物の身体部位に関する 17 点の位置座標を教師信号として、教師ありの深層学習で画像による人物の身体各部位の情報を抽出するための学習方式について研究している。従来の処理方式では、人物の各身体部位に対応する 17 点について、１点づつ取り上げて学習を行うこととなっており、莫大な計算コストを必要とされるだけでなく、画像における複数の人物の身体情報抽出は困難であった。そこで、本研究では、身体部位に対応する 17 点をグループ分けで深層学習を行うことを提案され、従来手法との性能比較を行い、その結果若干の性能改善が認められた。この部分の研究成果は、査読付きの国際論文誌 Pattern Recognition に掲載された。さらに、人物の身体情報抽出研究における画像の深層学習で、莫大な畳み込み演算の計算量問題を改善するために、本研究では異なる解像度の畳み込み演算を組み合わせることを提案している。この部分の研究成果は 26th Int. Conf. on Multimedia Modeling で発表された。

本論文は７章で構成されている。

第１章は、本研究の背景、従来の研究、本論文の目的と構成について説明している。

第２章には、画像分割や畳み込みニューラルネットワーク、人間の身体情報抽出に関する基本的な知識を整理している。

第３章は、自然環境下で撮影されているカメラ画像から撮影者自身の両手の領域情報を抽出する研究について説明している。

第４章から６章は、画像に撮影された人物の身体各部位情報抽出するための各種学習方式について説明している。

最後に第７章は，本論文のまとめと今後の課題について述べている。

氏名	Zhao Ying

以上で、本研究は、画像処理による人間の両手や身体部位情報抽出について、独自の学習方式を研究したものであり、より自然なヒューマンコンピュータインタラクションの実現について重要な知見を得たものとして価値ある集積であると認める。提出された論文はシステム情報学研究科学位論文評価基準を満たしており、学位申請者の Zhao Ying は、博士(工学)の学位を得る資格があると認める。