2-D Human Pose Estimation from Images based on Deep Learning：A Review

Yi Liu, Ying Xu, Shao-bin Li
School of Information Engineering
Communication University of China
Beijing, China

***Abstract*—Human pose estimation is an important research topic in the field of computer vision and artificial intelligence. This paper focuses on the state-of-art progress of 2-D human pose estimation methods based on deep learning. According to the neural network structure, these methods are classified as single CNN method, Multi-stage CNN method, Multi-branch CNN method, Recurrent Neural Network (RNN) method and Generative Adversarial Networks (GAN) method. We summarize and analyze their attributes and performance. The future development direction is also prospected.**

Keywords—deep learning; human pose estimation; 2-D image

I. INTRODUCTION

Human pose estimation is a significant research topic in computer vision and artificial intelligence. 2-D human pose estimation in still images is the basis of 3-D pose estimation in motion images. In recent years, although researchers are digging deeper in this area, they still face a series of challenges, including foreground occlusion, background chaos, illumination, complexity of poses, multi-person overlaps, calculation complexity etc.

The traditional 2-D human pose estimation methods [1,2,3,4,5,6] are mainly based on the Pictorial Structure(PS) model, exploiting the Kinematic Tree to characterize the distribution of human joints. Firstly, Histogram of Oriented Gradient (HOG) [7], Hue, Saturation and Value (HSV) color model [8] and other algorithms are used to extract the information about shape, color etc. Secondly, estimate the human pose combining with regression, generation, part relations and other reasoning models. The traditional methods have quite a few shortcomings, i.e. the tree model is not able to represent all the constraints, the extraction performance of handcraft feature is poor and the reasoning complexity is high.

In recent years, deep learning has been widely applied in image classification [9], object detection [10, 11, 13], speech detection, semantic analysis, human action recognition and pose estimation. Compared with traditional methods, methods based on deep learning train the networks with large amount of images, learn features from global space. It needn't detect or recognize the local feature of objects and has good robustness. This paper focuses on the state-of-the-art progress on deep learning based 2-D human pose estimation methods. According to the network structures, these methods are classified as single CNN method, Multi-stage CNN method, Multi-branch CNN method, RNN method and GAN method.

II. 2-D HUMAN POSE ESTIMATION

A. Single CNN methods

Tompson et al. [16] utilize a single CNN to extract multi-scale features of human body parts, which incorporates a high-order spatial model based on the Markov random field and represents the structural domain constraints among joints.

Chen et al. [18] propose a graphical model with image dependent pairwise relations (IDPRs), based on Pictorial Structure (PS), Mixtures-of-part and Conditional Random Fields (CRF). IDPRs models pairwise relations between body parts, detects body parts and predicts their spatial relationship.

Yang et al. [19] combine deformable mixture parts model with Deep Convolutional Neural Network (DCNN), considering the global consistency of poses to reduce the negative influence of part deformation. This method is generalized to tree model and loopy graph.

Chu et al. [20] model the spatial dependence between joints with bi-directional tree model and geometric transformation kernel, which enhances feature extraction in combination with CNN. On this basis, Xiao Chu et al. [21] utilize CRF to model the structural features of the human body part configuration, which further improved performance.

Carreira et al. [37] introduce a top-down feedback algorithm named Iterative Error Feedback (IEF). IEF extends CNN's learning hierarchical representation from input space to the output with a self-correcting model.

Pishchulin et al. [27] decrease the receptive field size of Fast R-CNN [10] to focus on local information, which transforms the part detection into a multi-label classification task, and conduct a bottom-up deduce in combination with DeepCut. The calculation is a little complex. On this basis, Insafutdinov et al [29] introduce the residual learning [14], which involves more context by increasing the depth of the network and the size of the receptive field.

Compared with traditional methods designing features in a handcraft way, CNN is better at feature learning. The single CNN methods extract features by exploiting CNN and model the configuration of joints using PS model.

B. Multi-stage CNN methods

Toshev et al. [15] use DCNN to reason human body joints in a holistic view, optimize for the occlusion caused by

978-1-5386-1803-5/18/$31.00 ©2018 IEEE

invisible joints and improve the accuracy by multiple cascading DCNN.

Shih En Wei et al. [25] design a cascaded CNN network to represent texture and spatial information with convolution layers. The size of the receptive field is increased to reduce the information loss of neighboring joints and study part relations and spatial context. Intermediate supervision is adopted to solve "gradient descent" caused by the increasing depth of network. This method is robust to severe part occlusion.

Based on residual learning and Fully Convolutional VGG [12], Newell et al. [28] design the hourglass residual module, which extracts multi-scale features through continuous down-sampling operations, and restores image resolution with up-sampling operations. It imitates the top-down and bottom-up reasoning process in pose estimation.

Chu and Yang et al. [30] design HRU (Hourglass Residual Units) based on the hourglass module to expand the receptive field and learn multi-scale features, multi-semantic features and features with different resolutions. What's more, they model interrelationship among adjacent areas with CRF and combine the global attention model with the local attention model, considering different granularities from local salient regions to global semantic space. Eventually, they improved the performance through multiple cascading HRU modules.

Yang et al. [31] propose PRMs (Pyramid Residual Module) which improves the scale variations of feature compared to that in Stacked Hourglass Network [28].

Bulat et al. [24] synthesizes the regression and detection results by cascading a part detection sub-network and a regression sub-network.

After training, each layer of a Multi-stage CNN network becomes relatively independent and is given a clear meaning in the structure and function. These methods solve the convergence difficulty of a large-scale network, thus gaining better performances compared to single CNN methods.

C. *Multi-branch CNN methods*

Fan et al. [22] develop a two-branch CNN architecture to extract global and local features, with both global and local information considered.

Li et al. [23] depend on CNN's generalization for detection and regression tasks. By setting the features extracted from a shared CNN as the common input of a detection sub-network and a regression sub-network, the model estimated human poses through synthesizing detection and regression results.

Cao et al. [26] extract features with the first 10 layers of VGG-19[12], then further process the features using a three-branch CNN architecture, whose one branch predicts joint locations, one models limb direction and orientation with PAF (Part Affinity Field), and the last one keeps initial features through a shortcut. This method concatenates and iteratively optimizes the output of three branches, gradually improving the accuracy of regression in a coarse-to-fine way. The multi-branch CNN methods combine the processing results of CNN branches to obtain rich representations of image information, so as to improve the performance.

D. *RNN methods*

Gkioxari et al. [33] learn from Seq2Seq (Sequence to Sequence) which models the correlations of body joints, adopt a recurrent network and chained prediction to estimate joint positions. It feeds the output of network back to the input. This method eliminates the over-simplification of CRF caused by designing correlations of joints in a handcraft way, thus enhancing the expression of the model.

Belagiannis and Zisserman et al. [34] incorporate an iterative recurrent module in CNN, leading to the increasing size of the effective receptive field and the advantage over occlusion prediction.

Lifshitz and Fetaya et al. [38] incorporate sequentially predictions and local geometric constraints in DCNN, utilize the information of the entire image to predict the distribution probability of joints, and calculate the image-dependent joint probability of joints according to the consensus 'voting'.

The RNN methods utilize the RNN network's property of timing memory. Although 2-D human pose estimation doesn't involve timing, RNN's preservation for previous information can be used to model the interrelationship between joints.

E. *GAN methods*

Chen et al. [35] adopt Generative Adversarial Networks (GAN) to help modify pose prediction. In order to well capture the structural correlations among body joints, the generator G is designed in a stacked multi-task way to predict poses and occlusion heat-maps. It effectively addresses occlusion.

Chou et al. [36] combine the residual learning with the GAN by constructing the generator and discriminator with a 4-stack hourglass network separately, which outperformed the 8-stack hourglass network without GAN [28].

GAN is based on unsupervised learning and completes the training through massive "confrontation" between the generator and the discriminator. It captures high-order correlations of data in the absence of target class label information. Compared with regression models, GAN fits data faster and more adequately, and generates better samples.

Some other methods combining several theories also work. Jain et al. [17] extract low-level features with CNN, and exploit multiple small CNNs to classify each single part independently. A pure bottom-up weak spatial model based on spatial prior is proposed to extract high-level features. Bulat et al. [32] design a residual binarization module which employs parallel branches in the residual module to extract hierarchical and multi-scale features. It also reduces parameters, computing time and calculating complexity at the cost of some accuracy.

III. Evaluation and Performance

A. *Datasets*

The most popular datasets in 2-D human pose estimation are MPII Human Pose dataset (MPII) and Leeds Sports Pose dataset (LSP). MPII contains 25,000 images of full body, each of which is annotated with 14 joints, covering daily activities, as shown in Figure 1. LSP contains 1000 training samples and

1000 test samples of full body, each of which is annotated with 14 joints, covering sporting events, as shown in Figure 2.

Fig. 1. Examples of MPII

Fig. 2. Examples of LSP

B. Evaluation Criteria

In 2-D human pose estimation, the joint is considered to be detected if the distance between the predicted joint and the ground-truth is within the standardized threshold. Accuracy precision differs when adopting different thresholds. Widely used evaluation criteria includes Percentage of Correct Parts (PCP), Percent of Correct Keypoints (PCK) and Mean Average Precision (mAP).

PCP selects limb length as the benchmark, evaluates the detection accuracy of head, torso, upper arm, lower arm, upper leg and lower leg. The ratio of strict-PCP is 0.5. PCK selects the normalized distance as the benchmark, evaluates the detection accuracy of seven joints, containing head, shoulder, elbow, wrist, hip, knee, and ankle. PCKh@0.5 means that the detection of a joint is considered successful within a threshold of 0.5 to the size of the head. mAP reflects the average PCKh detection rate of all the joints.

C. Performance

The paper compares the performance of methods mentioned above. Methods listed in Table I shows the average accuracy of all joints evaluated by strict-PCP, PCK and mAP. Other methods that only estimate part of joints are not listed.

Based on Table 1, single CNN methods have achieved good performances under strict-PCP. Multi-stage CNN methods gain excellent performances under PCK. Multi-branch CNN methods are more effective under mAP. It is worth mentioning that RNN and GAN methods have achieved the highest accuracy under strict-PCP and PCK respectively.

TABLE I. COMPARISON OF 2-D HUMAN POSE ESTIMATION METHODS

Author	*Method*	*Performance（%）*	
		MPII	*LSP*
strict-PCP			
Carreira et al[37]	IEF		72.5%
Chen et al[18]	Single CNN		75.0%
Chu et al[20]	Single CNN		81.1%
Yang et al[19]	Single CNN		81.8%
Chu et al[21]	Single CNN		83.1%
Lifshitz et al[38]	RNN		**84.2%**
PCK			
Carreira et al[37]	IEF	81.3% (PCKh)	
DeepCut [27]	Single CNNCNN	82.4% (PCKh)	87.1% (PCK)
Bulat et al[24]	Multi-stage CNN (VGG)	82.7% (PCKh)	83.5% (PCK)
Lifshitz et al[38]	RNN	85.0% (PCKh)	
Bulat et al[32]		85.5% (PCKh)	
Gkioxari et al[33]	RNN	86.1% (PCKh)	
WeiShihEn et al[25]	Multi-stage CNN	87.95% (PCKh)	
Belagiannis et al[34]	RNN	88.1% (PCKh)	85.2% (PCK@0.2)
DeeperCut [29]	Single CNN	88.5% (PCKh)	90.1% (PCK)
Newell et al[28]	Multi-stage CNN	89.4% (PCKh)	
Bulat et al[24]	Multi-stage CNN (ResNet)	89.7% (PCKh)	90.7% (PCK)
Chou et al[36]	GAN	91.8% (PCKh)	**94.0% (PCK@0.2)**
Chu et al[30]	Multi-stage CNN	91.5% (PCKh)	92.6% (PCK@0.2)
Yang et al[31]	Multi-stage CNN	92.0% (PCKh)	93.9% (PCK@0.2)
Chen et al[35]	GAN	**92.1% (PCKh)**	93.1% (PCK@0.2)
mAP			
Fan et al[22]	Multi-branch CNN		44.4%
Cao et al[26]	Multi-branch CNN	**79.1%**	

IV. CONCLUSION

This paper concentrates on the state-of-the-art progress of 2-D human pose estimation based on deep learning. Classification, introduction, and performance comparison of the methods are involved.

Different methods have their own merits and shorts. Single CNN methods combine CNN which extracted features, with traditional PS model which models the distribution of joints, leading to low complexity of network. Multi-stage CNN methods boost feature extraction using cascading CNNs, implicitly models the relationships of neighboring joints using the receptive field. However, problems brought by the

increasing depth of networks, such as gradient decent and parameters mounting, need to be solved. Multi-branch CNN methods synthesize results of multiple CNN branches, at the cost of complex calculation and training. RNN methods applies the sequential processing used in Natural Language Processing to model the interrelationships between body joints, expands the receptive field with iteration of the recurrent network, and solves the occlusion problem in a certain degree at a cost of massive calculation. GAN methods captures the high order correlation of the data in the absence of the target class label information. Such methods have provide the state-of-the-art best solution against the occlusion. Nonetheless, the drawback is still the complex calculation and the time cost.

Now, occlusion is still a pressing matter in 2-D human pose estimation. Some methods have solved it to a certain extent, nonetheless, lack the robustness. In addition, the high cost of memory and time caused by the algorithm complexity limits their application value. These problems are still leading to a great difficulty of 2-D human pose estimation research in the future.

REFERENCES

[1] Yang, Yi, and D. Ramanan. "Articulated Human Detection with Flexible Mixtures of Parts." IEEE Transactions on Pattern Analysis & Machine Intelligence 35.12(2013):2878.

[2] Cherian, Anoop, et al. "Mixing Body-Part Sequences for Human Pose Estimation." Computer Vision and Pattern Recognition IEEE, 2014:2361-2368.

[3] Puwein, Jens, et al. Foreground Consistent Human Pose Estimation Using Branch and Bound. Computer Vision – ECCV 2014. Springer International Publishing, 2014:315-330.

[4] Kiefel, Martin, and P. V. Gehler. "Human Pose Estimation with Fields of Parts." European Conference on Computer Vision Springer, Cham, 2014:331-346.

[5] Fu, Lianrui, J. Zhang, and K. Huang. "Beyond Tree Structure Models: A New Occlusion Aware Graphical Model for Human Pose Estimation." IEEE International Conference on Computer Vision IEEE, 2015:1976-1984.

[6] Dantone, Matthias, et al. "Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images." Pattern Analysis & Machine Intelligence IEEE Transactions on 36.11(2014):2131-2143.

[7] Sapp, Benjamin, A. Toshev, and B. Taskar. "Cascaded Models for Articulated Pose Estimation." Computer Vision - ECCV 2010, European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings DBLP, 2010:406-420.

[8] Ding, Xiao. "Human Pose Esyimation in Static Images." Diss. Liaoning University, 2015. (in Chinese)

[9] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." International Conference on Neural Information Processing Systems Curran Associates Inc. 2012:1097-1105.

[10] Girshick, Ross. "Fast R-CNN." IEEE International Conference on Computer Vision IEEE Computer Society, 2015:1440-1448.

[11] Ren, Shaoqing, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE Transactions on Pattern Analysis & Machine Intelligence 39.6(2017):1137-1149.

[12] Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." Computer Science (2014).

[13] Girshick, Ross, et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." (2013):580-587.

[14] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." Computer Vision and Pattern Recognition IEEE, 2016:770-778.

[15] Toshev, Alexander, and C. Szegedy. "DeepPose: Human Pose Estimation via Deep Neural Networks." (2014):1653-1660.

[16] Tompson, Jonathan, et al. "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation." Eprint Arxiv(2014):1799-1807.

[17] Jain, Arjun, et al. "Learning Human Pose Estimation Features with Convolutional Networks." Computer Science (2013).

[18] Chen, Xianjie, and A. Yuille. "Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations." Eprint Arxiv(2014):1736-1744.

[19] Yang, Wei, et al. "End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation." Computer Vision and Pattern Recognition IEEE, 2016:3073-3082.

[20] Chu, Xiao, et al. "Structured Feature Learning for Pose Estimation." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2016:4715-4723.

[21] Chu, Xiao, et al. "CRF-CNN: Modeling Structured Information in Human Pose Estimation." (2016).

[22] Fan, Xiaochuan, et al. "Combining local appearance and holistic view: Dual-Source Deep Neural Networks for human pose estimation." Computer Vision and Pattern Recognition IEEE, 2015:1347-1355.

[23] Sijin, L. I., Z. Q. Liu, and A. B. Chan. "Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network." IEEE Conference on Computer Vision and Pattern Recognition Workshops IEEE Computer Society, 2014:488-495.

[24] Bulat, Adrian, and G. Tzimiropoulos. Human Pose Estimation via Convolutional Part Heatmap Regression. Computer Vision – ECCV 2016. Springer International Publishing, 2016:717-732.

[25] Wei, Shih En, et al. "Convolutional Pose Machines." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2016:4724-4732.

[26] Cao, Zhe, et al. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." (2016).

[27] Pishchulin, Leonid, et al. "DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation." 2008.1(2015):4929-4937.

[28] Newell, Alejandro, K. Yang, and J. Deng. "Stacked Hourglass Networks for Human Pose Estimation." (2016):483-499.

[29] Insafutdinov, Eldar, et al. "DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model." European Conference on Computer Vision Springer, Cham, 2016:34-50.

[30] Chu, Xiao, et al. "Multi-Context Attention for Human Pose Estimation." (2017).

[31] Yang, Wei, et al. "Learning Feature Pyramids for Human Pose Estimation." (2017).

[32] Bulat, Adrian, and G. Tzimiropoulos. "Binarized Convolutional Landmark Localizers for Human Pose Estimation and Face Alignment with Limited Resources." (2017).

[33] Gkioxari, Georgia, A. Toshev, and N. Jaitly. "Chained Predictions Using Convolutional Neural Networks." European Conference on Computer Vision Springer, Cham, 2016:728-743.

[34] Belagiannis, Vasileios, and A. Zisserman. "Recurrent Human Pose Estimation." IEEE International Conference on Automatic Face & Gesture Recognition IEEE, 2017:468-475.

[35] Chen, Yu, et al. "Adversarial PoseNet: A Structure-aware Convolutional Network for Human Pose Estimation." (2017).

[36] Chou, Chia Jung, J. T. Chien, and H. T. Chen. "Self Adversarial Training for Human Pose Estimation." (2017).

[37] Carreira, Joao, et al. "Human Pose Estimation with Iterative Error Feedback." 2013.2013(2016):4733-4742.

[38] Lifshitz, Ita, E. Fetaya, and S. Ullman. Human Pose Estimation Using Deep Consensus Voting. Computer Vision – ECCV 2016. Springer International Publishing, 2016.