Matching-CNN Meets KNN: Quasi-Parametric Human Parsing

Si Liu[1,2], Xiaodan Liang[2,4], Luoqi Liu[2], Xiaohui Shen[3], Jianchao Yang[3], Changsheng Xu[5], Liang Lin[4], Xiaochun Cao[1], Shuicheng Yan[2],
[1] SKLOIS, IIE, Chinese Academy of Sciences [2]National University of Singapore [3]Adobe Research [4]Sun Yat-sen University [5] IA, Chinese Academy of Sciences

Both parametric [6] and non-parametric approaches [6, 7] have demonstrated encouraging performances in the human parsing task, namely segmenting a human image into several semantic regions (e.g., hat, bag, left arm, face). In this work, we aim to develop a new solution with the advantages of both methodologies, namely supervision from annotated data and the flexibility to use newly annotated (possibly uncommon) images, and present a quasi-parametric human parsing model.

Under the classic K Nearest Neighbor (KNN)-based nonparametric framework, the parametric Matching Convolutional Neural Network (M-CNN) is proposed to predict the matching confidence and displacements of the best matched region in the testing image for a particular semantic region in one KNN image. As shown in Fig. 1, given a testing image, we first retrieve its KNN images from the annotated/manually-parsed human image corpus. Then each semantic region in each KNN image is matched with confidence to the testing image using M-CNN. Reliable matching between an input image and a KNN region is challenging, because the matching needs to handle the large spatial variance of semantic regions. For example, the bags can be placed on the left, right or in front of the human body. The proposed M-CNN is able to achieve accurate multi-ranged matching. As shown in Fig. 1, M-CNN contains three paths, i.e., two single image convolutional paths and a cross image convolutional path. The *single image convolutional path* receives the input image or a particular KNN region, and produces its discriminative hierarchical feature representations layer by layer. The *cross image convolutional path* embeds cross image filters into every convolutional layer to characterize the multi-ranged matching. The cross image filters are applied to all feature maps in previous convolutional layers, including the single image feature maps and cross image feature maps. Because the scale of receptive fields of the feature maps increase when tracing up the M-CNN, the cross image matching filters capture the displacements from the near-range to the far-range. Therefore the feature maps from the cross image convolutional path can well represent the displacements. Because the feature maps generated by the two single image convolutional paths are excellent feature representations, their absolute difference maps are calculated as another measurement of the displacements. The difference maps are combined with the cross image feature maps and then link to the subsequent fully connected layers. Finally, the matching confidence and displacements are regressed. Then, the matched regions from all KNN images are further fused, followed by a superpixel smoothing procedure to obtain the ultimate human parsing result. Comprehensive evaluations over a large dataset with 7,700 annotated human images well demonstrate the significant performance gain from the quasi-parametric model over the state-of-the-arts [6, 7], for the human parsing task.

Table 1: Comparison of parsing performances with several architectural variants of our model (cross image matching filters embedded into different convolutional layers, with and without superpixel smoothing) and two state-of-the-arts.

Method	Accuracy	F.g. accuracy	Avg. precision	Avg. recall	Avg. F_1
Yamaguchi et al. [6]	84.38	55.59	37.54	51.05	41.80
PaperDoll [7]	88.96	62.18	52.75	49.43	44.76
Siamese [1]	85.24	56.42	50.27	48.88	47.08
M-CNN (w/o cross)	88.30	69.84	58.63	59.52	56.99
M-CNN (cross 5)	88.62	69.88	60.89	60.47	58.07
M-CNN (cross 5,4)	89.41	72.44	58.93	63.16	60.03
M-CNN (cross 5,4,3)	88.97	70.84	60.27	62.23	60.36
M-CNN	**89.57**	**73.98**	**64.56**	65.17	**62.81**
M-CNN (cross 5,4,3,2,1)	89.42	71.86	63.13	63.49	61.53
M-CNN(w/0 ss)	87.08	71.73	55.88	**65.32**	59.39

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

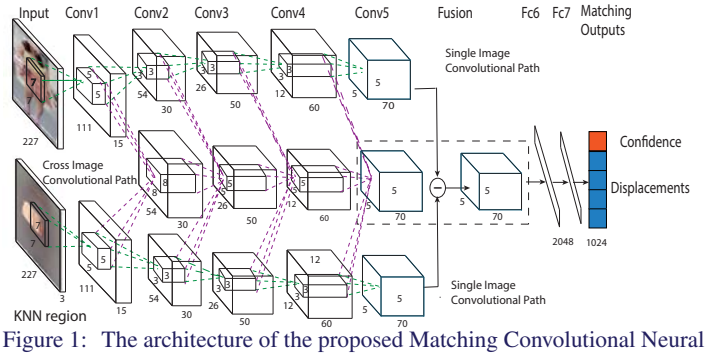

Figure 1: The architecture of the proposed Matching Convolutional Neural Network (M-CNN) with parameters shown.

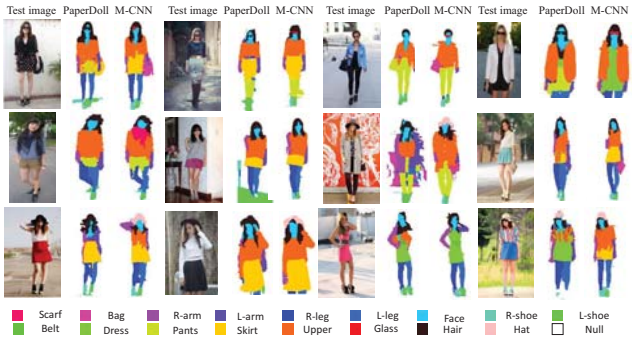

Figure 2: Comparison of our parsing results with the PaperDoll Method. For each image, we show the testing image, parsing results by PaperDoll [7], our "M-CNN" sequentially.

We use the dataset in [5] pixel-wisely labeled by the 18 categories defined by Daily Photos dataset [2]. We compare our M-CNN based quasi-parametric human parsing framework with two state-of-the-arts: Yamaguchi et al. [6] and PaperDoll [7]. Our "M-CNN" significantly outperforms these two baselines by over 21.01% for Yamaguchi et al. [6] and 18.05% for PaperDoll [7]. This verifies the effectiveness of our end-to-end M-CNN based quasi-parametric framework. We also extensively explore different CNN architectures to demonstrate the effectiveness of each component in M-CNN more transparently. The architecture of M-CNN is shown in Fig. 1 and other variants are constructed by gradually adding/eliminating the cross image filters in different layers. M-CNN contains 4 cross image matching filters from layers *conv*2 to *conv*5. "M-CNN (cross 5,4,3,2,1)" is obtained by adding $11 \times 11 \times 6$ cross image filters in the "conv1" layer to the "M-CNN". Fig. 2 shows the comparison between M-CNN and PaperDoll [7]. The results demonstrate that our method can successfully predict the label maps with small regions, which can be attributed to the reliable label transferring from the KNN regions.

[1] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[2] Jian Dong, Qiang Chen, Wei Xia, Zhongyang Huang, and Shuicheng Yan. A deformable mixture parsing model with parselets. In *ICCV*, 2013.

[3] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[5] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. 2015.

[6] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

[7] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013.