Multi-Objective Convolutional Learning for Face Labeling

Sifei Liu, UC Merced
Jimei Yang, UC Merced
Chang Huang, IDL, Baidu Inc.
Ming-Hsuan Yang, UC Merced

1. Implementation Details

The nonparametric shape prior is constructed based on five keypoints (as shown in Figure 3 in the manuscript) which can be computed more efficiently using [1]. Given a test image, 20 ($K = 20$) exemplars are selected by comparing the Euclidean distance of PCA coefficients in the keypoints subspace. The selected exemplars are then aligned to the test image through the similarity transformation. The least-squares optimization is used on the corresponding five keypoint pairs (exemplar vs. test image) to estimate their similarity transformation. The ground truth labels are then transformed using the the similarity transformation, and combined with weights to generate the prior. For any specific class, the prior is:

$$M = \sum_{k=1}^{K} \alpha_k M_K, \quad \alpha_k = \frac{\|UP - UP_k\|_2^2}{\sum_{l=1}^{K} \|UP - UP_l\|_2^2}, \tag{1}$$

where U is the eigenvector of keypoints on the validation set. In addition, P and P_k are respectively the detected five keypoints for test image and the k-th exemplar, and M_k is the ground truth binary label for the k-th exemplar. In (1), UP_k forms the projection of P_k in the subspace. The weights α are proportional to the Euclidean distance of the test image to the exemplars, where α_k are the weight for the k-th exemplar among the nearest K exemplars. The value of nonparametric prior M ranges from 0 to 1. We denote the input or activation of each layer in the proposed CNN as $h \times w \times d$, where h and w are spatial size, and d is the number of channels for the input, or dimensions for activation. The detailed pipeline with nonparametric prior input is illustrated as shown in Figure 1.

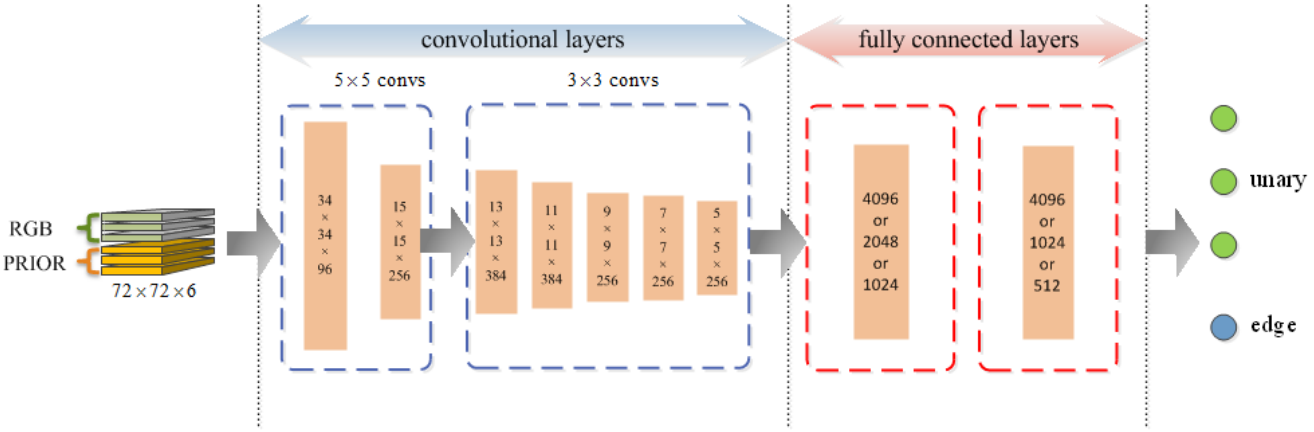


Figure 1. Proposed CNN classifier for the LFW-PL dataset (patch-based training phase) with input and activation size of each layer.

2. Improvement by Nonparametric Prior

We list the network performances with respect to FCs settings of $4096*4096$ (163MB), $4096*1024$ (119MB), $2048*4096$ (65MB), $1024*1024$ (38MB) and $1024*512$ (36MB), as shown in Table 1-4. The quantitative evaluations for MO-unary, MO-GC, MO-unary with prior and MO-GC with prior corresponde to the results in Figure 6 of the manuscript. Note that we only test and compare at the setting of $4096*1024$ (119MB) for **S-CNNs** and **S-CNNs with prior**.

Figure 2 shows the training convergence rate for the first 20 epochs between two multi-objective learning (MO) approaches, one of them (magenta) applies the nonparametric prior, the other (blue) does not. We illustrate one example with the same FC layer settings of $4096*1024$, where the total iteration numbers for 20 epochs is $60,000$. We demonstrate that other than improving the labeling performance, the proposed nonparametric prior can significantly speed up the convergence in the training process.

Table 1. Overall per-pixel accuracy on the *LFW-PL* dataset.

accuracy (%)	4096* 4096	4096* 1024	2048* 1024	1024* 1024	1024* 512
S-CNNs	-	92.92	-	-	-
MO-unary	93.05	93.45	92.74	92.91	92.70
MO-GC	93.41	93.77	92.89	93.23	93.05
S-CNNs with prior	-	94.25	-	-	-
MO-unary with prior	94.82	94.94	95.03	95.03	94.99
MO-GC with prior	94.95	95.12	95.19	**95.24**	95.16

Table 2. Overall F-measure of skin on the *LFW-PL* dataset.

F-skin (%)	4096* 4096	4096* 1024	2048* 1024	1024* 1024	1024* 512
S-CNNs	-	90.07	-	-	-
MO-unary	90.84	91.45	89.58	90.39	89.99
MO-GC	91.37	91.95	90.21	90.88	90.49
S-CNNs with prior	-	92.79	-	-	-
MO-unary with prior	93.61	93.64	93.73	93.73	93.75
MO-GC with prior	93.89	93.93	94.00	94.03	**94.05**

Table 3. Overall F-measure of hair on the *LFW-PL* dataset.

F-hair (%)	4096* 4096	4096* 1024	2048* 1024	1024* 1024	1024* 512
S-CNNs	-	73.73	-	-	-
MO-unary	76.56	78.03	74.84	76.76	76.17
MO-GC	77.45	79.06	75.90	77.64	77.42
S-CNNs with prior	-	78.03	-	-	-
MO-unary with prior	79.70	79.95	80.67	80.27	80.24
MO-GC with prior	80.47	80.70	81.09	**81.27**	80.70

Table 4. Overall F-measure of background on the *LFW-PL* dataset.

F-bg (%)	4096* 4096	4096* 1024	2048* 1024	1024* 1024	1024* 512
S-CNNs	-	95.18	-	-	-
MO-unary	95.56	95.84	95.51	95.48	95.37
MO-GC	96.79	96.03	95.73	95.68	95.59
S-CNNs with prior	-	96.63	-	-	-
MO-unary with prior	96.84	97.02	97.03	97.03	97.02
MO-GC with prior	96.90	**97.10**	**97.10**	**97.10**	**97.10**

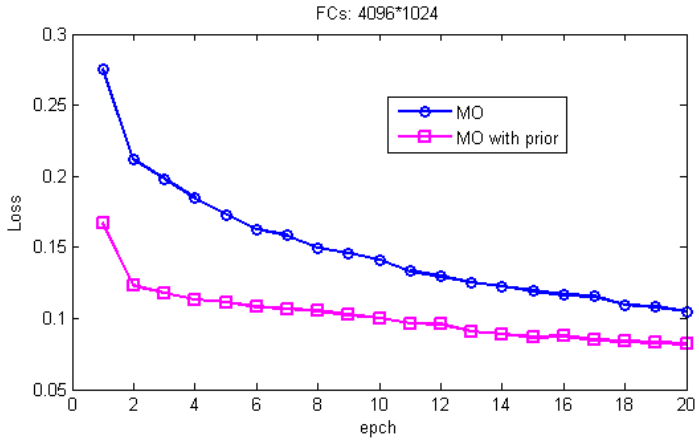


Figure 2. Training loss of multi-objective learning for FCs set to $4096 * 1024$. The nonparametric prior, colored in magenta, significantly speeds up the convergence in the training phase, compare to the one without prior as colored in blue.

3. Runtime

All models are trained and tested using the Caffe package on a single NVIDIA Tesla K10c GPU. Our proposed method (MO-GC with prior) with FCs setting of $1024 * 1024$ takes approximately 120 ms to forward propagate an 250×250 6-channel input (See Figure 1). To generate original-sized output maps with 16 shifted versions of an input and to infer the final labeling results, it takes less than two seconds for the full pipeline. The configurations and code will be released to the public.

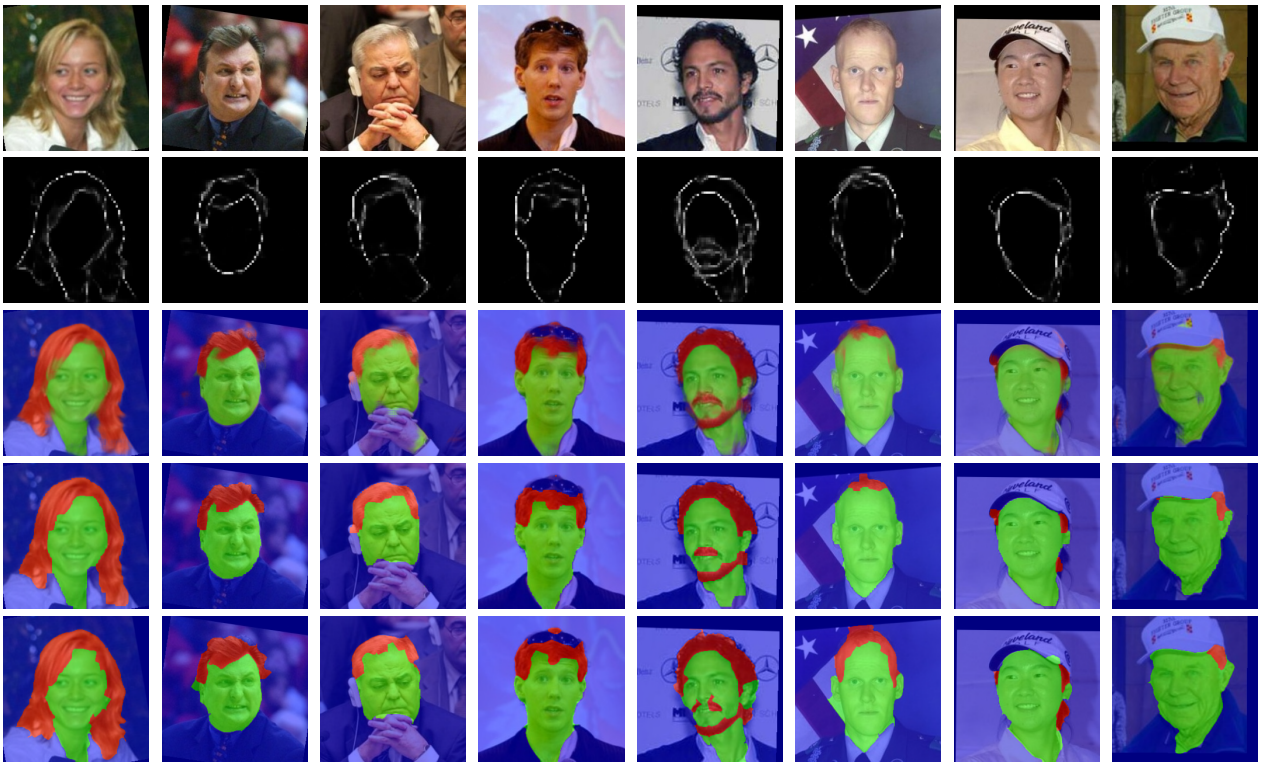
4. More Results on the LFW-PL Dataset

Figure 3 and 4 show more experimental results using challenging images including blurry hair region with low contrast, occlusions and mustache. For unary output on the third row, we show a "soft mask" with values ranging from 0 to 1 for each class. Specifically, the hair region (red) reveals its natural properties of transparency by showing a smooth probability map. For edge output on the second row, we also illustrate a probabilistic output that ranging from 0 to 1, which is directly used on the inference step. Our generated edge is clean (with much little on the background) and accurate, which further helps infer labeling results with better class boundaries as shown on the forth row. Although our approach is not specifically designed to handle occlusions, it handles such factors well as shown in Figure 3(a)(c) and Figure 4(d).

The superpixel-based manual annotation results by the LFW-PL dataset, although practical and efficient, does not contain pixel-level ground truth data. Some typical examples are shown in Figure 3(d)(g) and Figure 4(b)(g) where the boundary regions are not well defined by superpixels, and inaccurate annotations are thus generated. Furthermore, humans may not be able to annotate details well, e.g., the mustache region in Figure 3(e) and the low-contrast hair region in Figure 4(f). In such cases, the proposed method is also able to generate reasonable probability maps with fine details for more accurate labeling results.

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3. Face labeling results and semantic edge maps from the LFW-PL dataset. First row: test images; Second row: edge term output; Third row: unary term output; Forth row: labeling result by graphcut inference, denoted as GC; Fifth row: ground truth. Best viewed in colors.



(a) (b) (c) (d) (e) (f) (g) (h)

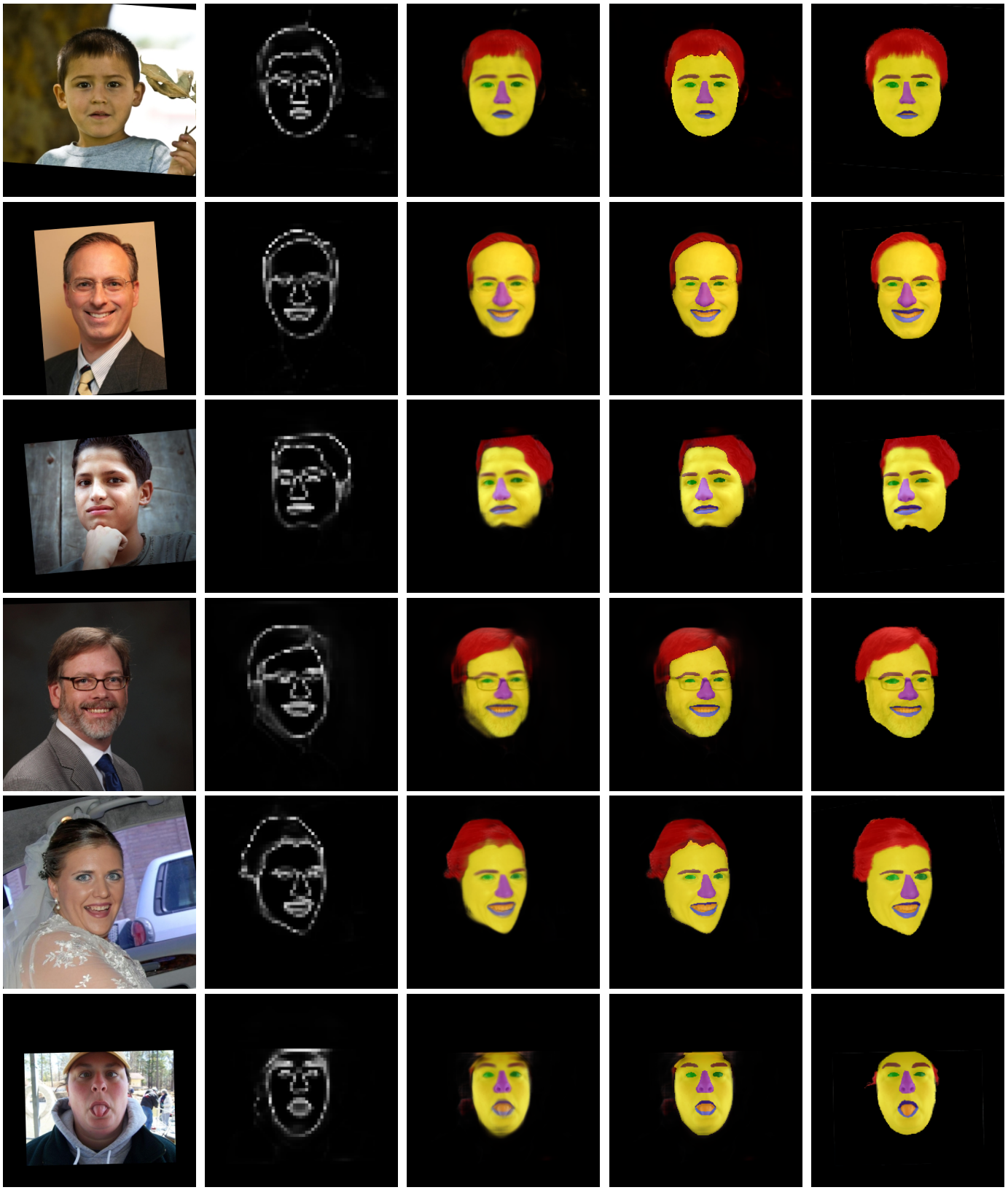
Figure 4. Face labeling results and semantic edge maps from the LFW-PL dataset. First row: test images; Second row: edge term output; Third row: unary term output; Forth row: labeling result by graphcut inference, denoted as GC; Fifth row: ground truth. Best viewed in colors.

5. More Results on the HELEN Dataset

Figure 5 shows more experimental results on the HELEN dataset with 11 labels such that the hair can be illustrated. Unlike the superpixel-based annotation in the LFW-PL dataset, the hair in the HELEN dataset is annotated by matting with a "soft mask" that ranging from 0 to 1, as shown in the fifth row in Figure 5. To be consistent with the ground truth, we also visualize hair regions with "soft masks" generated by unary probabilistic output maps, while keeping the other classes with "hard masks", as shown on the fourth row of Figure 5. Our approach generates visually pleasant labeling results (including hair) and semantic edges.

References

[1] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 1



(a) image (b) edge (c) unary (d) GC (labeling) (e) ground truth

Figure 5. Face labeling results and semantic edge maps from the HELEN dataset. GC denotes labeling result by graphcut inference. Best viewed in colors.