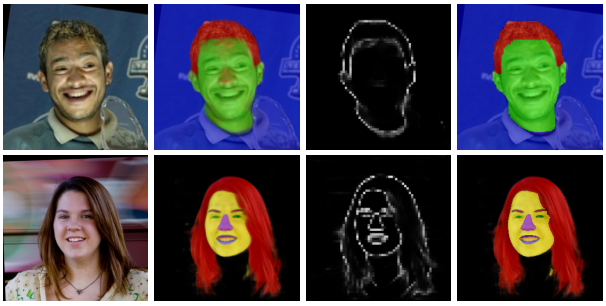
Multi-Objective Convolutional Learning for Face Labeling

Sifei Liu[1], Jimei Yang[1], Chang Huang[2], Ming-Hsuan Yang[1],
[1]University of California, Merced, [2]Baidu Research.



(a) input (b) unary (c) edge (d) label

Figure 1: Face labeling on the LFW [2] and Helen [1] (a) input images. (b) pixel-wise label likelihoods. (c) semantic edge maps. (d) face labeling results.

Face labeling parses an input face image into semantic regions, e.g. eyes, nose and mouth for future processing. Unlike existing methods that define a set of landmarks along face contours and facial components [7], face labeling provides a more robust representation by assigning a semantic label to every pixel of a face image, and has several advantages: First, it is not sensitive to pose, shape, illumination variations and occlusions. Second, it solves the hair region parsing, which is never addressed in landmarks-based methods. For face labeling, three main challenges need to be addressed: (a) How to obtain the per-pixel labeling by combining a per-pixel likelihood and a labeling dependency of neighboring pixels? (b) How to introduce facial priors to regularize the labeling model? (c) How to efficiently infer labelings?

In this work, we formulate face labeling as a conditional random field with unary and pairwise classifiers. A multi-objective convolutional learning method is developed by decomposing the structured loss of conditional random fields (CRFs) into two distinct, non-structured losses: one encoding the unary label likelihoods and the other encoding the pairwise label dependencies:

$$\begin{aligned} &\min_{\omega}\{O_u(\omega,\omega_u),O_b(\omega,\omega_b)\},\\ &\begin{cases} O_u(\omega,\omega_u)=\mathbb{E}(\sum_{i\in\mathcal{V}}L_u(y_i,\mathbf{x}_i,\omega,\omega_u))+\Psi(\omega,\omega_u)\\ O_b(\omega,\omega_b)=\mathbb{E}(\sum_{i,j\in\mathcal{E}}L_b(z_{ij},\mathbf{x}_{ij},\omega,\omega_b))+\Phi(\omega,\omega_b)\end{cases}\end{aligned} \tag{1}$$

where $O_u(\omega,\omega_u)$ is the expected loss for the unary classifier, and $O_b(\omega,\omega_b)$ is the expected loss for the binary classifier over all the training samples. We denote the parameters of the shared CNN network by ω, and those not shared from the topmost intermediate layer of CNN by (ω_u,ω_b). In addition, $\Psi(\omega,\omega_u)$ and $\Phi(\omega,\omega_b)$ are regularization terms. The network is updated through combining gradients of both the softmax and logistic loss functions for backpropagation. Weight sharing is enforced between them so that the network is strengthened by learning from both objectives. More details and expressions can be found in Section 3.3. Compared to structured loss CNNs, our method has two advantages. First, the training process is as efficient as existing CNNs with non-structured losses. Second, by converting the edge term into a logistic loss (edge versus non-edge), semantic image boundaries are learned for effective labeling. Our training pipeline is based on a sliding window input [4, 6] with overlapping patches, as shown in Figure 2. In the testing stage, we convert the original energy function, including the unary term $E_u(y_i)$ and the pairwise term $E_b(y_i,y_j)$, into a submodular one, so that the GraphCut algorithm can be used for efficient inference,

$$\min E(\mathbf{Y})=\sum_{i\in\mathcal{V}}E_u(y_i)+\sum_{i,j\in\mathcal{V}}E_b(y_i,y_j)\mathbb{I}(y_i\neq y_j), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Since convolutional operations share computations between overlapped patches, an efficient strategy introduced

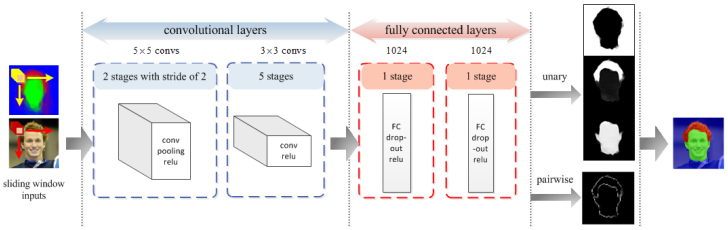


Figure 2: Proposed CNN classifier with sliding window based inputs.

in [4] is proposed by replacing the fully connected layers with equivalent convolutional layers. The full-convolutional model is applied directly to a test image. In addition, an efficient adaptive inference method is proposed to ensure full-sized labeling.

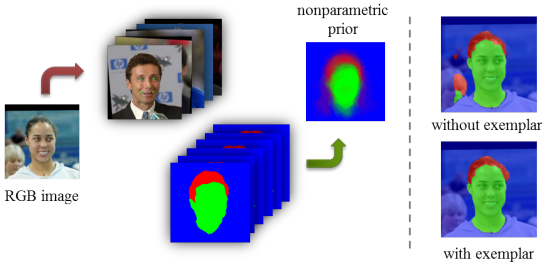


Figure 3: An nonparametric prior is proposed based on label transfer, as shown on the left. A typical labeling improvement is show on the right.

We also integrate a global facial prior into our learning model, as shown in Figure 3. The global facial prior is estimated by transferring face labeling masks (ground truth) from exemplars through landmark detection [5]. Unlike existing methods [3, 5, 8] that use this nonparametric prior at the inference stage, our method uses it as additional input channels, other than raw RGB image intensities, to train a CNN, which significantly improves the performance and reduces the network size.

Experiments on face labeling tasks show that the proposed multi-objective learning and the nonparametric prior significantly improves the labeling performance. We show that accurate labeling results on challenging images can be obtained by the proposed algorithm for real-world applications.

[1] http://www.ifp.illinois.edu/ vuongle2/helen/.

[2] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller. Augmenting CRFs with Boltzmann machine shape priors for image labeling. In *CVPR*, 2013.

[3] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *IEEE PAMI*, 33(12):2368–2382, 2011.

[4] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[5] Brandon Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *CVPR*, 2013.

[6] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *arXiv preprint arXiv:1406.2984*, 2014.

[7] Zhu Xiangxin and Ramanan Deva. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.

[8] Jimei Yang, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Exemplar cut. In *CVPR*, 2013.

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