Exemplar-Based Face Parsing

Brandon M. Smith[1] Li Zhang[1] Jonathan Brandt[2] Zhe Lin[2] Jianchao Yang[2]
[1]University of Wisconsin - Madison [2]Adobe Research

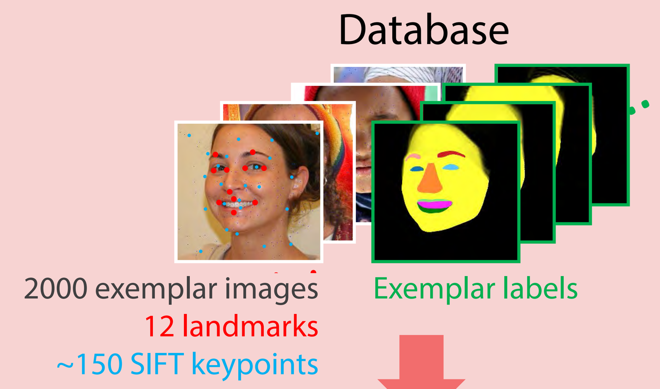
This work is supported in part by NSF IIS-0845916, NSF IIS-0916441, a Sloan Research Fellowship, a Packard Fellowship for Science and Engineering, Adobe Systems Incorporated, and an NSF Graduate Research Fellowship.

Motivation

A common task in face image analysis is parsing an input face image into facial parts, e.g., left eye and upper lip. Most previous methods accomplish this task by marking a few landmarks or contours on the input face image. Instead, we seek to mark each pixel on the face with its semantic part label; that is, our algorithm parses a face image into its constituent facial parts.

	Previous Landmarks, Contours	Ours Per-Pixel Label Probability
Pros	• Vectorized representation	• Encodes ambiguity • Generalizes to hair, teeth, ears, etc. across datasets
Cons	• Ambiguous localization • Inconsistent definitions across datasets	• Not vectorized, but can be combined with landmarks and contours

Our Approach



Runtime Pre-Processing

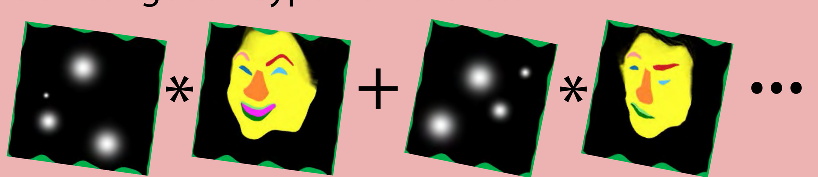
Extract dense SIFT descriptors in the input image. Search for a subset of top exemplar faces, each associated with a similarity transformation that aligns the exemplar face to the input face.

Step 1: Nonrigid Exemplar Alignment

For each keypoint in each top exemplar, perform a local search in the input image to find the best match; record the matching score. Warp the label map of each exemplar nonrigidly using a displacement field interpolated from the match location offsets.

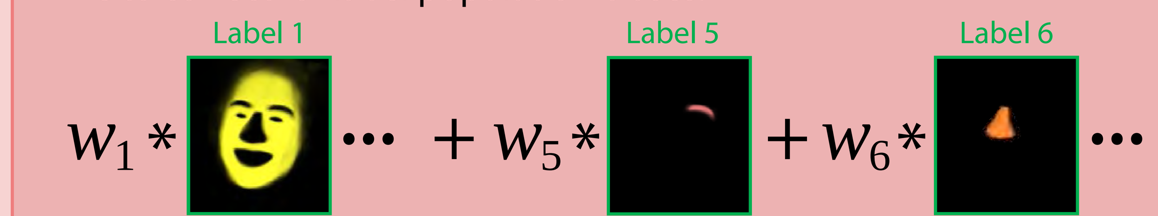
Step 2: Exemplar Label Aggregation

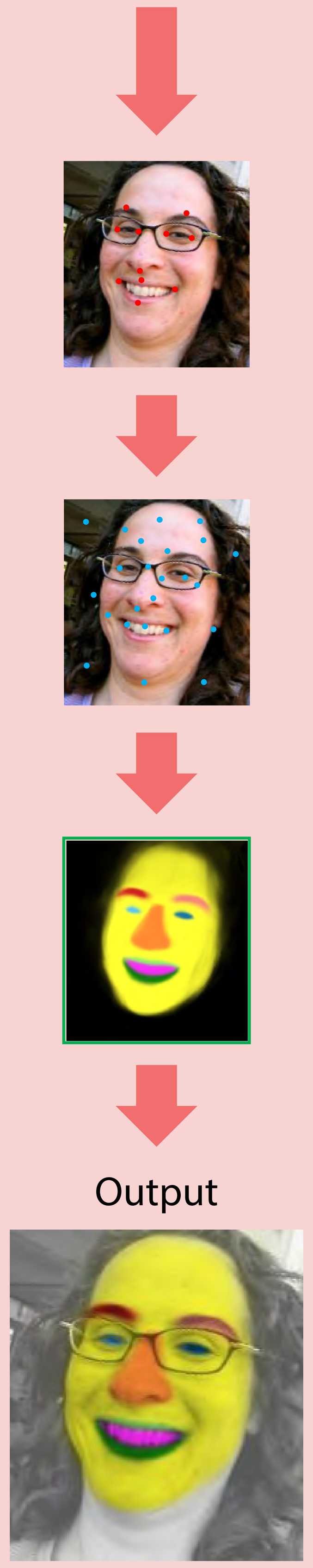
Aggregate warped label maps using weights derived from the keypoint matching scores in Step 1. The weights are spatially varying and favor exemplar pixels near good keypoint matches.

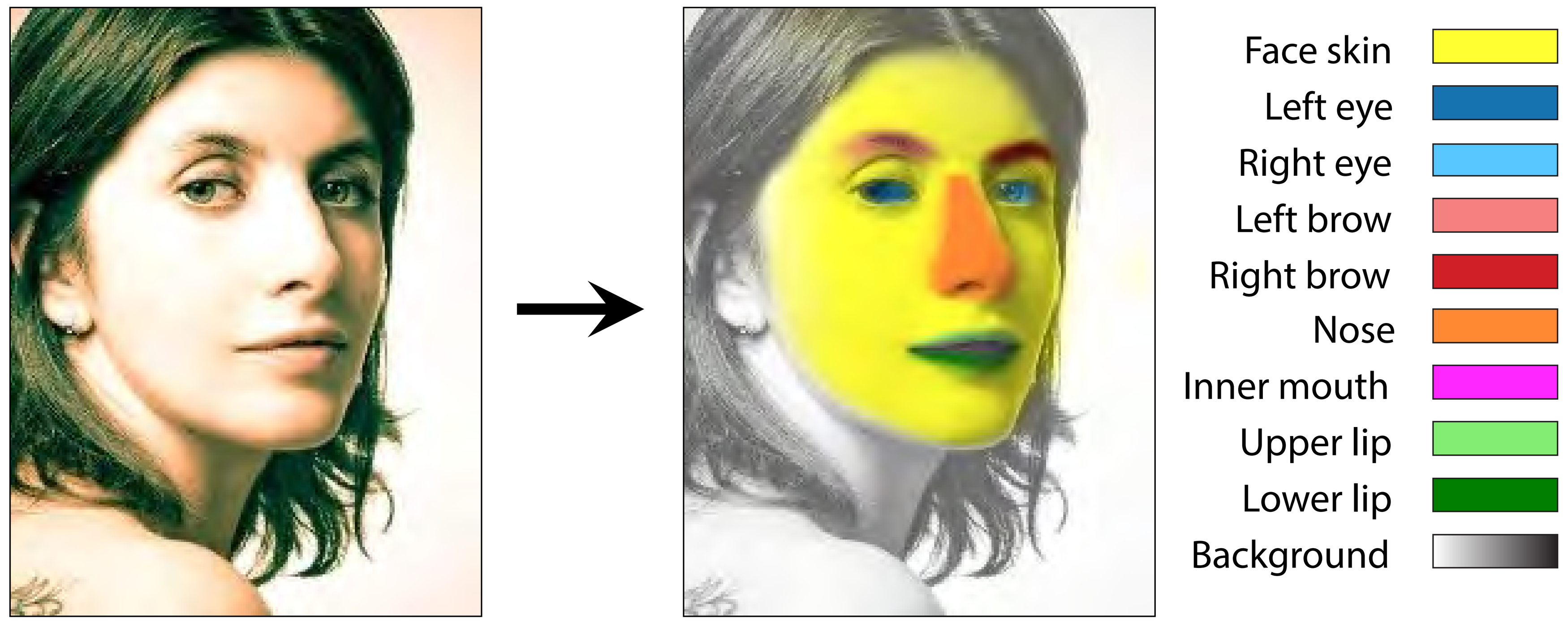


Step 3: Pixel-Wise Label Selection

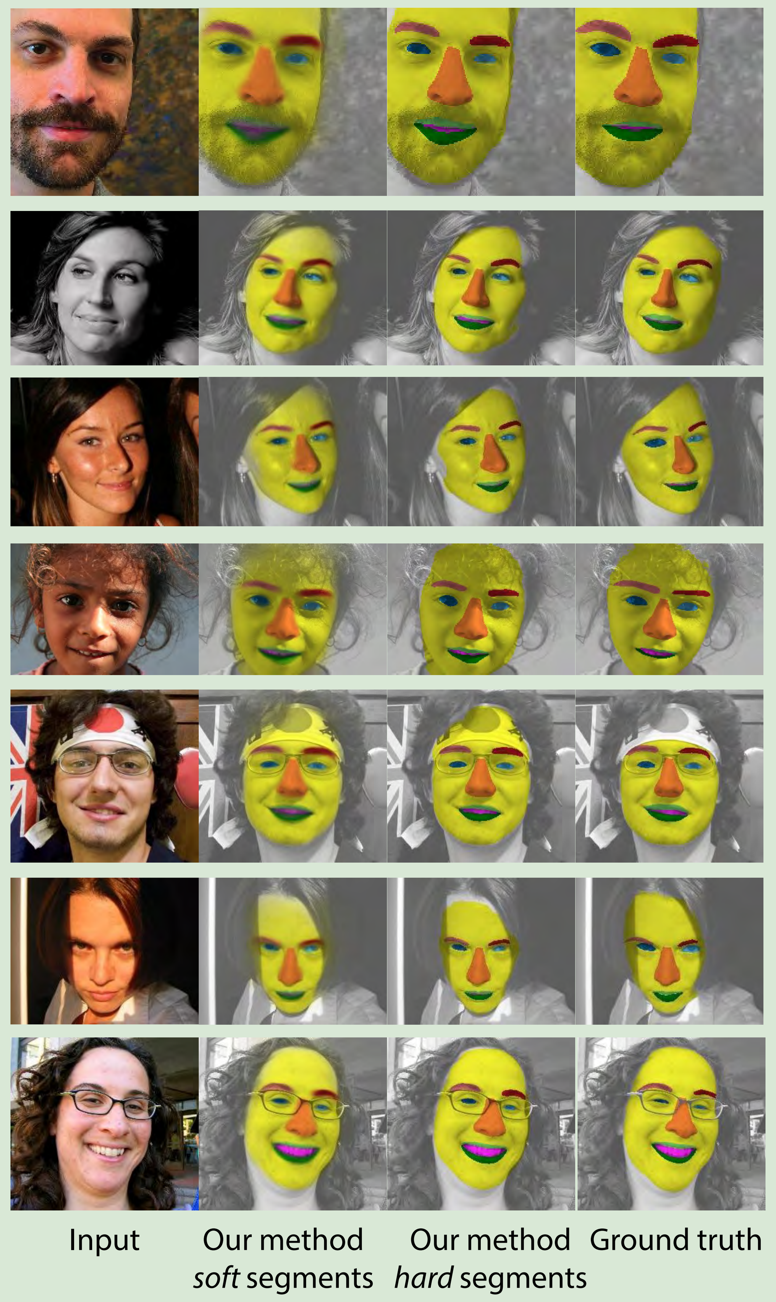
Produce a label probability vector at each pixel by attenuating each channel in the aggregated label map. The attenuating weights are trained offline to correct for label population biases.



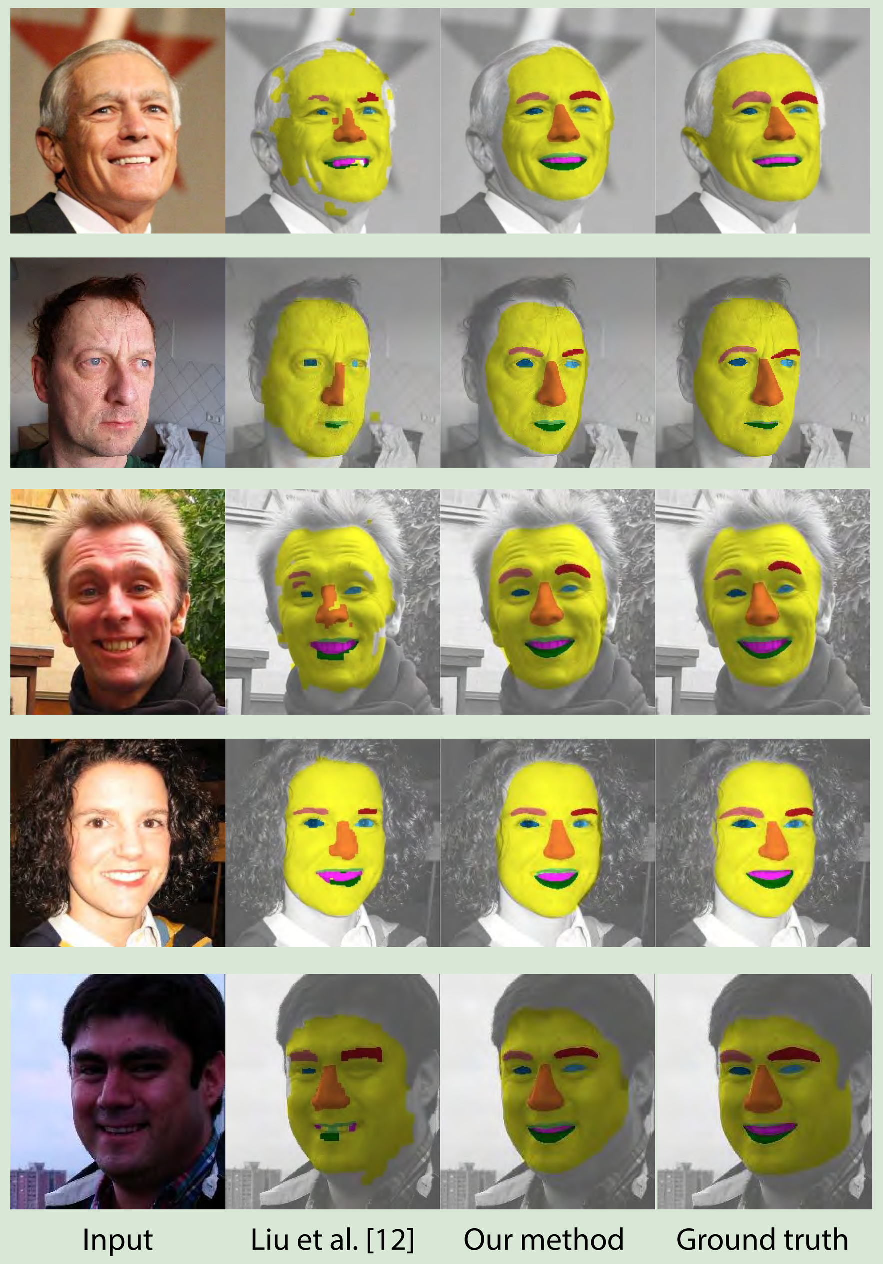




Qualitative Results

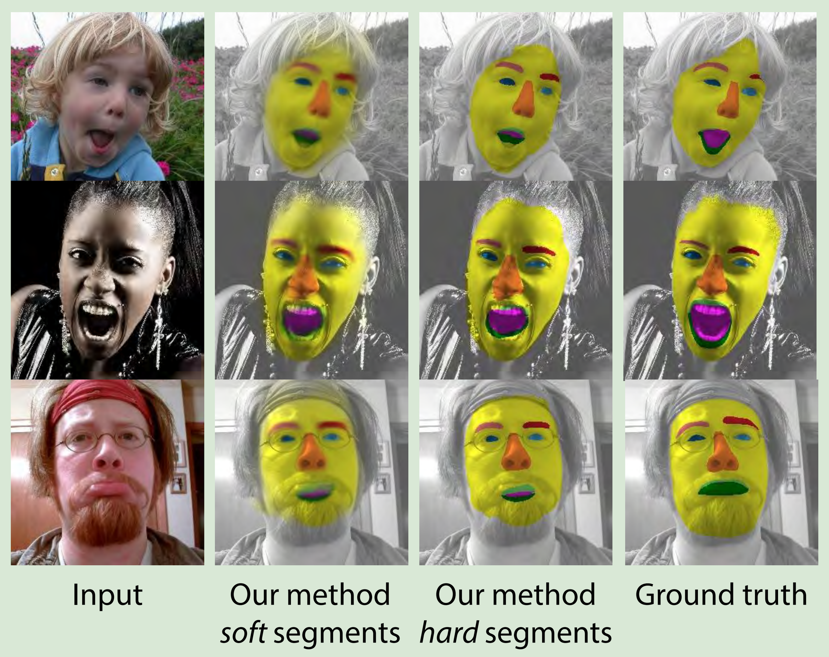


Our algorithm generally produces accurate results.



The segments generated by Liu et al.'s nonparametric scene parsing algorithm are visibly less accurate, especially in the mouth region. This suggests that a general scene parsing approach is not well suited to faces.

Failure Cases on Mouths Due to Insufficient Exemplars



Large segmentation errors occur infrequently, but when they do occur, errors are almost always localized to the mouth region. Unusual mouth expressions like those shown above are not represented well in the exemplar images, which results in poor label transfer from the top exemplars to the test image.

Quantitative Results

Confusion Matrix Comparison

	left eye	right eye	nose	left brow	mouth	right brow	background
left eye	**.90**			.01			.09
right eye		**.93**				.01	.06
nose			**.88**		.01		.11
left brow	.03			**.91**			.06
mouth					**.90**		.10
right brow		.02				**.89**	.09
background			.01		.04		**.95**

(a) Results from Liu et al. [15]

	left eye	right eye	nose	left brow	mouth	right brow	background
left eye	**.990**			.003			.007
right eye		**.990**				.002	.008
nose			**.992**	.001		.001	.006
left brow	.002			**.988**			.010
mouth			.001		**.983**		.016
right brow		.003				**.982**	.015
background	.002	.002	.004	.006	.005	.006	**.975**

(b) Our results

Based on the confusion matrix, our results look much more accurate than the same results from Liu et al. [15]. However, this metric can be deceiving (see right).

(c) Estimated (d) Ideal

The result in (c) exemplifies the problem with the label weights used to maximize the diagonal of the confusion matrix. We instead show accuracy using the F-measure (harmonic mean of precision and recall) and we optimize label weights to maximize the F-measure.

F-Measures for LFW Images

Method	Eyes	Brows	Nose	Mouth	Overall
Warrell & Prince [21]	0.443	0.273	0.733	0.653	n/a
Zhu & Ramanan [22]	0.520	n/a	n/a	0.635	n/a
Saragih et al. [18]	0.684	0.651	0.903	0.753	0.793
Gu & Kanade [4]	0.735	0.722	0.900	0.801	0.820
Ours	**0.765**	**0.752**	**0.914**	**0.881**	**0.863**

Comparison with a face parsing algorithm (Warrell & Prince), and three face alignment algorithms (segments were derived from the contours generated by these algorithms).

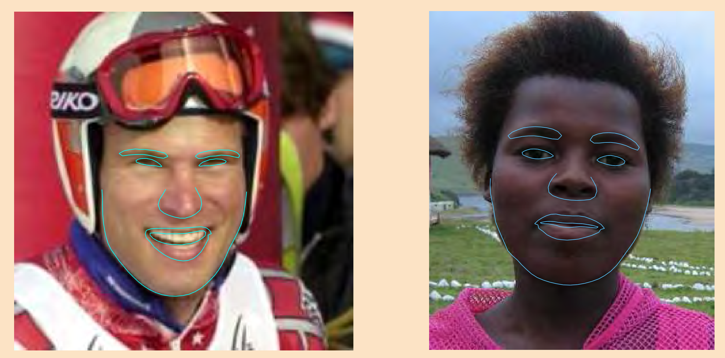
F-Measures for Helen Images

Method	Eyes	Brows	Nose	In Mouth	Upper Lip	Lower Lip	Mouth(all)	Face Skin	Overall
Zhu & Ramanan [22]	0.533	n/a	n/a	0.425	0.472	0.455	0.687	n/a	n/a
Saragih et al. [18]	0.679	0.598	0.890	0.600	0.579	0.579	0.769	n/a	0.733
Liu et al. [12]	0.770	0.640	0.843	0.601	0.650	0.618	0.742	**0.886**	0.738
Gu & Kanade [4]	0.743	0.681	0.889	0.545	0.568	0.599	0.789	n/a	0.746
Ours, omit Steps 1, 3	0.766	0.687	0.896	0.678	0.637	**0.703**	0.853	0.861	0.779
Ours, omit Step 3	0.772	0.708	0.914	0.659	0.639	0.697	0.850	0.872	0.790
Ours, full pipeline	**0.785**	**0.722**	**0.922**	**0.713**	**0.651**	0.700	**0.857**	0.882	**0.804**

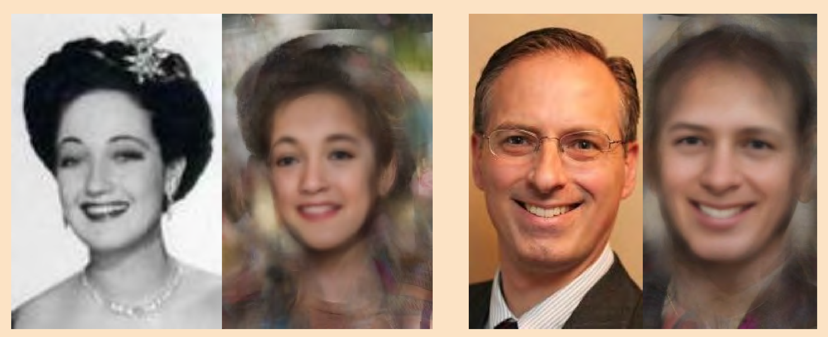
Liu et al. is a nonparametric scene parsing algorithm. The only area where Liu et al.'s system is more accurate than ours is on the face skin. The difference is primarily due to our algorithm incorrectly hallucinating skin in hair regions, while Liu et al.'s system does not. In general, we see that our algorithm compares favorably to all previous works on this dataset, and our full pipeline performs best overall.

Extensions of Our Approach

Contour Estimation



Face Image Synthesis and Reconstruction



We can synthesize the input face by replacing the exemplar label vectors with the color channels from the exemplar images.

Hair Segmentation

Our automatically generated "seeds" for hair are shown in purple; background is shown in blue. Hair mattes are computed from these seeds using an automatic matting algorithm.