Towards Unconstrained Face Recognition

Gary B. Huang Manjunath Narayana Erik Learned-Miller
University of Massachusetts Amherst
Amherst, MA
{gbhuang,narayana,elm}@cs.umass.edu

Abstract

In this paper, we argue that the most difficult face recognition problems (unconstrained face recognition) will be solved by simultaneously leveraging the solutions to multiple vision problems including segmentation, alignment, pose estimation, and the estimation of other hidden variables such as gender and hair color. While in theory a single unified principle could solve all these problems simultaneously in a giant hidden variable model, we believe that such an approach will be computationally, and more importantly, statistically, intractable. Instead, we promote studying the interactions among mid-level vision features, such as segmentations and pose estimates, as a route toward solving very difficult recognition problems. In this paper, we discuss and provide results showing how pose and face segmentations mutually influence each other, and provide a surprisingly simple method for estimating pose from segmentations.

1. Introduction

Work has recently begun on the difficult problem of face recognition in unconstrained environments [1, 2, 5, 10]. While there has been tremendous progress in face recognition under carefully controlled conditions [15], machine performance on the problem of unconstrained face recognition is still poor. Recently, a database specifically for studying the problem of unconstrained face recognition, Labeled Faces in the Wild (LFW), has been published [4].

It is interesting to note that the best performers on unconstrained face recognition data sets (discussed in more detail below) do not use any face structure, any head structure, or any high-level features that explicitly encode knowledge of face or head parts or structure. For example, the best performing models do not have an explicit representation of eyes, nose, mouth, hair, skin, or any other face part. They do not explicitly estimate pose, gender, or other hidden variables either.

From one point of view, such generic models are quite impressive: they have won the accuracy competition so far with no hand-specification or learning of parts. On the other hand, it seems that some representation of "natural" hidden variables such as pose or face parts, is likely to improve results if done in the right way. As we will show in this paper, an analysis of the errors of the best face recognizers suggest they could be greatly improved by using knowledge of certain high level nuisance variables such as pose, gender, hair color, and so on.

In this paper, in accordance with the goals of the workshop, we describe a general long-term strategy for integrating the solutions to a variety of vision problems related to face recognition. In particular, we believe that estimating alignment, pose, gender, and head parts such as hair and skin must all be done jointly with recognition in order to achieve the highest recognition rates.

In addition to introducing this long term strategy, we provide a variety of specific results toward this end. Specifically, we give results for hair-skin-background segmentation on the LFW database, we show how ground-truth segmentations allow us to estimate pose quite easily, how pose improves segmentation results, and how segmentation estimates give reasonably good estimates of pose.

The structure of the rest of the paper is as follows. In Section 2, we review recent work on the unconstrained face recognition problem. In addition, we provide an informal analysis of the errors typical of these algorithms. These errors naturally suggest our current approach. In Section 3, we discuss the problem of pose estimation and how it has been addressed previously. We observe that perhaps the simplest possible features to establish pose have never been used in a pose estimation algorithm. In Section 4, we discuss our approach to pose estimation, which relies on an image segmentation into hair (including facial hair), skin, and background. We show that such a segmentation gives excellent estimates of pose, and is easily interpretable. In Section 5, we then discuss our method for automatically generating the segmentations used for pose estimates. While they are not as good as the ground truth poses, they are still good enough to provide significant information about pose. In

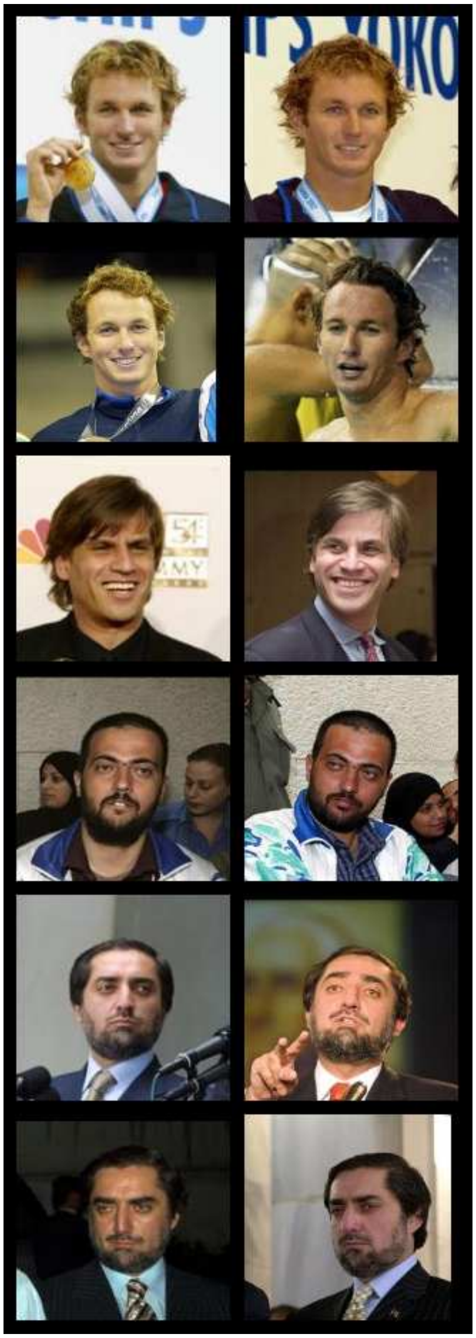

Figure 1. **Labeled Faces in the Wild**. These are the first six matching pairs in the database. Our long term goal is to be able to do pair matching under these challenging conditions.

Section 6, we show that pose, in turn, can improve segmentation estimates. Finally, in Section 7, we discuss other interactions among pose, alignment, segmentations, and other hidden variables such as gender, accessories, hair-length, and so on, that we believe will be leveraged for better results on difficult face recognition problems.

2. Unconstrained Face Recognition

The long term goal of our work is to dramatically improve the performance of face recognition in unconstrained photos. We are less interested in photos of highly unusual situations (like strong lighting from below with no lighting from above), but more interested in the natural variation of pose, expression, lighting, and other appearance factors that occur in the everyday world. We refer to this as the unconstrained face recognition problem. Recently, a large database of face images called Labeled Faces in the Wild (or LFW) has been published [4], specifically for studying the problem of unconstrained face recognition. LFW is used in the work here as a test bed for this type of problem.

The best *recognition* performance to date on LFW has been achieved using the randomized forests algorithm of Nowak et al. [10]. It is interesting to note that this algorithm works well not only on classifying pairs of faces as "matching" or "not matching" but also on many other types of data such as pairs of cars. While the performance of this algorithm is very impressive, it raises the question of whether algorithms that more explicitly model the structure of faces have the potential to do better. By "explicitly modeling" the face, we do not mean to imply that an algorithm should be hand built without learning the structure, but that an explicit attempt to model the structure within faces (in either a supervised or unsupervised fashion) may lend significant power to a face classification algorithm.

To support this hypothesis, consider in Figure 2 the pairs of LFW images that are misclassified (for most locations on the precision-recall curve) by the Nowak recognizer. The left two columns show pairs of pictures that were incorrectly labeled as "not matching" while the right two columns show pairs of pictures that were incorrectly labeled as "matching". It is perhaps not surprising that the matching pairs that were incorrectly labeled as mismatched were in significantly different poses. Recognition across pose is an extremely difficult problem, and may be one of the most difficult aspects of face recognition.

However, much more surprising are the mismatched images which are incorrectly labeled as matching. To a person, such mismatched pairs are egregious errors, and would simply never happen with a human in the loop. In the first case, the hair of the two women is completely different. In addition, the woman on the right is clearly older than the woman on the left. In the second case, while slightly more subtle, it is clear that the man on the right has significantly less hair than the man on the left. In addition the man on the left has a moustache. Such features are obvious to human observers, and yet are not used by the best current recognition algorithms.

One reason for this may be that it is difficult to estimate these features definitively. However, modern probabilistic methods will allow us to give estimates of the likelihood of

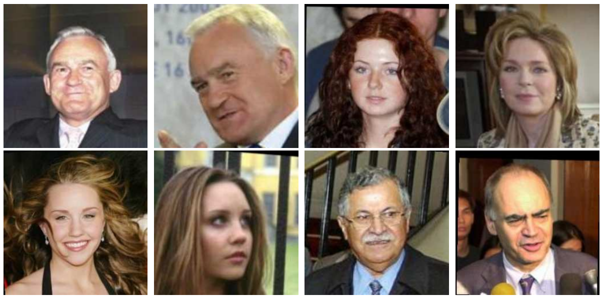

Figure 2. **Errors by the top performing face recognition algorithm.** The pairs on the left were incorrectly identified as "not matching". The pairs on the right were incorrectly identified as "matching". The latter errors are particularly egregious, and exhibit the ignorance of the top recognition algorithms of higher level features such as hair color, hair texture, degree of baldness, and presence of moustache.

these features (e.g., the probability of a moustache), which can allow an algorithm to benefit from the features even when there is remaining uncertainty.

3. Estimating Pose

Understanding the pose of a person's face seems as though it is likely to be a natural part of the recognition process. We note informally that it is particularly easy for people to understand the approximate pose of almost any recognizable face. The problem of identifying pose can be formulated in many ways. One may try to identify the precise roll, pitch, and yaw angles of the head. Often pose algorithms, assuming little roll or pitch, simply focus on the yaw angle (rotation about the vertical axis). Simplifying even further, one may classify pose into left-facing, right-facing, or frontal. Often, alignment (or translation) is considered a part of pose, but here we shall consider pose as simply the rotational part of the head position.

Like establishing alignment, understanding the pose of a head is valuable conditioning information, as discussed below. But how can we establish the pose of a person's head under very general imaging conditions?

Many pose estimation methods [3, 12] focus on facial features such as eyes, ears, and corners of the mouth. These methods crop out the center of each face and learn or estimate a function from pixels or estimated facial features to the final pose. We believe these algorithms are, in fact, eliminating the parts of the image most informative of pose. In particular, we will show below that good pose estimates can be obtained very simply from a good segmentation of the image into skin, hair, and background. We will show that if we can obtain such a segmentation automatically, we can obtain very strong pose information in addition to high level features about skin and hair that should be useful in recognition.

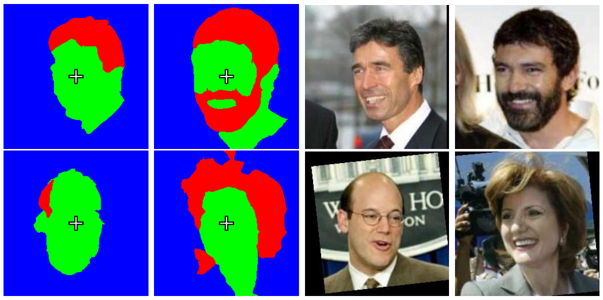

Figure 3. **Hand-segmented images.** It is easy to guess the approximate pose of these faces from their segmentations, by using the relative position of hair and skin segments as well as the proportion of the face and skin to the right of the center.

Some pose algorithms do not focus on facial features, but instead use very general learning techniques on the entire image [11]. While the results of such algorithms have been quite good, we believe similar and more interpretable results can be obtained by using soft segmentations as higher level features into a simpler algorithm. Some drawbacks of such low-level learning methods for such complex problems is that results are difficult to reproduce, they require large amounts of training data and relatively long training times.

4. Pose Estimates from Segmentation

In this section, we consider the question of how well we can estimate pose given segmentations of face images into three regions: skin (of the central subject's face and possibly the neck), hair (including beards and moustaches) and background. The background may contain other faces, and other regions that are difficult to distinguish from facial skin, such as the skin on a subject's arms.

This work stems from a very simple observation, which is that it is often quite easy for a person to estimate the approximate pose of a person merely from a segmentation of the person's photo. We show four such images in Figure 3.

More specifically, it is interesting to note that if one marks the point half way between the eyes as an "origin" of the face, then frontal faces tend to have a nearly equal balance of skin pixels and hair pixels on either side of the origin. As a person's head turns to the side, more skin and hair become visible to the observer on one side of the head, and less skin and hair become visible on the other side. This leads to a very simple algorithm for estimating pose from a segmentation.

Further simplifying the process is the fact that the LFW database was developed so that each face appears centered in the image according to a detection by the Viola-Jones face detector [13]. Since this face detector tends to center faces about the middle of the eyes, it is even easier to use the proposed method.

Actual	Right	Frontal	Left
Right facing	19	3	0
Frontal	4	26	10
Left-facing	3	6	29

Table 1. Pose estimates from hand-labeled segmentations: training data.

Actual	Right	Frontal	Left
Right facing	26	4	0
Frontal	2	16	18
Left-facing	0	3	31

Table 2. Pose estimates from hand-labeled segmentations: test data.

We use a very simple polynomial regression scheme for estimating pose (yaw only) in which our two basic features are the first moments of skin and hair pixels about the center line. Starting with 100 labeled segmentations such as those shown in Figure 3, we compute the "first moments" of hair and skin about the vertical line running through the middle of the image (shown by a cross). The "hair feature" f_h is given by

$$f_h = \frac{\sum_{H_R} p_x}{\sum_H p_x},$$

where H_R is the set of pixels labeled "hair" in the right half of the image, H is the set of all pixels labeled hair, and p_x is the horizontal distance of a pixel from the center line of the image. A "skin feature" s_h is computed similarly.

Using these features and their squares and cross terms, we perform a polynomial regression on pose values (given in degrees) on a training set. Since our goal is ultimately to use pose to help in classification, rather than to perform accurate pose estimates, we evaluate our pose estimates by classifying each face as left-facing, right-facing, or frontal.

Using this simple regression scheme, we obtained the training and testing performance shown in Tables 1 and 2.

These pose estimates are adequate for many purposes. In particular, note that in the test data there were no confusions at all between left and right-facing poses. Since poses can be estimated fairly well from segmentation data, the next obvious question is how well one can estimate the segmentations from which these poses were derived.

To estimate image segmentations, we started with the superpixel representations provided as part of the LFW database. (These superpixel representations also made it particularly easy to do manual labeling of our image segmentation data.) The basic intuition behind our segmentation method is that, when segmenting many images at once, information can be shared across segmentations. For example, we expect hair to be in roughly the same locations in most images. The same is true for hair and for background superpixels. This observation has been leveraged by other authors [7].

5. Estimating Segments

We use a Conditional Random Field (CRF) [6] to estimate a segmentation. Our CRF encodes the probability of a segmentation Y given features X of an image. $Y = \{y_1, \ldots, y_n\}$ where n is the number of superpixels in the image, and y_i can take on one of three values corresponding to "background", "face", and "hair". X consists of node features X^n and edge features X^e. For each superpixel i, we compute F_n features, so X_i^n is a vector of length F_n. Similarly, for each pair of neighboring superpixels i, j, we compute F_e features, so $X_{i,j}^e$ is a vector of length F_e.

We use a log-linear CRF, with node energies $\psi(y_i, X_i^n)$ and edge energies $\psi(y_i, y_j, X_{i,j}^e)$ as follows

$$\psi(y_i = l, X_i^n) = \sum_{f=1}^{F_n} (W_l^n)_f (X_i^n)_f \qquad (1)$$

$$\psi(y_i = l_1, y_j = l_2, X_{i,j}^e) = \sum_{f=1}^{F_e} (W_{l_1,l_2}^e)_f (X_{i,j}^e)_f \qquad (2)$$

W^n is a set node weights, one vector of length F_n for each label l. W^e is a set of edge weights, one vector of length F_e for each pair of labels (l_1, l_2). We use a symmetric edge potential by letting $W_{l_1,l_2}^e = W_{l_2,l_1}^e$.

Putting this together, the probability of a segmentation Y given X is

$$p(Y|X) = \frac{\exp(-\sum_{i=1}^n \psi(y_i, X_i^n) - \sum_{(i,j)} \psi(y_i, y_j, X_{i,j}^e))}{Z(X)} \qquad (3)$$

where the second sum is taken over neighboring superpixels, and $Z(X)$ is the partition function that normalizes the distribution.

For node features, we compute color, position, and texture features, as well as a constant bias feature. For color, we compute a normalized color histogram for each superpixel in Lab space, using 64 bins computed from kmeans clustering. For position, we overlay an 8x8 grid on the image and compute the proportion of pixels within each square for each superpixels. For texture, we compute 64 textons as in [8] and compute a normalized histogram for each superpixel. Thus, we have $F_n = 193$.

For edge features, we compute $F_e = 3$ features. We compute Euclidean distance in Lab space between the mean colors of neighboring superpixels. We compute the probability of boundary (Pb) [9] at each pixel, and sum the values over all pixels along the border of two superpixels. We also compute the χ^2 test for the texture histograms h_1 and h_2 of neighboring superpixels, using

$$\chi^2 = \frac{1}{2}\sum_{i=1}^{64}\frac{(h_1(i)-h_2(i))^2}{h_1(i)+h_2(i)} \tag{4}$$

In practice, we found that using edge potentials gave approximately a half percentage increase in accuracy over a CRF with only node potentials (corresponding to logistic regression). For the node potential only CRF, adding texture gave approximately a half percentage increase over only using color and position features.

Due to cycles in the CRF caused by the edge potentials, computing the partition function $Z(X)$ is intractable. Therefore, to learn the weight potentials, we optimize the log likelihood using the Bethe approximation for the partition function and loopy belief propagation to approximate marginals for each y_i [14]. We use an implementation of L-BFGS[1] to do the optimization, and also add a Gaussian prior on the weights for regularization.

To estimate a segmentation, we use loopy belief propagation to compute the maximum posterior marginals (MPM).

Some resulting segmentations are shown in Figure 4. These segmentations are typical of the results of our algorithm. The top two segmentations are quite accurate, and provide important information for obtaining mid-level features for classification such as hair color and whether there is a beard or not. The third segmentation, while not great, still allows estimation of pose and may still be useful for classification. The last segmentation is somewhat misleading, and throws off both pose estimates and estimates of other quantities such as amount of hair, leading to poor estimates of hidden variables such as gender.

To assess the accuracy of our segmentation estimates, we adopted an L1 error on the segmentation estimates, essentially penalizing each superpixel according to the difference between 1.0 and the probability of the correct label. That is, if a superpixel was given a probability of 0.8 of being hair, and it was in fact hair, then a penalty of 0.2 would be incurred. Using this scheme, we were able to obtain segmentations that were over 90% correct using our CRF estimation procedure.

It is of course interesting to ask how good our pose estimates are when they are based not on perfect hand-labeled segmentations, but rather when they are based on our estimated segmentations. Tables 3 and 4 summarize these results which are nearly as good as those based on manual segmentations. Thus, our estimated segmentations provide almost as good pose estimates, using our regression scheme, as the hand-labeled data.

Actual	Right	Frontal	Left
Right facing	11	10	1
Frontal	7	17	16
Left-facing	1	12	25

Table 3. Pose estimates from estimated segmentations: training data.

Actual	Right	Frontal	Left
Right facing	15	15	0
Frontal	2	18	16
Left-facing	0	10	24

Table 4. Pose estimates from estimated segmentations: test data.

6. Improving Segmentations with Pose Information

In addition to estimating pose from segmentations, it is possible to use the estimated pose to refine segmentations. If we are given the pose of someone's head, this clearly changes our prior notion of where we would expect to see hair, skin, and background relative to the origin of the face (or midpoint between the eyes).

To see whether pose information improved segmentations, we trained two separate models for segmentation, one based on left-facing faces and one based upon right-facing faces. Single node potentials based on position, which can be thought of as something like a prior on the probability of each particular segmentation label, were learned for each model separately. These models are shown in Figure 5.

Using these separate models for segmentation did improve the segmentations in accuracy by about 0.5%. While this is not a dramatic increase, it shows that pose and segmentation can each help each other.

7. Conclusions

We have demonstrated a number of phenomena:

- That segmentations of hair, skin, and background provide enough information to make reasonably good pose estimates, both from a human observer's point of view and using a simple regression technique,

- That pose can be used to improve segmentations by developing separate segmentation models conditioned on pose, and

- A CRF-estimated segmentation is good enough to provide a great deal of information about skin, hair, and background, and in addition can be used to estimate pose. We are not aware of other work which focuses specifically on the problem of segmenting a face into

[1] `http://vis-www.cs.umass.edu/~weinman/code.html`

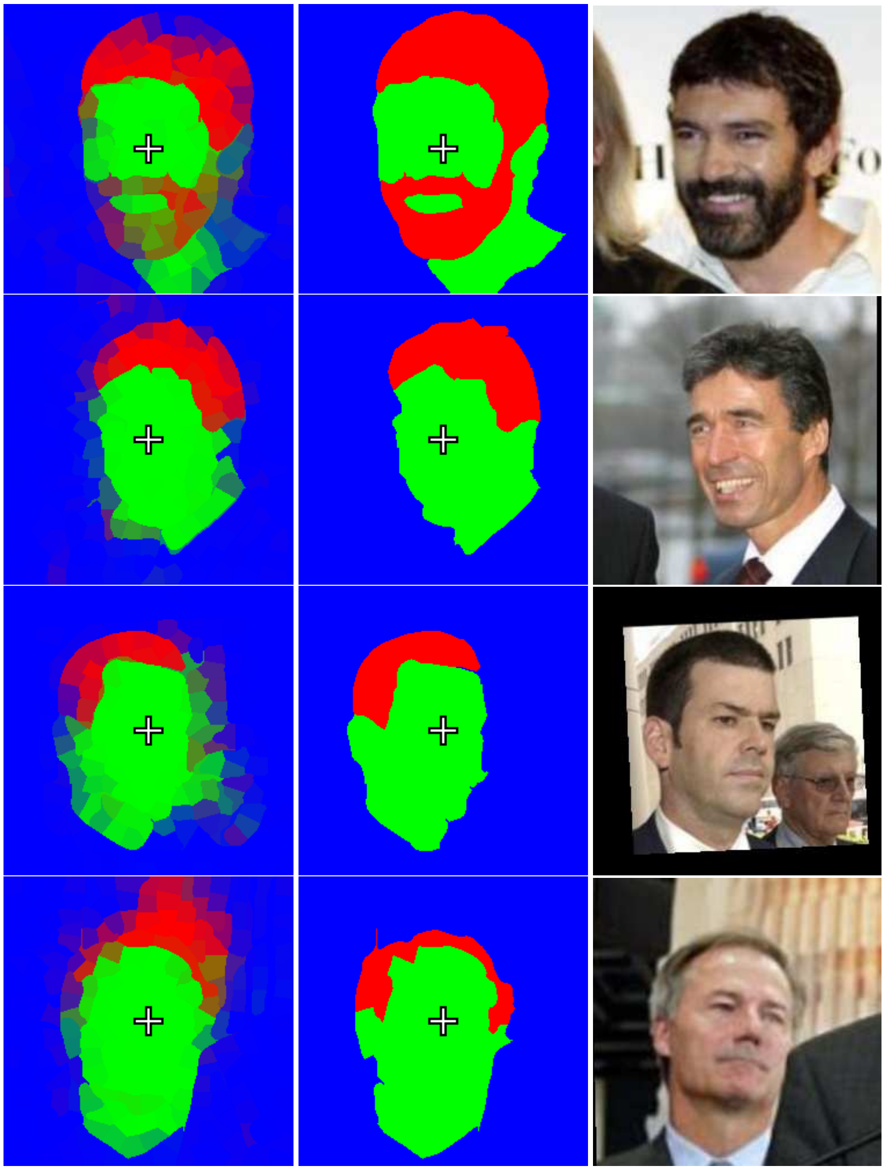

Figure 4. **Estimated segmentations.** These images show the estimated probability of each region, with green representing skin, red representing hair (including facial hair) and blue representing background. Mixed probabilities are simply shown as a mixture of the red, blue, and green channels, so yellow, for example, would show a pixel with probability 0.5 of hair and probability 0.5 of skin.

skin, hair, and background. We note this is quite different, and significantly harder, than segmentation of the skin only, as has been done in a number of papers.

We believe intermediate level features, such as segmentations, that can be estimated quite well when done simultaneously on a set of images (rather than one image at a time) will provide important new sources of information for the difficult problem of face recognition in unconstrained environments.

We have just started to tap this information. We believe it will be extremely useful in estimating a number of other highly informative hidden variables such as gender, age, beardedness, degree of balding, color of hair, color of skin, and so on. These are hidden variables which may be difficult to get at using a monolithic and undifferentiated neural network architecture.

We believe that building up mid-level features in this way will give us not only interpretability, but computational efficiency, statistical efficiency, and ultimately, increased accuracy.

References

[1] A. Ferencz, E. Learned-Miller, and J. Malik. Building a classification cascade for visual identification from one example.

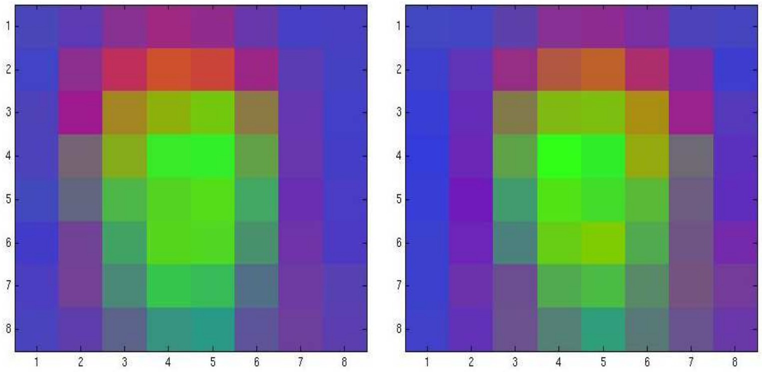

Figure 5. **“Priors” of segmentation label based on left-facing and right-facing subsets of training data.**

In *ICCV*, 2005.

[2] A. Ferencz, E. Learned-Miller, and J. Malik. Learning hyper-features for visual identification. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[3] F. J. Huang and T. Chen. Tracking of multiple faces for human-computer interfaces and virtual environments. In *IEEE International Conference on Multimedia and Exposition*, 2000.

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[5] V. Jain, A. Ferencz, and E. Learned-Miller. Discriminative training of hyper-feature models for object identification. In *Proceedings of the British Machine Vision Conference*, 2006.

[6] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, 2001.

[7] N. Loeff, H. Arora, A. Sorokin, and D. Forsyth. Efficient unsupervised learning for localization and detection in object categories. In *Advances in Neural Information Processing Systems*, 2005.

[8] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *International Conference on Computer Vision*, 1999.

[9] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using brightness and texture. In *Advances in Neural Information Processing Systems*, 2002.

[10] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. In *cvpr*, 2007.

[11] M. Osadchy, Y. L. Cun, and M. L. Miller. Synergistic face detection and pose estimation with energy-based model. *Journal of Machine Learning Research*, 8:1197–1215, 2007.

[12] T. Vatahska, M. Bennewitz, and S. Behnke. Feature-based head pose estimation from images. In *Proceedings of the IEEE-RAS 7th International Conference no Humanoid Robotics*, 2007.

[13] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 2004.

[14] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. Technical Report TR-2001-22, Mitsubishi Electric Research Laboratories, 2001.

[15] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature survey. Technical Report CAR-TR948, University of Maryland, 2000.

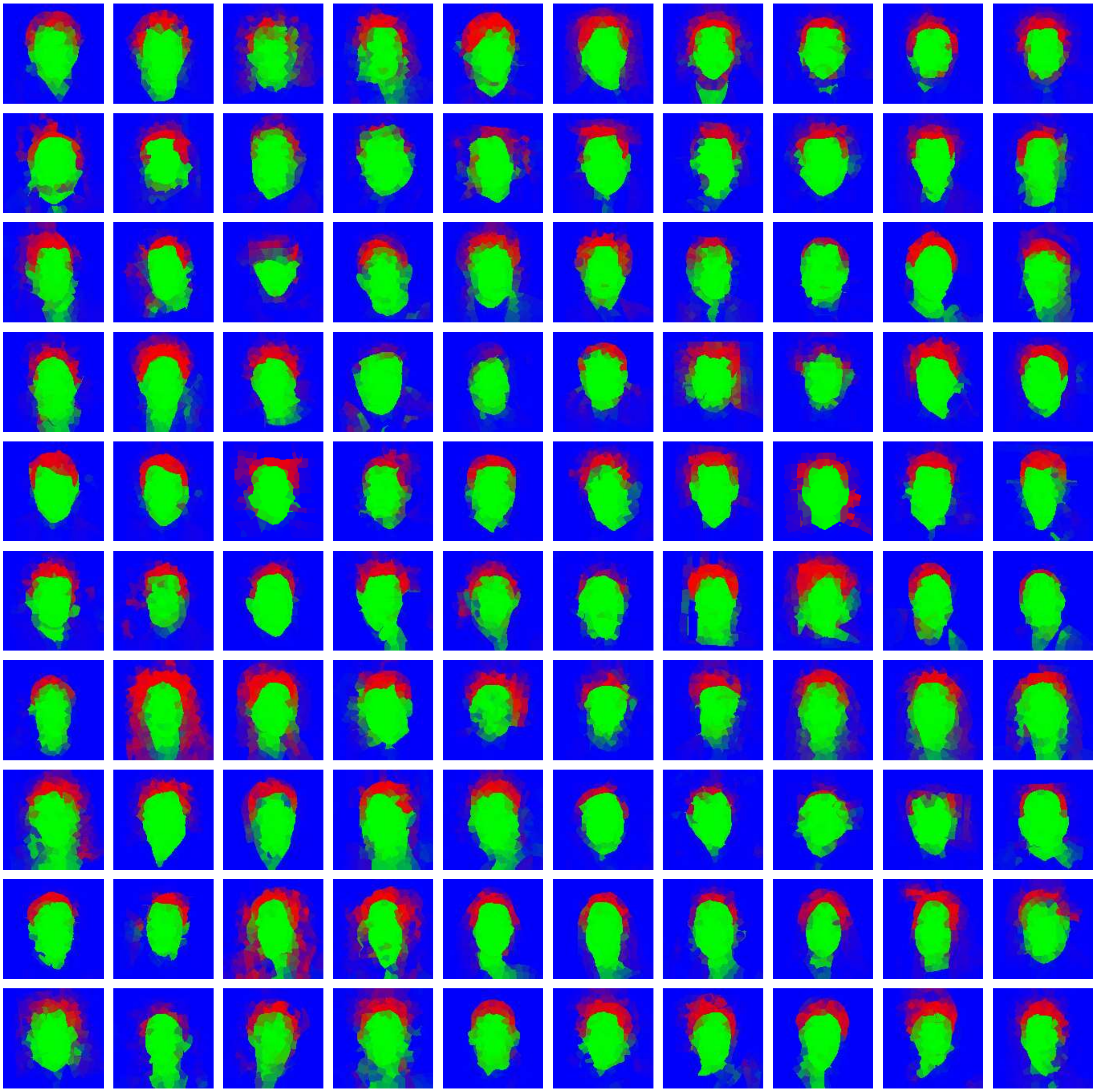

Figure 6. **Additional results of our CRF segmentation algorithm on the LFW database.**