

## **A Conditional Random Field for Automatic Photo Editing**

Matthew Brand, Patrick Pletscher

TR2008-035 July 2008

### **Abstract**

We introduce a method for fully automatic touch-up of face images by making inferences about the structure of the scene and undesirable textures in the image. A distribution over image segmentations and labelings is computed via a conditional random field; this distribution controls the application of various local image transforms to regions in the image. Parameters governing both the labeling and transforms are jointly optimized w.r.t. a training set of before-and-after example images. One major advantage of our formulation is the ability to marginalize over all possible labeling and thus exploit all the information in the distribution; this yield better results than MAP inference. We demonstrate with a system that is trained to correct red-eye, reduce specularities, and remove acne and other blemishes from faces, showing results with test images scavenged from acne-themed internet message boards.

*CVPR 2008*

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.



# A conditional random field for automatic photo editing

Matthew Brand  
Mitsubishi Electric Research Labs  
Cambridge, MA 02139 USA  
<http://www.merl.com/people/brand>

Patrick Pletscher  
ETH Institute of Computational Science  
8092 Zurich, Switzerland  
[pletscher@inf.ethz.ch](mailto:pletscher@inf.ethz.ch)

## Abstract

*We introduce a method for fully automatic touch-up of face images by making inferences about the structure of the scene and undesirable textures in the image. A distribution over image segmentations and labelings is computed via a conditional random field; this distribution controls the application of various local image transforms to regions in the image. Parameters governing both the labeling and transforms are jointly optimized w.r.t. a training set of before-and-after example images. One major advantage of our formulation is the ability to marginalize over all possible labelings and thus exploit all the information in the distribution; this yields better results than MAP inference. We demonstrate with a system that is trained to correct red-eye, reduce specularities, and remove acne and other blemishes from faces, showing results with test images scavenged from acne-themed internet message boards.*

## 1. Overview and Background

Some kinds of photographic touch-up, notably red-eye reduction, have been successfully automated. Other kinds, such as hole repair, are amenable to automatic infill but first require manual segmentation of flaws. Most others, such as blemish and specularities removal, require manual detection and repair, because they ultimately rely on some understanding of the scene and *how we would wish it to appear*. Our goal is to capture this relationship via machine learning and fully automate the work of touch-up inside of a camera.

We propose to learn a mapping from a training set of “before” and “after” images, where “after” images contain local image repairs that correct flaws in the scene (*e.g.*, pimples) or image acquisition (*e.g.*, specularities). The mapping revolves around a labeling of texture patches in the image as “healthy skin”, “lips”, “acne”, “background”, etc. More precisely, regions of the source image should be labeled according to the kind of repair they need. This approach is related to nonphotorealistic rendering (NPR) and super-resolution (SR) works where inference about the surfaces

constituting an imaged scene is used to calculate how the image would appear with small changes to lighting, view-point, geometry, materials, etc. (*e.g.*, [2, 5, 10, 7, 11]). In these literatures, the source image is divided into patches and a Markov random field (MRF) generates a joint distribution over the image and possible labelings of the patches, usually on the basis of patch and label statistics observed in a training set. Parameter estimation and inference in MRFs ranges from hard to NP-hard [1, 13, 6], and it is telling that this approach has been more successful in NPR than in SR, or, for that matter, image labeling.

In our view, the chief technical problem is that it is unlikely that the true scene can be estimated precisely and reliably, especially from impromptu images snapped by novice photographers using low-quality cameras in unfavorable lighting. Consequently we cannot expect any inference process to give a single set of reliable repair labels.

Our mathematical framework addresses this problem with three principal contributions: First, our objective seeks a target image that is most likely with respect to *all* possible interpretations of the image, *i.e.*, we seek to *marginalize out the labeling*. Secondly, we introduce a *weak scene grammar* which improves statistical inference and learning (because faces and many other kinds of scenes cannot be assembled arbitrarily). Thirdly, our system learns to *recruit repair texture* from elsewhere in the image.

We formulate our system as a discriminatively trained conditional random field (CRF) [8, 12]. Unlike MRFs, CRFs model the likelihood of the output given the input, not the joint likelihood of the output and the input, thus the representational capacity of the model is not wasted describing patterns in the training data that are irrelevant to the inference task. Due to the difficulty of training CRFs on cyclic graphs, CRFs have only recently gained traction in computer vision [7, 9, 16, 14]. Our objective adds the challenge of integrating over all labelings. In the reduction to practice, we make this mathematical framework tractable by developing lower bounds on the objective that allow fast inference and learning methods.

In contrast to photo-processing systems that use spe-

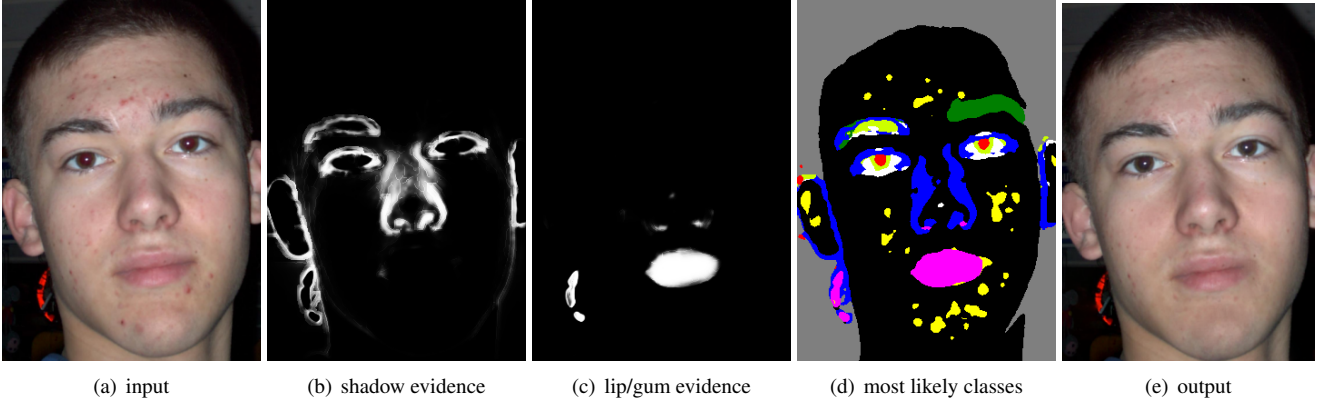


Figure 1. Automatic touch-up applied to an image found on the internet (a). A low-level pixel classifier estimates evidence for 10 classes of image transforms, each specialized for texture classes that roughly correspond to shadowed skin (b), lips/gums (c), eyebrows, sclera, iris, pupils, teeth, background, etc. This evidence is combined with a weak scene grammar and learned statistics of face images in a conditional random field, yielding class likelihoods for each pixel. In this paper, we visually represent the distribution over labellings with a false-color image that indicates, at each pixel site, the label having maximum marginal likelihood (henceforth, maximum-of-marginals). (d). Class likelihoods in turn inform local texture filtering or replacement, here yielding a resynthesized image with red-eye corrected and blemishes removed, but birthmarks preserved (e). The entire process is automatic and trained discriminatively from before-and-after images.

cially engineered heuristics to find and correct each kind of flaw (*e.g.* [3]) our framework uses no shape or geometric information and performs no search in the image. I.e., the facts that pupils are round and symmetrically arranged about the nose, that lips have corners, etc., play no role in our system. These invariants *could* be exploited by providing additional evidence features to the CRF, but this paper explores image editing as a purely pixel-level process.

## 2. Inference Framework

Let source image  $\mathbf{X}$  and target image  $\mathbf{Y}$  be tiled with patches, represented as vectors  $\mathbf{x}_i \in \mathbb{R}^p$  and  $\mathbf{y}_i \in \mathbb{R}^q$  respectively. These may be as small as individual pixels. Each patch can be described with a label  $\ell_i$  drawn from a finite set  $\mathcal{L}$  of repairs. Associated with the tiling is a graph  $\{\mathcal{V}, \mathcal{E}\}$  indicating how labelings of adjacent patches might be constrained by logical or statistical information. A joint labeling  $\boldsymbol{\ell} \in \mathcal{L}^N$ ,  $N = |\mathcal{V}|$  is an assignment of labels to all patches in an image. We define the score of a labeling as

$$s(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta}) = \sum_{i \in \mathcal{V}} e_{\ell_i}^\top \theta_1 f_i(\mathbf{X}) + \sum_{(ij) \in \mathcal{E} | i > j} e_{\ell_i}^\top \theta_2 e_{\ell_j}. \quad (1)$$

Here  $\ell_i$  is the label at the  $i^{\text{th}}$  patch,  $e_{\ell_i}$  is a canonical indicator vector (*e.g.*, if  $\ell_i = 3$ ,  $e_{\ell_i} = [0, 0, 1, 0, \dots, 0]$ ); evidence vector  $f_i(\mathbf{X})$  contains feature weights computed for the  $i^{\text{th}}$  image patch; and  $\theta_1, \theta_2$  are parameter matrices.  $\theta_1$  scores compatibility between a label and evidence;  $\theta_2$  scores compatibility between labels at patches that are connected in the graph.

A conditional random field (CRF) [8, 12] is simply a mapping of the score to a normalized probability distribu-

tion over all possible joint labelings of the graph. In our case,

$$p(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{Z} \exp s(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta}) \quad (2)$$

with partition function  $Z = \sum_{\boldsymbol{\ell}' \in \mathcal{L}^N} \exp s(\boldsymbol{\ell}'|\mathbf{X}, \boldsymbol{\theta})$ . We will use  $\boldsymbol{\theta}$  henceforth to refer to all system parameters.

We first constrain this distribution to penalize “ungrammatical” scene interpretations—*e.g.*, an eyebrow adjoining teeth—by clamping the relevant parameter to  $-\infty$  (in reduction to practice, some constant  $\ll 0$ ). These weak “scene grammar” constraints usefully sharpen the distribution over labelings, but do not force face-structured interpretations.

We now use this distribution over labelings to build a distribution over synthesized target images. We associate each label with a unique “repair”—a unique local image transform that predicts the texture in a target image patch given the corresponding source image patch. In this paper each transform is a (noisy) affine map from local spatial gradients  $d\mathbf{x}_i$  of the source image to those of the target image ( $d\mathbf{y}_i$ ), plus some random additive noise  $\boldsymbol{\nu}_\ell$ . Thus for each patch  $i$  and label  $\ell_i$ ,

$$d\mathbf{y}_i = \mathbf{W}_{\ell_i} [d\mathbf{x}_i^\top, 1]^\top + \boldsymbol{\nu}_\ell, \quad (3)$$

with learned matrix  $\mathbf{W}_\ell$  giving the affine transform associated with the label  $\ell$ . The distribution of  $\boldsymbol{\nu}_\ell$  determines an *prediction* likelihood function  $p(d\mathbf{y}_i | \ell_i, d\mathbf{x}_i, \boldsymbol{\theta})$ , where  $\boldsymbol{\theta}$  refers to all CRF and transformation parameters. We use  $\boldsymbol{\nu}_\ell \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , making the prediction likelihood a gaussian density.

Assuming predicted patch gradients are independent given an input image  $\mathbf{X}$  and a full labeling  $\mathbf{L}(\mathbf{X})$ , and a

gaussian noise model for the the integration of the target image from its gradients, the conditional likelihood of a target image given source image and labeling is

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{L}(\mathbf{X}), \boldsymbol{\theta}) \propto \prod_i p(dy_i|\ell_i, dx_i, \boldsymbol{\theta}). \quad (4)$$

Of course, for most source images there is a good deal of uncertainty about the correct labeling, even to human eyes—otherwise, everyone would be a talented touch-up artist. The optimal approach under such uncertainty is to solve for a target image that has greatest conditional likelihood with respect to the entire distribution of possible image labelings, i.e., we marginalize out the labelings and therefore seek to maximize

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \propto \sum_{\boldsymbol{\ell} \in \mathcal{L}^N} p(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta}) \prod_{i \in \mathcal{V}} p(dy_i|\ell_i, dx_i, \boldsymbol{\theta}). \quad (5)$$

In general this calculation is intractable because the sum runs over an exponential number of labellings. However we can construct a lower bound which allows inference and estimation to be cast in terms of convex optimization problems: We apply the log-sum inequality to r.h.s. eqn. (5) and rearrange the sums to show that the conditional log-likelihood  $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$  is lower-bounded by a weighted sum of the prediction log-likelihoods (errors w.r.t. eqn. 3), where the weights at each patch are precisely the marginal likelihoods of the labels at that patch

$$\begin{aligned} & \log \sum_{\boldsymbol{\ell} \in \mathcal{L}^N} p(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta}) \prod_{i \in \mathcal{V}} p(y_i|\ell_i, \mathbf{x}_i, \boldsymbol{\theta}) \\ & \geq \sum_{\boldsymbol{\ell} \in \mathcal{L}^N} p(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta}) \sum_{i \in \mathcal{V}} \log p(y_i|\ell_i, \mathbf{x}_i, \boldsymbol{\theta}) \\ & = \sum_{i \in \mathcal{V}, j \in \mathcal{L}} \log p(y_i|\ell_i = j, \mathbf{x}_i, \boldsymbol{\theta}) \sum_{\boldsymbol{\ell} \in \mathcal{L}^N | \ell_i = j} p(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta}) \\ & = \sum_{i \in \mathcal{V}, j \in \mathcal{L}} (\log p(y_i|\ell_i = j, \mathbf{x}_i, \boldsymbol{\theta})) \cdot p(\ell_i = j|\mathbf{X}, \boldsymbol{\theta}) \end{aligned} \quad (6)$$

Thus inference and estimation will revolve around computing the label marginal  $p(\ell_i = j|\mathbf{X}, \boldsymbol{\theta})$ , which is the conditional likelihood of a specific label at a specific patch, marginalized over all possible image labelings. For this there are computationally attractive approximations with bounded suboptimality. Note that the bound tightens when the distribution over labelings  $p(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta})$  has low entropy, and reaches equality (in the unlikely scenario) where the probability mass of the CRF distribution is concentrated on a single labeling.

Both inference and parameter estimation can be done effectively by maximizing this lower bound on the likelihood of the target image. Indeed, all the problems we discuss below will be addressed as convex optimizations:

**Inference:** To find the optimal target image  $\mathbf{Y}|\mathbf{X}$ , equation (6) tells us to choose target patch gradients that

have minimal prediction error  $\log p(dy_i|\ell_i = j, dx_i, \boldsymbol{\theta})$  with regard to the marginal likelihoods of the patch labels. With normally distributed  $\boldsymbol{\nu}$ , this can be solved in a straightforward sparse least-squares problem where we seek  $\{\mathbf{y}_i\}_{i=1 \dots N}$  that minimize the squared difference between the two sides of equation (3), summing over all patches and labels, and weighted by the label marginals.

**Estimation:** (transforms) Given paired training images  $\{\mathbf{X}, \mathbf{Y}\}$  and label marginals for  $\mathbf{X}$ , the same system of linear equations can be solved for the optimal affine transforms  $\{\mathbf{W}_\ell\}_{\ell=1 \dots |\mathcal{L}|}$ . This maximizes the lower bound, holding the CRF parameters fixed.

**Estimation:** (CRF parameters) Holding the affine transforms fixed, the lower bound is maximized when all the marginal probability mass is concentrated on the transforms giving the least prediction error. Therefore we can increase the lower bound by training the CRF on the discrete labeling having least prediction error. Formally,  $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \geq p(\mathbf{Y}|\boldsymbol{\ell}_{\max}, \mathbf{X})p(\boldsymbol{\ell}_{\max}|\mathbf{X}, \boldsymbol{\theta})$  where  $\boldsymbol{\ell}_{\max}$  is the labeling giving lowest prediction error of a training pair  $\mathbf{X}, \mathbf{Y}$  according to eqn. 3. We choose  $\boldsymbol{\ell}_{\max}$  to maximize  $p(\mathbf{Y}|\boldsymbol{\ell}_{\max}, \mathbf{X})$ , then optimize  $p(\boldsymbol{\ell}_{\max}|\mathbf{X}, \boldsymbol{\theta})$  through any supervised training method for CRFs, e.g., ascending the gradient

$$\frac{d}{d\theta_2} \log p(\boldsymbol{\ell}_{\max}|\mathbf{X}, \boldsymbol{\theta}) = \{ \langle e_i e_j^\top \rangle_{p(\boldsymbol{\ell}_{\max})=1} - \langle e_i e_j^\top \rangle_{p(\boldsymbol{\ell}|\mathbf{X}, \boldsymbol{\theta})} \}_{ij},$$

and similarly for  $\theta_1$ , replacing  $e_j$  with  $f_j(\mathbf{X})$ . Here  $\langle e_i e_j \rangle_{p(\cdot)}$  is the expectation, under distribution  $p(\cdot)$ , of the outer product between the label indicator vectors at any two patches connected in the graph, i.e., it is the marginal pairwise probabilities of labels  $i$  and  $j$  over all edges in the graph. The gradient is thus the difference between marginal pairwise statistics of the labels in  $\boldsymbol{\ell}_{\max}$  the marginal pairwise probabilities in the distribution specified by  $\boldsymbol{\theta}$ .

Finally, because our bounding arguments are strongest with low-entropy distributions, we use a  $L_1$  prior on the parameters:  $p(\theta_{ij}) \propto e^{-|\theta_{ij}|}$ , which promotes sparsity in the compatibility matrices. In gradient ascent this is effected by shifting all free parameters in  $\theta_1, \theta_2$  a constant decrement toward zero on each iteration.

Training proceeds by alternating optimization of the transformation parameters and the CRF parameters. One can train using only “before” ( $\mathbf{X}$ ) and “after” ( $\mathbf{Y}$ ) images. However, we found that first training the CRF matrices  $\theta_1, \theta_2$  on “label hint” images considerably speeds the subsequent training of the full system.

Training adjusts the label categories discriminatively so that prediction accuracy is increased, so the resulting (label) classes are optimized for image touch-up rather than for tissue identification. In that light, we stress that labeling images in this paper are *not* face segmentations—they merely indicate the most heavily weighted transform (maximum of marginals) at each pixel.

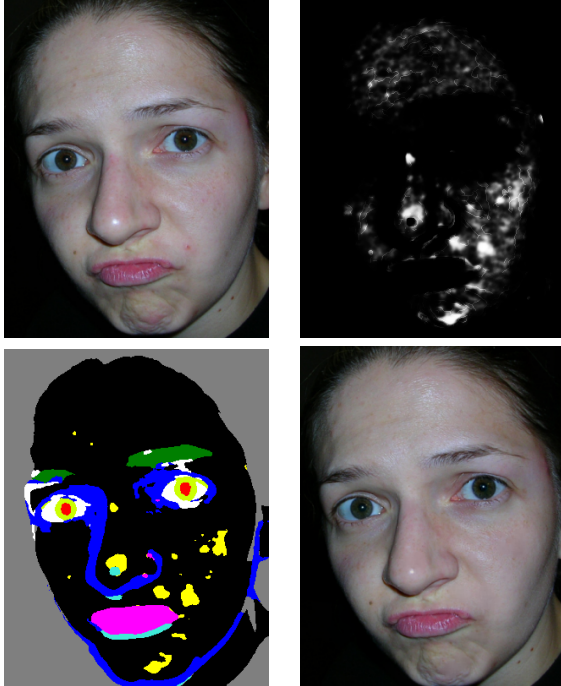


Figure 2. In this test image the blemish-detecting SVM has a strong response over most of the face and nose, including birthmarks and specularities. The CRF determines that most of these pixels are more likely to belong to other classes. In the final result the birthmark is kept; the acne on the forehead, cheek, and chin is removed; and the specularity is moderated.

### 3. Implementation and Results

Face images were collected from user message board at web site <http://acne.org/> with an eye to having a variety of skin types, ages, and complexions. Most are informal self-portraits taken with low resolution phone cams or webcams under generally unfavorable lighting conditions. Poses, lighting, and lens distortions vary widely. From each training image a blemished or unblemished image was made via manual touch-up, as well as a “hint” labeling image. Images were roughly cropped to the head, as can be done by current in-camera face-detectors. To reduce lighting and gamut bias, the colorspace of each image was affinely projected onto PCA axes estimated from pixels in its central region as in [4]. Roughly 25 images were used for training, and 50 for testing. No test individuals appear in the training set.

For convenience of experimentation, we divided the processing pipeline into three parts: a low-level pixel classifier that yields feature vectors for the CRF; a CRF that uses these vectors as evidence; and the least-squares solver that combines the affine transforms and re-integration of the gradient patches to yield the final predicted image. These are first trained separately on the “before”, “after”, and “label hint” images. They are then combined as a favorable pa-

rameter initialization of equation (5), and jointly improved by alternating estimation of  $\theta_1, \theta_2$  and  $\{\mathbf{W}_i\}$  to a point of diminishing returns. In our experiments, 4-8 iterations were needed to obtain good results on the held-out test images.

For the initial low-level classifier, a linear support vector machine (SVM) was estimated to distinguish patches “belonging” to each pair of classes. Each image patch is represented by a vector of oriented multiscale Gabor filter responses, computed in the top two dimensions of the image’s colorspace PCA. Following standard practice for multiclass settings [15], the pairwise SVM scores (distance of a point from the separating hyperplane, measured as multiples of the margin) are mapped via tanh sigmoids to pseudo-probabilities, which are multiplied to yield the class pseudo-likelihoods. These are then used as evidence in the CRF. As shown in the response images in figure 1, these are informative but unreliable class indicators.

The CRF parameters were initialized to the pairwise statistics of the labels in the hint images and the low level class pseudo-likelihoods.

To estimate the affine transforms, the source and the target images were moved into the gradient domain by finite differencing, then all target patches belonging to a label in the hint image were affinely regressed onto the corresponding source patches, yielding class-specific affine transforms. Random nearby patches of source texture with other class labels were also incorporated into the regression, where they were used to the degree that they predict the local target texture better than the local source texture. A parallel process is employed in inference. This allows the system to learn to recruit nearby texture when replacing undesirable such as skin sores with healthier looking skin texture. For classes with insufficient samples, the least-squares regression was lightly regularized with small diagonal values. We experimented with patch sizes  $1 \times 1$ ,  $2 \times 2$ ,  $4 \times 4$ ,  $6 \times 6$ ,  $8 \times 8$ , all giving visually good results, with  $6 \times 6$  yielding lowest prediction errors. It is interesting to note that averaging the label marginals in the  $1 \times 1$  CRF over  $6 \times 6$  results in marginals extremely close to those of the  $6 \times 6$  CRF. For the purposes of illustration, the false-color images in this paper are generated with a  $1 \times 1$  patch size CRF.

Thus initialized, the modules were combined and jointly optimized as described in the previous section.

Inference proceeds as follows: Given a novel source image, low-level class scores are computed from the multiclass SVM and the resulting “evidence images” are used to compute the CRF marginal likelihood of each label at every patch via loopy belief propagation [17]. For each source patch, the software locates nearby “alternative” patches with high probability of being skin, eyebrow, lips, etc. The gradients of all source patches and their alternative patches are then passed to the least-squares solver. The solver simultaneously applies the affine transforms and re-integrates

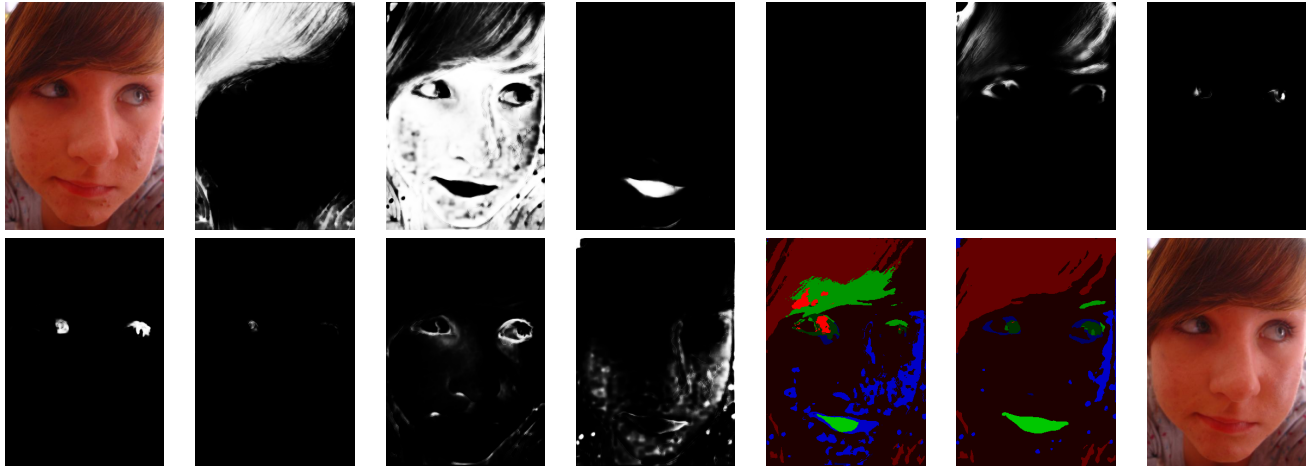


Figure 3. Contextual inference improves over special-purpose classifiers. From left to right, an image with very different lighting than those of the training set; the class-specific response images of the multiclass SVM. The last three images show the multiclass SVM pixel classification, the more accurate CRF MAP labeling, and the system output, which uses the full marginal distribution on labels and thus avoids problems with the SVD and CRF MAP labelings.

the resulting gradients to produce the target image.

The inferential parts of this pipeline are illustrated in figure 1. Figure 2 highlights the role of the CRF in assessing poor evidence. Note that the system correctly distinguishes pimples from freckles of the same color, removing the former and preserving the latter. Figure 3 shows that the CRF produces significantly better labelings than the low-level classifier, but even when both produce flawed labelings, marginalizing over the CRF’s full distribution yields a good output image. Figure 6 shows several real-world images and their automatic touch-ups. All are test images, kept out of the training set.

Often, but not always, the MAP and maximum-of-marginals image agree (see figure 5), and give a reasonable segmentation. However, using such “hard” labelings for touch-up generally does not produce results as good as those from the “soft” marginal labeling. Finally, figure 4 shows an example of a bloop— a dark nostril mistaken for a pupil—that was eliminated by the introduction of weak grammatical constraints.

In our experiments, training took less than two hours—relatively quick considering that the model contains a multiclass CRF. Inference took roughly 1/2 minute per photo.

#### 4. Discussion

In this paper we constructed a conditional random field on local signal edits by conditioning a set of parameterized continuous local transforms on a multiclass discrete CRF, then marginalizing out all “hidden” variables by integrating over all CRF labelings. Marginalization allows us to use all the information in the CRF distribution. Because the full marginalization is intractable, we developed a lower bound that revolves around node marginals, and in-



Figure 4. A labeling “bloop” (left) where the dark nostril is mistaken for a pupil and dark skin around the eyes is mistaken for eyebrows. A weak scene grammar largely eliminates such mistakes (right).

ference/estimation procedures that reduce to a small number of convex optimizations in practice. With as little as 10 training images, the system does a good job of removing unwanted blemishes and imaging artifacts from photographs.

This framework is fairly general and could be applied to sampled continuous signals of any kind and dimension, using any family of parameterized transforms. If the log prediction error is differentiable, the parameters governing all stages of the inference process can be jointly optimized.

Experiments with other lower bounds based on MAP and maximum-of-marginal labelings produced inferior results, partly because these approximations are less informative, and partly because such “hard” labelings produced undesirable artifacts in image patches that straddle more than one tissue type. Nonetheless, fast inference algorithms make these approximations attractive and we hope to make them



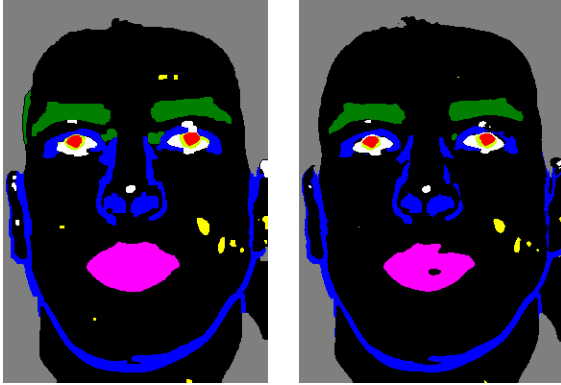


Figure 5. A MAP labeling (left) versus a maximum-of-marginals labeling. The MAP label is slightly smoother, but the two are almost identical, showing that at the end of training the CRF parameters and the input induce a low-entropy distribution over scene interpretations.

viable by increasing the training set and classes.

Our current implementation could be improved in a number of ways: One might obtain even better label likelihoods by training the CRF directly on image features, instead of multiclass SVM outputs (at cost of considerably more computation). The source-to-target transforms could be constrained to ignore pixels of foreign source texture (*e.g.*, a hair over the forehead). Additional classes could be introduced, for example, to allow different treatment of eyelashes and eyebrows. A stronger grammar could impose labeling restrictions to enforce the geometric schema of faces. The system also makes subtle errors, for example, smoothing out the corners of the lips because there is nonzero probability mass for the hypothesis that these are pimples. Resolving this will require a larger repertoire of evidence features, particularly those that implicitly carry some information about whether the observed texture is appropriately located in the geometry of the face. However, these are mainly engineering matters, whereas our goal for this paper is to exhibit a simple and minimal system that performs quite well “in the wild” of casual imaging devices and uncontrolled imaging conditions.

## References

- [1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.
- [2] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.
- [3] M. Gaubatz and R. Ulichney. Automatic red-eye detection and correction. *Image Processing*, 2002.
- [4] D. J. Heeger and J. R. Bergen. Pyramid-based texture analysis/synthesis. In *Proc. SIGGRAPH*, pages 229–238, 1995.
- [5] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. Salesin. Image analogies. In *Proc. SIGGRAPH*, pages 327–340, 2001.
- [6] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. PAMI*, 26(2):147–159, 2004.
- [7] S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *Proc. NIPS*, 2003.
- [8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.
- [9] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Proc. NIPS*, 2004.
- [10] R. Rosales, K. Achan, and B. J. Frey. Unsupervised image translation. In *Proc. ICCV*, pages 472–478, 2003.
- [11] T. A. Stephenson and T. Chen. Adaptive markov random fields for example-based super-resolution of faces. *EURASIP Journal on Applied Signal Processing*, 2006.
- [12] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2006.
- [13] M. Wainwright, T. Jaakkola, and A. Willsky. Map estimation via agreement on trees: Message-passing and linear-programming approaches. *IEEE Trans. Information Theory*, 51(11):3697–3717, 2005.
- [14] J. Winn and J. Shotton. The layout consistent random field. In *Proc. CVPR*, 2006.
- [15] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.
- [16] H. Xuming, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *Proc. CVPR*, pages 695–702, 2004.
- [17] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Proc. NIPS*, pages 689–695, 2000.





Figure 6. Sample outputs. At top, left to right: Original test images; the most likely label at each pixel; resynthesized images after automatic touch-up, and the test image with some of the blemishes circled. On some displays and printers the color gamut may obscure the more subtle blemishes; readers may wish to view the images with magnification.