Integrating Local Classifiers through Nonlinear Dynamics on Label Graphs with an Application to Image Segmentation

Yutian Chen Andrew Gelfand Charless C. Fowlkes Max Welling
Bren School of Information and Computer Sciences
University of California, Irvine, CA 92697, USA
{yutianc,agelfand,fowlkes,welling}@ics.uci.edu

Abstract

We present a new method to combine possibly inconsistent locally (piecewise) trained conditional models $p(\mathbf{y}_\alpha|\mathbf{x}_\alpha)$ into pseudo-samples from a global model. Our method does not require training of a CRF, but instead generates samples by iterating forward a weakly chaotic dynamical system. The new method is illustrated on image segmentation tasks where classifiers based on local appearance cues are combined with pairwise boundary cues.

1. Introduction

The use of Markov Random Fields (MRF) and their discriminatively trained cousin Conditional Random Fields (CRF) is ubiquitous in computer vision. For instance, CRFs have been used extensively in image segmentation and labeling [2, 9, 7], as well as object recognition [12], image denoising and image restoration [13], and stereo [11].

The standard approach for learning such models is to first propose local potential functions $\psi_\alpha(\mathbf{y}_\alpha, \mathbf{x}_\alpha, \mathbf{w}_\alpha)$ on local overlapping subsets of variables $\mathbf{y}_\alpha$, conditioned on some input variables $\mathbf{x}_\alpha$ and parameterized by parameters $\mathbf{w}_\alpha$. The exponentiated sum of all these potentials then compiles into an undirected graphical model, usually with a very high treewidth. This high treewidth renders inference in such models computationally very expensive. Unfortunately, inference is a necessary ingredient in tuning the parameters $\mathbf{w}_\alpha$ from data and applying the model to new test data.

The question naturally arises if there is an alternative procedure by which we can still exploit the graphical model structure but avoid the computationally costly learning procedure. Our first observation is that we often have powerful local classifiers that we would like to directly embed in our joint model. For instance, in image segmentation one can train a discriminative classifier from local appearance cues: $p(\mathbf{y}_\alpha|\mathbf{x}_\alpha)$ where we can think of α as an indexing a super-pixel (if the classifier does not directly provide a probabilistic output we can easily calibrate its score into a probability afterwards.) Moreover, we can also obtain information about pairs of super-pixels by detecting boundaries and representing this in a probability distribution $p(\mathbf{y}_\beta|\mathbf{x}_\beta)$ where β now represents a pair of super-pixels. Figure 1 illustrates the power of combining such local classifiers into a global estimate. The question is, how can we combine these conditionally trained local probability models into a global model?

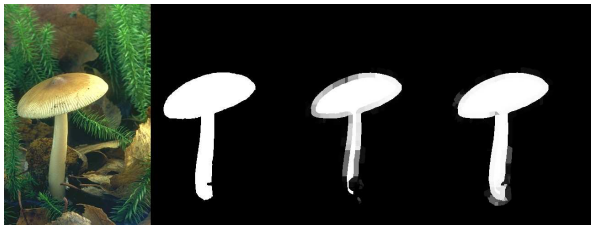

Figure 1. Combining an independently trained local color model with boundary model for segmentation. Images from left to right are the actual image, the ground truth segmentation, the probability of foreground using local color information and the result of combining local color and boundary cues using herding.

The approach we take is to enforce that a joint model must respect these local distributions as marginals as much as possible. Because the local models were trained independently, this may not always be possible. In technical terms: the set of marginal probability distributions lies outside the marginal polytope. In this case, we would like to find the joint distribution that has marginals inside the marginal polytope that are as close as possible (in some sense to be defined) to the input marginals. We will prove that the proposed procedure orthogonally projects the inconsistent marginals onto the marginal polytope.

Instead of returning a joint model parameterized by a single estimate $\mathbf{w}^\star_\alpha$, we propose an algorithm - called herding - that generates a sequence of parameters $\mathbf{w}_\alpha$'s and labels $\mathbf{y}_\alpha$'s [15]. This sequence is produced by running a deterministic nonlinear dynamical system jointly in parameter

space and label space. Input to this dynamical system are the local marginals (or more generally average feature values). The output label sequence can then be used to segment the image. If the input marginals are consistent and some easy to check conditions are satisfied, then one can show that the generated label sequences actually respect the input marginals: $|p(\mathbf{y}_\alpha|\mathbf{x}_\alpha) - \hat{p}_T(\mathbf{y}_\alpha|\mathbf{x}_\alpha)| \sim \mathcal{O}(\frac{1}{T})$ where $\hat{p}_T$ is the empirical marginal computed over a label sequence of length T. (Note that convergence is faster than the usual $\mathcal{O}(\sqrt{1/T})$ rate for stochastic averages.) Perhaps the most surprising result is that this can often be achieved without solving NP-hard inference problems. Instead, one only needs repeated local maximizations on the underlying graphical model (but with different parameters) until a simple local condition related to the "perceptron cycling theorem" is met [6]. For inconsistent input marginals this no longer holds but we can prove that the label sequence corresponds to a set of consistent marginals which are closest (in the L_2 sense) to the input marginals. In this case suboptimal maximization may lead to a suboptimal projections, providing a natural tradeoff between accuracy and computation.

We illustrate these ideas for the task of image segmentation where we integrate local appearance and boundary cues into a joint graphical model. We show that our algorithm is competitive with the traditional approach of learning the CRF directly. We predict that our approach is much more broadly applicable and could be useful for the computer vision community. The reason for our optimism is that 1) it readily integrates locally trained conditional models, 2) it is very easy to implement and 3) approximations to global maximization may still lead to good performance.

2. Conditional Random Fields

A conditional random field is a standard approach to combining local features into a global conditional distribution over labels. One first defines potential functions $\psi_\alpha(\mathbf{y}_\alpha, \mathbf{x}_\alpha)$ where $\mathbf{y}_\alpha$ is a subset of the labels associated with potential ψ_α and $\mathbf{x}_\alpha$ are input features. The label subsets indexed by α are assumed to overlap and form a loopy graph over the labels $\mathbf{y}$. For instance, the subsets could correspond to all pixels and all pairs of pixels in an image. These local potentials are combined in a CRF which specifies a joint probability distribution over all labels by

$$P_{\text{crf-1}}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} \exp\left[\sum_\alpha w_\alpha \psi_\alpha(\mathbf{y}_\alpha, \mathbf{x}_\alpha)\right] \qquad (1)$$

$\{\mathbf{w}_\alpha\}$ are model parameters, one associated with each potential function, that are typically learned from data. Maximum likelihood training of such a model with respect to some dataset with empirical distribution $\hat{P}(y|x)$ has the intuitive property that expectations of the potential functions under the model match those of the training data

$$\mathbb{E}_{P_{\text{crf-1}}}[\psi_\alpha] = \mathbb{E}_{\hat{P}}[\psi_\alpha] \qquad (2)$$

While this is quite elegant, it poses a practical problem that each potential function ψ_α has a distinct parameter w_α to be estimated. In typical image labeling problems, there may be thousands of weights to learn (e.g. one for every pixel and pair of pixels for every image). In other words, there is less than one pixel of information per parameter, leading to extreme overfitting.

To avoid this explosion of parameters, a typical approach is to share parameters across potential functions. For instance, if we have pixel-wise and pairwise potential functions we could use a single parameter λ to trade off their relative importance.

One of the main questions we wish to address in this paper is how to most effectively use the information of local discriminative classifiers $p(\mathbf{y}_\alpha|\mathbf{x}_\alpha)$ whose parameters are trained on all the pixels of an image or a training set of images. In the CRF approach one can incorporate these local classifiers by taking the log probabilities as local potential functions so that $\psi_\alpha(\mathbf{y}_\alpha) = -\log(p(\mathbf{y}_\alpha|\mathbf{x}_\alpha))$. For example, [5] use the following CRF model for segmentation

$$E_{P_{\text{crf-2}}}(\mathbf{y}|\mathbf{x}) = -\sum_i \log(p(y_i|x_i)) - \lambda \sum_{i,j} \log(p(y_i \neq y_j|x_i, x_j)) \qquad (3)$$

where local unary models $p_i(y_i|x_i)$ and pairwise models $p_{ij}(y_i \neq y_j|x_i, x_j)$ are trained independently and then a single weight λ is fit that calibrates the relative importance of the unary and pairwise terms.

Unfortunately, in such a locally trained model, there is no longer reason to expect that the model matches the training data in the previous sense that $\mathbb{E}_{P_{\text{crf-2}}}[\psi_\alpha] = \mathbb{E}_{\hat{P}}[\psi_\alpha]$ since we have only one parameter to tune and a very large collections of these moment constraints (nr. of pixels plus nr. of neighboring pairs of pixels).

Instead, we might like to impose that our global model at least still approximately matches the constraints,

$$\mathbb{E}_P[\psi_\alpha] = \mathbb{E}_{p_\alpha}[\psi_\alpha] \qquad (4)$$

where $\hat{P}$ has been replaced with p_α. For features of the form $\psi_{\alpha,\mathbf{z}_\alpha}(\mathbf{y}_\alpha) = \mathbb{I}[\mathbf{y}_\alpha = \mathbf{z}_\alpha]$, this condition implies that the joint distribution marginalizes down to local distributions

$$\sum_{\mathbf{y}\setminus\mathbf{y}_\alpha} P_{\text{crf-2}}(\mathbf{y}|\mathbf{x}) = p_\alpha(\mathbf{y}_\alpha|\mathbf{x}_\alpha) \qquad (5)$$

However, with independently trained local classifiers, no joint model can achieve this as the p_α are likely to be mutually *inconsistent*.

The Herding approach which we describe in the next section provides an elegant solution to this problem. Given locally trained discriminative models, it produces a sequence of states $\ldots \mathbf{y}^t, \mathbf{y}^{t+1} \ldots$ that on average satisfy the marginalization condition (Eqn. 4) when the local models are consistent. If the locally trained models are not consistent, we show that the same procedure still produces a sequence whose average behavior matches that of the closest consistent model. We thus gain some of the flexibility of the general CRF formulation (Eqn. 1) in matching moments while retaining the parsimony of piecewise training local discriminative models.

2.1. Herding Local Models

The herding approach we advocate follows the second method in that we try to identify a joint probability distribution over some features ψ_α that approximately marginalizes to averages $\mathbb{E}_{p_\alpha}[\psi_\alpha]$. Let us first assume that the p_α were in fact consistent and let us define the following dynamical system,

$$\mathbf{y}^t = \arg\max_{\mathbf{y}} \sum_\alpha w_\alpha^{t-1} \psi_\alpha(\mathbf{y}_\alpha, \mathbf{x}_\alpha) \tag{6}$$

$$w_\alpha^t = w_\alpha^{t-1} + \eta_\alpha \left(\mathbb{E}_{p_\alpha}[\psi_\alpha] - \psi_\alpha(\mathbf{y}_\alpha^t, \mathbf{x}_\alpha) \right) \tag{7}$$

In the experiments we use $\psi_{\alpha,\mathbf{z}_\alpha} = \mathbb{I}[\mathbf{y}_\alpha = \mathbf{z}_\alpha]$, i.e. a separate feature for every state $\mathbf{z}_\alpha$ in every region α.

It can now be shown [6] that if at every iteration we can guarantee that the following condition holds,

$$C^t = \sum_\alpha w_\alpha^{t-1} \left(\mathbb{E}_{p_\alpha}[\psi_\alpha] - \psi_\alpha(\mathbf{y}_\alpha^t, \mathbf{x}_\alpha)) \right) \leq 0 \tag{8}$$

then it follows that

$$|\mathbb{E}_{p_\alpha}[\psi_\alpha] - \frac{1}{T} \sum_{t=1}^{T} \psi_\alpha(\mathbf{y}_\alpha^t, \mathbf{x}_\alpha)| = \mathcal{O}(\frac{1}{T}) \tag{9}$$

We note that this convergence rate is optimally fast given that we approximate the probabilities with Monte Carlo averages, and in particular much faster than the typical $\mathcal{O}(\sqrt{1/T})$ convergence for stochastically generated averages.

In summary, this deterministic dynamical system generates sequences $...(\mathbf{w}_t, \mathbf{y}_t), (\mathbf{w}_{t+1}, \mathbf{y}_{t+1}), ...$ of parameters and states in such a way that the states come from some joint distribution $P(\mathbf{y}|\mathbf{x})$ which has moments $\mathbb{E}_{p_\alpha}[\psi_\alpha]$. Unlike CRF models the entropy of this joint model is not expected to be maximal, although empirically it is often close. Perhaps surprisingly, for many problems local maximizations initialized at the last iteration are often sufficient to satisfy condition 8 at every iteration so that hard inference is sidestepped. It should be noted that this is not always the case, in particular when the constraints are hard or impossible to satisfy as may arise for image segmentation.

We also emphasize that the dynamical system defined through equations 6 and 7 do not return a parameterized model. The sequence $...(\mathbf{w}_t, \mathbf{y}_t), (\mathbf{w}_{t+1}, \mathbf{y}_{t+1}), ...$ never converges to a fixed point and one should rather think of this as a deterministic process to generate "representative points". In fact, it can be shown that the dynamical system is weakly chaotic meaning that the sequence over $\mathbf{y}_t$ is not periodic but that there is also not extreme sensitivity to initial conditions.

Not having an explicit model is not a problem for the applications we have in mind. For instance, in image segmentation the dynamical system will generate a sequence of segmentations of the input image. From this sequence we can extract the final segmentation by averaging.

2.2. Herding with Inconsistent Marginals

We now describe how to handle inconsistent marginals in herding. When $\mathbb{E}_{p_\alpha}[\psi_\alpha]$ doesn't reside inside the marginal polytope $\mathcal{M}$, then by definition there doesn't exist a joint distribution $P(\mathbf{y}|\mathbf{x})$ with moments $\mathbb{E}_{p_\alpha}[\psi_\alpha]$. If we want to train a CRF without regularization, the parameters will diverge. For herding this means that the condition in Equ. 8 cannot always be satisfied, and the norm of parameters $\mathbf{w}_t$ will also linearly diverge. Nevertheless, we can still obtain a stationary joint distribution of states $\mathbf{y}^t$ from the herding sequence. The potential numerical problems caused by the divergence of $\mathbf{w}_t$ can be easily prevented by taking an additional normalization step $\mathbf{w} \leftarrow \mathbf{w}/K$ for some K. This global scaling will not affect the state sequence $\mathbf{y}^t$ in any way. The most important consequence of inconsistent marginals is that the moments of the joint distribution don't converge to $\mathbb{E}_{p_\alpha}[\psi_\alpha]$ any more. Instead, we prove in this paper that the moments orthogonally project onto the marginal polytope.

In the following we will denote the collection of expectations $\mathbb{E}_{p_\alpha}[\psi_\alpha],\ \forall \alpha$ as $\bar{\boldsymbol{\psi}}$ and the sample average of the features generated by herding up to time T as $\tilde{\boldsymbol{\psi}}^T = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\psi}(\mathbf{x}_t)$. We now claim that the following property holds:

Proposition 1. *Assume $\bar{\boldsymbol{\psi}}$ is outside the marginal polytope $\mathcal{M}$ and the stepsize η_α is constant. Let $\bar{\boldsymbol{\psi}}_{\mathcal{M}}$ be the L_2 projection of $\bar{\boldsymbol{\psi}}$ on $\mathcal{M}$. Then the average features of herding $\tilde{\boldsymbol{\psi}}^T$ converge to $\bar{\boldsymbol{\psi}}_{\mathcal{M}}$ at the rate of $1/T$.*

For a proof see appendix A.

When η_α depends on the feature index α, we can construct an equivalent herding sequence with a constant stepsize and new features $\{\sqrt{\eta_\alpha}\psi_\alpha\}$. Then proposition 1 still applies except that the L_2 distance is weighted by $\sqrt{\eta_\alpha}$. So the stepsizes control the relative importance of features. When we consider $\boldsymbol{\psi}$ as $\psi_{\alpha,\mathbf{z}_\alpha}(\mathbf{y}_\alpha) = \mathbb{I}[\mathbf{y}_\alpha = \mathbf{z}_\alpha]$, then the marginal probabilities of herding samples will converge to the closest consistent marginals in $\mathcal{M}$.

As an immediate consequence of proposition 1, herding always improves an initial set of moments $\bar{\psi}$ (which drive herding dynamics through equations 6 and 7) in the following sense:

Corollary 2. *For any true expectations $\bar{\psi}_{true}$ and any approximate values $\bar{\psi}$, the limit of the empirical average of the herding sequence won't increase the L_2 error. Specifically, $\|\bar{\psi}_{\mathcal{M}} - \bar{\psi}_{true}\|_2 < \|\bar{\psi} - \bar{\psi}_{true}\|_2$ when $\bar{\psi} \notin \mathcal{M}$, and $\tilde{\psi}^T \to \bar{\psi}$ otherwise.*

3. Application: Image Segmentation

Inconsistent marginals are common in image segmentation. Conditional probabilities of groups of variables can come from various sources such as color, texture, context, etc. We consider an example of two types of possibly inconsistent probabilities in this paper: unary conditional probabilities $\{p_i(y_i|x_i)\}$ on super-pixels and pairwise conditional probabilities $\{p_{ij}(y_i \neq y_j|x_i, x_j)\}$ on neighboring super-pixels. The former provides local class distributions, and the latter suggests the existence of boundaries.

The CRF approach uses the energy defined in Eqn. 3 with a single parameter λ. The best assignment with lowest energy is inferred as the segmentation output, and the best value of λ is estimated on a validation set using grid search.

Our herding algorithm follows equations 6 and 7 with two types of features: $\mathbb{I}(y_i = c)$ and $\mathbb{I}(y_i \neq y_j)$. The step size η_α is scale free in the sense that multiplying all η_α by the same factor doesn't change the output of label sequence $\mathbf{y}^t$, and so without loss of generality we may set the step-sizes for unary features to 1, and those for pairwise features as λ. The value of λ is used to trade off the strength of these two sources of information. The final segmentation for herding is obtained by maximization and averaging,

$$y_i^* = \arg\max_c \sum_t \mathbb{I}[y_i^t = c] \ \ \forall i \tag{10}$$

Notice that the role of the parameter λ is different between the CRF and herding approaches. In the CRF, λ controls the strength of smoothness. Increasing λ always increases smoothness. However, herding tries the respect all the probabilities, and λ measures how our much attention we pay to each of these two sources of information. Increasing λ not only increases the smoothness where $p_{ij}(x_i \neq x_j)$ is small, but also forces an edge where $p_{ij}(x_i \neq x_j)$ is large. As a special case, for a system of N super-pixels with $\lambda \gg 1$ and $p_{ij}(x_i \neq x_j) = 0$ for all neighbors in the label graph, the pairwise term dominates the system, and all super-pixels will take on the same value.

4. Experiments

We apply herding to image segmentation on two datasets, PASCAL VOC 2007 segmentation competition and GrabCut to illustrate its effectiveness. It is compared to the multiclass classifier with local appearance cues only, and to the traditional CRF approach.

4.1. PASCAL VOC 2007

On the PASCAL VOC 2007 dataset, we follow a similar experiment setting as that of [5] and perform segmentation on the level of super-pixels. Each image is first over-segmented by the global probability of boundary (gPb) method [1]. The threshold is set to 0 to make sure most boundaries are retained. SIFT features are then extracted and quantized in order to build a visual dictionary. A local multiclass SVM is trained to provide unary marginals $p_i(y_i|x_i)$ using histograms of the visual words in each super-pixel and its neighbors at a distance at most N. The larger N is, the more context information is available for the local classifier, less noise in the feature histogram but also the more blurred the boundaries between super-pixels become. By increasing N, the segmentations of the local classifier changes from inaccurate and noisy but with clear sharp boundaries to more accurate and smooth but with blurred boundaries (see the results of the local method of $N = 0$ in figure 4 and $N = 3$ in figure 2). The gPb algorithm provides the probability of a boundary between two super-pixels, i.e. the pairwise marginals $p_{ij}(y_i \neq y_j|\mathbf{x})$. The VOC test set includes 210 images, and the "trainval" set is split randomly into a training set of 322 images and a validation set of 100 images. The local classifier is trained on the training set, and the (hyper-)parameters of the CRF and herding are estimated on the validation set.

For the local models, we predict the super-pixel labels based on the output of SVMs. For the CRF models, the MAP label is inferred using the graphcut algorithm from [3, 5, 4, 8] with an energy as in equations 3. The parameter λ is estimated by grid search on the validation set. For the herding method, the maximization step in Eqn. 6 is also executed using the graphcut. Because the original gPb score is trained on the BSDS dataset and a lot of boundaries belonging to irrelevant categories of objects in the VOC dataset are not considered, gPb should be calibrated first. The calibrated pairwise probability is computed as $P_{VOC}(y_i \neq y_j|\mathbf{x}) = P_{BSDS}(y_i \neq y_j|\mathbf{x})^\alpha$, where α controls how sparse the boundaries in the VOC dataset are. The parameters λ and α are estimated on the validation set by first fixing $\alpha = 1$, estimating λ by grid search and then fixing λ and estimating α. More iterations can be done for better performance. Notice that for CRF, the function of λ and α appears in the same position in the pairwise term $\lambda\alpha \log(P(y_i \neq y_j|x_i, x_j))\mathbb{I}(y_i \neq y_j)$, and a second parameter is therefore redundant.

Figure 2 shows some examples of the test images, results of different algorithms as well as their posterior probabilities. The local classifiers are trained on features from

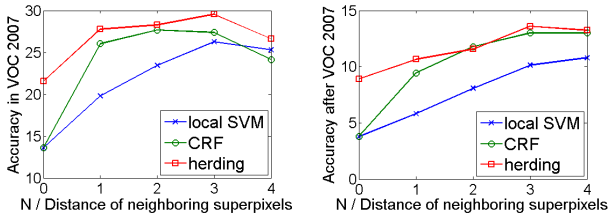

Figure 3. Average accuracy of segmentations by the local SVM classifier (cross), CRF (circle) and herding (square) with different number of neighboring superpixels used for extracting local features. N denotes the maximal distance of the neighboring superpixel used. The left plot uses the 2007 segmentation benchmark criteria (average recall). The plot on the right uses the 2010 criteria on the 2007 dataset (average overlap).

a neighborhood of $N = 3$. So the unary class distribution is already smoothed to some extent (compared to figure 4 for the case of N=0). But herding still leads to better smoothness and locates the boundaries more accurately. Most boundaries occur in the place with strong pairwise probabilities. CRF provides similar benefits as herding for regularizing the local classifiers.

We evaluate the performance of these three models by two measurements. The first one is the average accuracy adopted by VOC 2007 Competition. It measures the average recall of pixels for each category. The second measurement is the one adopted by VOC competition after 2007. It measures the average of the intersection over union ratio for each category. The results of both evaluation methods are shown in figure 3. The results show that both herding and CRF increase the accuracy in most cases, and herding always achieves the best accuracy except for $N = 2$ by the second measurement. The reduction of the advantage of herding compared to CRF in the second measurement may be due to the fact that false positive detections appear frequently in the background which doesn't reduce the recall of the background category by much, but will reduce the intersection over union ratio of the detected category.

Remarkably, herding performs much better than the local method when $N = 0$. The accuracy is improved from 14% to 22% on the first measurement and 4% to 9% on the second measurement, while CRF doesn't help at all. The local classifier performs poorly because the histogram feature is computed from very few pixels as discussed in [5]. Thus regularization on the pairwise term should improve the prediction. It turns out that the optimal value of λ for herding is about 1.1×10^3 which means the importance of the pairwise feature is $\sqrt{\lambda} \approx 33$ times higher than the unary feature, matching our expectation. On the other hand, the best value for CRF is only about 1.1. The difference in the choice of λ leads to the significant difference in the segmentations as shown with a typical example in figure 4. Herding outputs a highly smoothed result with clear boundary while CRF doesn't noticeably change the decision of the local classifier.

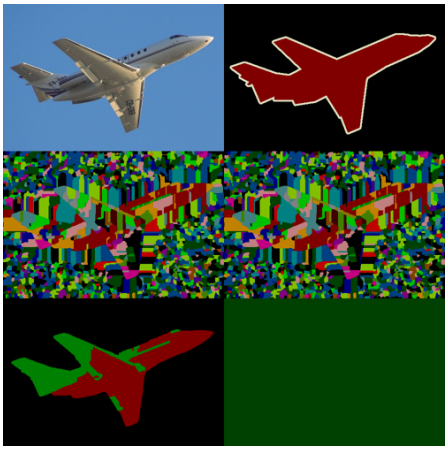

Figure 4. A typical example of segmentations when $N = 0$. The top 2 images are the original image and the ground truth segmentation. The remaining 4 images are respectively the segmentation of the local model, CRF, herding and a CRF with linear potential functions. The local model is so noisy because the histogram of SIFT features is computed from very few pixels.

Two properties of herding previously stated in section 3 would help explain the distinct choices of λ. Firstly, with a large λ, herding tries to average the distribution of superpixels in a smooth area. Although the local SVMs give very noisy results, the average distributions still contain strong signals about the true category. In contrast, the CRF computes the product of distributions which makes the noise in the final distribution even worse. So CRF has to choose a small λ. To verify this hypothesis, we train a CRF with energy as a linear function of features $P(y_i|x_i)$ and $P(y_i \neq y_j|x_i, x_j)$, that also computes the average of distributions when λ is large. The new CRF chooses a large λ (≈ 22) as expected and the accuracy is improved to 16% and 7% respectively. However figure 4 shows that the result is oversmoothed because of the high penalty of boundaries. Secondly, herding not only increases smoothness in flat areas but also encourages boundaries at strong edges. That's why herding still captures the shape of the object correctly even with a large value of λ.

4.2. GrabCut

We also ran herding on the GrabCut data set, which consists of 50 images of a foreground object on a natural background [1]. The objective on the GrabCut images is to perform hard foreground/background segmentation. We divided the image set into 30 training and 20 test images.

The GrabCut data set contains two labeled trimaps for each image, where a trimap is a labeling of pixels as foreground, background or undetermined (see Figure 5). The 'lasso' trimap contains a rough segmentation of the image

[1] http://www.research.microsoft.com/vision/cambridge/segmentation/

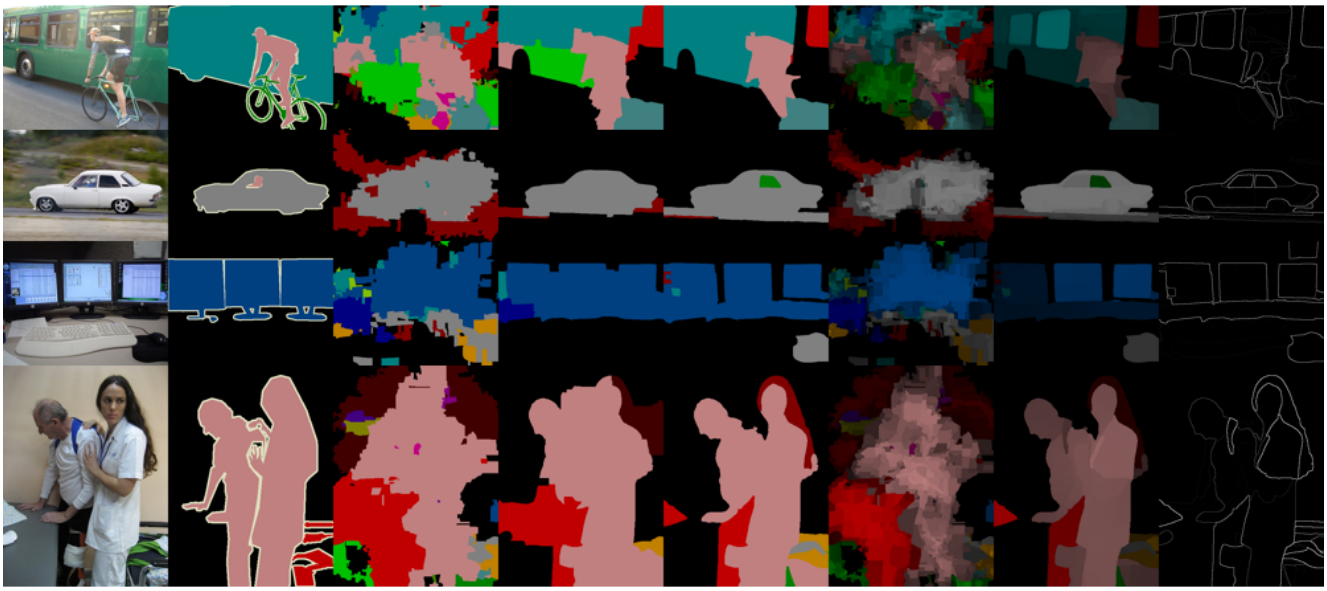

Figure 2. Examples of segmentation on Pascal VOC 2007 data set. Images on each line starting from left to right are respectively: the original image, ground truth segmentation, results of local classifier, CRF and herding, results with intensity proportional to the posterior probability of the local classifier and herding, and the herding estimate of the pairwise probability of the existence of a boundary (the corresponding posterior probability for CRF cannot be easily obtained). Neighboring superpixels of a distance up to 3 hops are used for training local SVM. Best viewed in color.

		background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	Average
Recall	Local	46	7	15	**10**	8	**10**	31	51	34	**17**	6	**16**	41	23	58	50	18	21	**15**	36	**39**	26
	CRF	56	**20**	12	2	**15**	7	33	52	**59**	8	**10**	8	31	20	68	**55**	12	15	**15**	**49**	36	28
	Herd	**62**	3	**16**	3	7	7	**38**	**58**	50	15	2	11	**58**	**24**	**70**	54	**20**	**23**	14	47	**39**	**30**
Overlap	Local	50	**2**	6	**8**	2	0	13	21	14	2	**2**	4	8	10	24	20	6	**8**	**5**	12	10	11
	CRF	**65**	1	**8**	0	2	0	15	**30**	**17**	3	0	4	5	10	**37**	**24**	**9**	**8**	**5**	**18**	**13**	13
	Herd	60	**2**	4	4	**3**	**5**	**23**	28	15	**4**	0	**5**	**20**	**12**	31	22	6	**8**	3	**18**	12	**14**

Table 1. Accuracies per category and the average accuracy of PASCAL VOC 2007 dataset. Each model uses the N value that maximizes the average test accuracy. Top table shows recall (PASCAL 2007 benchmark) the bottom table shows overlap (PASCAL 2010 benchmark)

obtained using a lasso or pen tool. The 'expert' trimap is a ground truth labeling obtained by tracing the boundary of the foreground image at the pixel level. Performance on the GrabCut data set is assessed via the segmentation error rate (SER), which is the number of misclassified pixels in the set of undetermined pixels (where the set of undetermined pixels does not include the undetermined pixels in the expert trimap).

As with the PASCAL data set we train a binary classifier to provide unary marginals and use the gPb method to provide pairwise marginals. Since we have a trimap for each image and can easily identify pixels that are foreground and background, we train a different binary (foreground/background) classifier for each image [2]. In particular, we consider two color models. Color Model (CM) 1 uses RGB values as features; CM 2 uses 64-dimensional feature vectors constructed by a building a histogram of RGB values in the 5-by-5 neighborhood around each pixel. We then oversegment each image using the gPb method and identify pairwise marginals between superpixels i and j ($P(y_i \neq y_j|\mathbf{x})$) as in the PASCAL experiments section. The unary marginals $P(y_i = 0|\mathbf{x})$ and $P(y_i = 1|\mathbf{x})$ for each superpixel are found by averaging the predictions of CM 1 (or CM 2) on each pixel comprising superpixel i. An example of this process is shown in Figure 5.

We compare three different methods on the GrabCut data: 1) Local model; 2) CRF model; and 3) Herding. In the local model, each superpixel is labeled foreground if $P(y_i = 1|\mathbf{x}) > P(y_i = 0|\mathbf{x})$; and background otherwise. For the CRF model and herding, we construct a pairwise MRF that considers only superpixels containing undeter-

[2] As with [2], the color model is trained using background pixels that border the undetermined region.

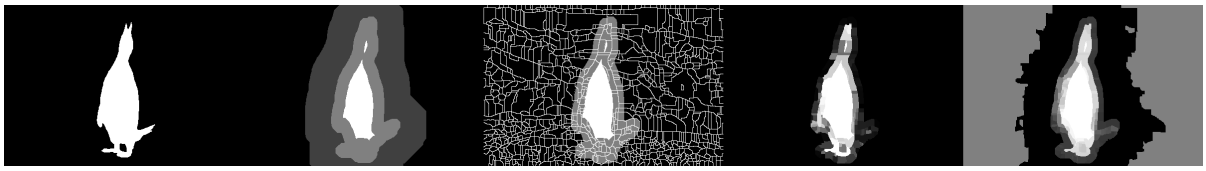

Figure 5. Example of segmentation on GrabCut data set. Images from left to right are the 'expert' ground truth trimap, 'lasso' trimap, over segmentation into superpixels, unary marginal probabilities from local classifiers and unary marginal probabilities after herding. In the trimaps, black pixels are background, white pixels are foreground and light gray pixels are undetermined. The dark gray pixels in the 'lasso' trimap were used as background pixels in training CM 1 and 2. In the last two images, whiter values indicate larger $P(y_i = 1|\mathbf{x})$.

mined pixels and superpixels that neighbor undetermined superpixels. In other words, we ignore superpixels that do not border the undetermined region. In the CRF model, the MAP state is inferred by minimizing the energy in Eqn. 3. The value of λ was determined by performing 5-fold cross-validation on the training set. For the CRF model using CM 1 $\lambda = 100$, while for CM 2 $\lambda = 1$. We consider the same set of features for herding as in the PASCAL data set. The value of λ was set for herding on using 5-fold validation as well. For herding with both CM 1 and 2, $\lambda = 10$. In these experiments $\alpha = 1$ for herding.

For herding, we consider there to be more than a single foreground or background class. That is, we divide $P(y_i = 0|\mathbf{x})$ into K replicas, $P(y_{i_1} = 0|\mathbf{x})$,...,$P(y_{i_K} = 0|\mathbf{x})$, where $P(y_{i_k} = 0|\mathbf{x}) = P(y_i = 0|\mathbf{x})/K$. This suggests that the background and foreground regions are comprised of K object sub-classes that occur with equal probability. Creating duplicate classes in this manner enables herding to explain strong boundaries that occur between neighboring superpixels that may both be background or foreground. Such strong boundaries occur because the gPb method is a general boundary (edge) finding method trained without knowledge of the foreground/background segmentation task.

Results from the local model, CRF model and herding are shown in Table 4.2. The segmentation error rate was computed across the set of undetermined pixels in the test set of images. From these results we see that the addition of pairwise marginals that encode boundary information gives a big gain over the local, independent model. The results for herding are shown with $K = 5$. It should be noted that herding with different values of K gives different performance. For example, with $K = 1$ (i.e. a single background or foreground class) herding SER is 7.68% for CM 1 and 6.97% for CM 1 - slightly worse than the CRF model. While with $K = 2$ the SER of herding is 6.47% for CM 1 and 6.36% for CM 2 - comparable to that of the CRF. This suggests that K should be set to the maximum number of object sub-classes in the foreground or background.

Segmentation Error Rate (SER)					
Local Model		**CRF**		**Herding**	
CM 1	CM 2	CM 1	CM 2	CM 1	CM 2
10.45%	10.46%	6.35%	6.58%	**6.28%**	**6.29%**

Table 2. Segmentation error rate for local model, CRF and herding on the GrabCut data set.

5. Conclusion

In this paper we illustrate a new technique for combining local, discriminatively trained classifiers over groups of (super-) pixels into a joint model over labels. The method is an alternative to conditional random fields [10] or max margin networks [14]. However the method follows a markedly different philosophy in that it never learns a joint model but rather generates representative points of some (unknown) joint distribution $P(\mathbf{y}|\mathbf{x})$. An important theoretical contribution of this paper relative to previous work [6] is that we prove that inconsistent marginals will be orthogonally projected onto the marginal polytope. This makes this technique quite unique as a way to combine inconsistent local classifiers.

We illustrate our new algorithm on image segmentation. While we do not claim to have developed the state of the art algorithm for that task, the results do show the utility of our technique in a computer vision setting. We hope that these ideas will thus find a broader range of applications within this discipline.

Acknowledgement

This work is supported by NSF grants 0447903, 0914783, 0928427, 1018433 and ONR/MURI grant 00014-06-1-073, as well as a grant from the UC Labs Research Program.

A. Proof of Proposition 1

Proof. Since herding is scale free with respect to the step-size, we can assume $\eta_\alpha = 1$ without loss of generality. We first construct another sequence of weights $\tilde{\mathbf{w}}_0 = \mathbf{w}_0, \tilde{\mathbf{w}}_t =$

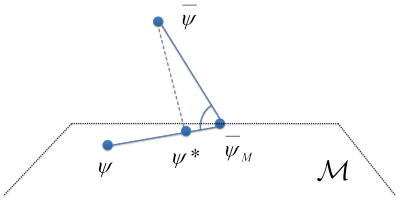

Figure 6. Inconsistent Moments

$\tilde{\mathbf{w}}_{t-1} + \bar{\psi}_{\mathcal{M}} - \psi(\mathbf{y}_t)$, where $\mathbf{y}_t$ is drawn by the original herding sequence. Applying the update equation recursively gives $\tilde{\mathbf{w}}_T = \tilde{\mathbf{w}}_0 + T(\bar{\psi}_{\mathcal{M}} - \bar{\psi}_T)$. Then to prove the proposition, it suffices to prove that the new sequence $\{\tilde{\mathbf{w}}_t\}$ satisfies the perceptron cycling theorem (PCT) [6], and consequently $\|\tilde{\mathbf{w}}_t\|$ is bounded.

Let's give the following lemma before moving on

Lemma 3. $(\bar{\psi} - \bar{\psi}_{\mathcal{M}})^T(\psi - \bar{\psi}_{\mathcal{M}}) \leq 0, \quad \forall \psi \in \mathcal{M}$

Proof. Assume there exists a point $\psi \in \mathcal{M}$ $s.t.$ the inequality doesn't hold, then the angle $\angle\psi\bar{\psi}_{\mathcal{M}}\bar{\psi}$ is an acute angle, and hence as shown in figure 6 we can always find a point on the segment $\overline{\psi\bar{\psi}_{\mathcal{M}}}$, ψ^*, such that $\|\psi^* - \bar{\psi}\|_2 < \|\bar{\psi}_{\mathcal{M}} - \bar{\psi}\|_2$. This contradicts with the fact that ψ^* is in $\mathcal{M}$ and $\psi_{\mathcal{M}}$ is the projection of $\bar{\psi}$. □

According to the herding algorithm (equation 6), $\mathbf{w}_T^T\mathbf{y}_{T+1} \geq \mathbf{w}_T^T\psi(\mathbf{y}), \forall \mathbf{y}$. Since $\bar{\psi}_{\mathcal{M}} \in \mathcal{M}$, it can be represents by a convex combination of $\{\psi(\mathbf{y})\}$ with a set of coefficients $\{\pi_i\}$, i.e. $\bar{\psi}_{\mathcal{M}} = \sum_i \pi_i\psi(\mathbf{y}_i), \pi_i > 0, \forall i$, and $\sum_i \pi_i = 1$, and thus we know

$$\mathbf{w}_T^T\bar{\psi}_{\mathcal{M}} = \sum_i \pi_i \mathbf{w}_T^T\psi(\mathbf{y}_i) \leq \mathbf{w}_T^T\psi(\mathbf{y}_{T+1}) \tag{11}$$

Observing $\tilde{\mathbf{w}}_T = \mathbf{w}_T - T(\bar{\psi} - \bar{\psi}_{\mathcal{M}})$ with Lemma 3 gives

$$\begin{aligned}&\tilde{\mathbf{w}}_T^T(\psi(\mathbf{y}_{T+1}) - \bar{\psi}_{\mathcal{M}}) \\ &= \mathbf{w}_T^T(\psi(\mathbf{y}_{T+1}) - \bar{\psi}_{\mathcal{M}}) - T(\bar{\psi} - \bar{\psi}_{\mathcal{M}})^T(\psi(\mathbf{y}_{T+1}) - \bar{\psi}_{\mathcal{M}}) \\ &\geq 0 \end{aligned} \tag{12}$$

which shows that the sequence $\{\tilde{\mathbf{w}}_t\}$ satisfies PCT, and proves the proposition.

□

Remark 4. *As long as the PCT condition for the new sequence, i.e. Eqn. 12, holds for all T, the global maximization isn't necessary. However, $\bar{\psi}_{\mathcal{M}}$ is usually unknown, and hence the condition is difficult to verify.*

Since any true expectation $\bar{\psi}_{true}$ must be inside the marginal polytope, according to Lemma 3, $\angle\psi_{true}\bar{\psi}_{\mathcal{M}}\bar{\psi}$ is an obtuse angle. This leads to Corollary 2.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour Detection and Hierarchical Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5), 2011.

[2] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *ECCV*, pages 428–441, 2004.

[3] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, September 2004.

[4] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 20(12):1222–1239, November 2001.

[5] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 670–677. IEEE, 2010.

[6] A. Gelfand, L. van der Maaten, Y. Chen, and M. Welling. On herding and the cycling perceptron theorem. In *Advances in Neural Information Processing Systems 23*, pages 694–702, 2010.

[7] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labelling. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[8] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, February 2004.

[9] S. Kumar and M. Hebert. Hebert discriminative random fields. *International Journal of Computer Vision (IJCV)*, 68(2):179–201, 2006.

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289, 2001.

[11] C. Pal, J. Weinman, L. Tran, and D. Scharstein. On learning conditional random fields for stereo. *International Journal of Computer Vision*, pages 1–19, 2010.

[12] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1848 – 1852, 2007.

[13] M. F. Tappen, C. Liu, E. H. Adelson, and W. T. Freeman. Learning gaussian conditional random fields for low-level vision. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[14] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Neural Information Processing Systems (NIPS-03)*, Vancouver, CA, 2003.

[15] M. Welling. Herding dynamic weights to learn. In *Proc. of Intl. Conf. on Machine Learning*, 2009.