Low Level Vision via Switchable Markov Random Fields

Dahua Lin
CSAIL, MIT
dhlin@mit.edu

John Fisher
CSAIL, MIT
fisher@csail.mit.edu

Abstract

Markov random fields play a central role in solving a variety of low level vision problems, including denoising, inpainting, segmentation, and motion estimation. Much previous work was based on MRFs with hand-crafted networks, yet the underlying graphical structure is rarely explored. In this paper, we show that if appropriately estimated, the MRF's graphical structure, which captures significant information about appearance and motion, can provide crucial guidance to low level vision tasks. Motivated by this observation, we propose a principled framework to solve low level vision tasks via an exponential family of MRFs with variable structures, which we call Switchable MRFs. The approach explicitly seeks a structure that optimally adapts to the image or video along the pursuit of task-specific goals. Through theoretical analysis and experimental study, we demonstrate that the proposed method addresses a number of drawbacks suffered by previous methods, including failure to capture heavy-tail statistics, computational difficulties, and lack of generality.

1. Introduction

Markov random fields provide a powerful framework for statistical image models. They have been widely used in a variety of low level vision problems, including denoising [6, 13, 22, 27], inpainting [15], segmentation [2, 6, 8, 24], and motion perception [9, 14]. Gaussian MRFs model relations between neighboring pixels through pairwise quadratic potentials. Efficient inference algorithms have contributed to their popularity in vision applications. However, as observed in previous work [12, 26], local derivative responses in natural images are highly non-Gaussian, and often exhibit heavy-tailed behavior. Gaussian models, which fail to capture this phenomena, tend to over smooth object boundaries and textures. High order MRFs [13, 20–22] were proposed to address this issue. However, these models generally require sophisticated sampling schemes [13, 18, 27] for parameter estimation. Also, the reliance on derivative filters make them vulnerable to high level of noises. While approximate variants [8, 16, 20, 21] have been developed, reliable learning and inference over high order MRFs remains a challenge.

Another family of methodologies is adaptive filtering. Representative work along this line includes anisotropic diffusion [10], adaptive weight smoothing [11], bilateral filtering [23], and non-local mean filtering [3, 4, 7]. The basic idea underlying these methods is to steer the filter kernel so as to preserve image structures such as edges and textures. It has been shown [5, 6] that they are closely related to Gaussian MRFs with inhomogeneous graphical structure and can be solved efficiently by iterative diffusion – a primary advantage over high-order MRFs.

An important issue remains, namely, that of determining the appropriate graphical structure. Whereas, most existing techniques rely on a graph structure that is determined in an ad-hoc manner, our purpose here is to develop an integrated methodology combining structural inference with traditional approaches to MRFs. Consequently, we suggest a novel MRF framework for low level vision tasks with the following desiderata. (1) *Adaptivity.* The graphical structure of the random fields should be adapted to the image structure, and thus provide a more accurate prior model, *e.g.* a graph that avoids connecting across boundaries mitigates over-smoothing. (2) *Statistical Consistency.* We aim to avoid heuristics as used in some previous methods by relying on a consistent and unified probabalistic formulation. (3) *Generality.* Previous applications of MRFs were tailored to a particular task, such as denoising, segmentation, or motion estimation. However, different tasks are often related. For example, image restoration would lead to more reliable motion estimation, while the temporal correspondence derived from the estimated motion would in turn benefit the former. Hence, it is desirable to have a task-neutral component, through which different tasks can share information. (4) *Efficiency.* Generally, increasing model complexity, such as incorporating high order interactions, leads to a more expressive representation, but often at the expense of increased computational complexity. We aim to develop a method to enhance the model's flexibility without sacrificing the efficiency enjoyed by pairwise MRFs.

With these goals in mind, we develop a formalism that

treats both the task-specific solutions and the underlying graph structure as random variables. The basic idea is to introduce switching variables that control the graphical structure of the MRF. Specifically, the graph in our formulation comprises three types of links: *observation links*, *spatial links*, and *temporal links*, whose presence/absence are controlled by switches. Via a prior distribution over the switches, we establish an exponential family of MRFs. Inference over switches adapts the spatial links to local image structures, and steers temporal links towards the motion direction. We show that with a proper switch prior, the heavy tail characteristics typically reflected in natural images can be effectively captured.

We note several important aspects of the model. (1) Rather than doing MAP estimation, we perform Bayesian inference with an aim of acquiring an approximate of the posterior. As we shall in experiments, by taking into account the uncertainty of different models, this approach improves the reliability. (2) We do not require a separate training set to learn model parameters as in many other approaches. Both the MRF structure and the task-specific solutions are directly inferred from the target image/video via the solution to a variational formulation. (3) Different low level vision tasks can be incorporated into the framework and share the underlying graph. In this way, the information from different tasks can be used to optimize the graph structure, which in turn offers useful guidance to the task-specific inference. For example, the result of motion estimation can be used to guide the connection across frames for image denoising. (4) The variational algorithm is efficient, and can be readily scaled up to handle large-volume data, such as videos.

As one of the contributions of this paper, we also derive a variational algorithm to perform efficient inference over large Gaussian MRFs. The algorithm is based on tree-reweighted approximation. Unlike the standard TRW message passing [25] that is devised for discrete distributions, it is tailored to Gaussian MRF, and guaranteed to converge.

2. Related Work

The research on Markov random fields and its application to low level vision problems has a long history. Despite the prevalence of pairwise MRFs, study on natural image statistics [12, 17] shows that they are not appropriate priors of images, as they failed to capture the heavy-tail characteristics that are typical in natural scenes.

A number of models have been developed to address this issue, among which Field of Experts (FoEs) [13] is representative. FoEs is a patch-based high order MRF model based on the product of experts framework. Steerable Random Fields [15] extends this idea by incorporating Gaussian scale mixtures [12] and steerable filters. Roth and Black [13] proposed using contrastive divergence sampling for parameter estimation, which tends to be slow for moderate-size problems. Consequently, Tappen [21] proposed variational mode learning, which optimizes a loss function instead of working on the original MLE problem. Weiss and Freeman [26] derived tractable bounds of the log partition function for a specific class of potentials, and proposed the basis rotation algorithm to optimize them. Li and Huttenlocher [9] suggest stochastic optimization to learn the model parameters. Schmidt et al. [18] argue that tailored to specific applications, these modified techniques may lead to boosted performance of a particular task, however, they lack the generality and versatility of generative MRFs.

Adaptive filtering is another important category of methods, originally developed as an improved variant of classical filtering techniques. Anisotropic diffusion [10], which involves the convolution with a space-variant filter kernel depending on image content. Black and Rangarajan [1] proposed to unify line processes and outlier processes to improve image restoration. Another representative approach is bilateral filtering [23]. The basic idea is replace each pixel with an adaptively weighted combination of its neighbors. This is further generalized by Buades et al. [3, 4] to non-local mean filtering, which no longer restricts the reference pixels to the immediate neighborhood. Gilboa and Osher [6] systematically examine various bilateral filtering and non local filtering techniques and reveal its intrinsic connections to graph-based diffusion. These methods use specific techniques to adaptively control the links between neighboring pixels, which, however, did not provide a generic model to exploit different information, *e.g*. motion.

Conditional random fields are also used in low level vision. Tappen *et al* proposed the Gaussian CRF [22] to mitigate over-smoothing without waiving the computational benefit of Gaussian models. Such a goal is partly similar to ours. However, there are two essential differences. (1) CRFs by their nature are discriminative models and task-specific, whereas our formulation is generative, and thus enjoys the generality and versatility of generative models. (2) Training a CRF in itself is nontrivial. As stated above, the formulation presented here jointly infers both the graphical structure and the task-specific solutions directly from the image or video, without requiring a separate training stage.

3. Switchable Markov Random Fields

In this section, we first revisit a pairwise MRF model for video restoration, and by analyzing its limitations, we motivate our approach. Next, we formally present the generic model of switchable MRFs in section 3.1, and derive the variational inference algorithm in section 3.2. In section 3.3, we discuss several properties of the model.

We consider a classic MRF model for video restoration, *i.e*. inferring the pixel values of a video from noisy observations. This model comprises three different types of

links: *observation-links* connect nodes to their corresponding measurements, while *spatial links* and *temporal links* enforce spatial and temporal coherence, respectively. Let $\mathbf{x}$ and $\mathbf{y}$ denote the vector of all pixel values and that of all measurements. Then the joint probability distribution is $p(\mathbf{x},\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$. Here, $p(\mathbf{x}) = \frac{1}{Z}\exp(-E_{int})$ is the scene prior, where the energy $E_{int} = E_S + E_T$ consists of two parts: the *spatial energy* E_S and the *temporal energy* E_T, as given by

$$E_S = \frac{w_S}{2}\sum_{t=1}^{T}\sum_{i\in I}\sum_{j\in N(i)}(x_{t,i}-x_{t,j})^2; \quad (1)$$

$$E_T = \frac{w_T}{2}\sum_{t=1}^{T-1}\sum_{i\in I}(x_{t+1,i}-x_{t,i})^2. \quad (2)$$

Here I is the set of all node indices in a frame, and $N(i)$ is the neighborhood of node i. Assuming that each observation $y_{t,i}$ is independently generated from a Gaussian distribution conditioned on $x_{t,i}$, we have $p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T}\prod_{i\in I}p(y_{t,i}|x_{t,i})$, with

$$p(y_{t,i}|x_{t,i}) = \frac{1}{(2\pi\sigma_y^2)^{1/2}}\exp(-(y_{t,i}-x_{t,i})^2/(2\sigma_y^2)). \quad (3)$$

Given $\mathbf{y}$, the posterior distribution of $\mathbf{x}$ is also Gaussian, as $p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z_{pos}(\mathbf{y})}\exp(-(E_{int}+E_{obs}))$, with

$$E_{obs} = \frac{1}{2\sigma_y^2}\sum_{t=1}^{T}\sum_{i\in I}(x_{t,i}-y_{t,i})^2. \quad (4)$$

This formulation has several drawbacks. First, spatial links in E_S may connect pixels across boundaries, leading to over-smoothing. To tackle this problem, we introduce a binary indicator $\theta_{t,ij}^{(S)}$ for each term in E_S, and set it to zero when the corresponding pair of pixels is across an edge. In practice, we treat it as a Bernoulli-distributed random variable, due to uncertainties arising from noisy observations. Moreover, we allow links between pixels that are not immediate neighbors of each other, which provides more flexibility in graph construction and leads to enhanced reliability.

Second, linking the nodes at the same locations across frames may not be appropriate in the presence of motion. We address this by steering temporal links in the motion direction. We do this by replacing each term $(x_{t+1,i}-x_{t,i})^2$ in E_T with $\sum_{j\in R_{t+1}(i)}\theta_{t,ij}^{(T)}(x_{t+1,j}-x_{t,i})^2$. Here, $R_{t+1}(i)$ is a set of all possible destination nodes in next frame for $x_{t+1,j}$. $\theta_{t,ij}^{(T)}$ is the indicator of whether the point at node i at time t is moved to node j at time $t+1$. The indicators for the links from the same source node are controlled by a discrete random variable from a multinomial distribution.

Third, in some cases (*e.g.* images contaminated by shot noise), Gaussian distributions may not adequately capture the observation noise. A robust way to handle this is to explicitly model outliers, as

$$p(y_{t,i}|x_{t,i}) = (1-c)\mathcal{N}(y_{t,i}|x_{t,i},\sigma_y^2) + cq_0. \quad (5)$$

Here, for each pixel, the observation can be generated from an outlier distribution q_0 with probability $(1-c) > 0$. This is equivalent to replacing each term in E_{obs} with

$$\frac{\theta_{t,i}^{(O)}}{2\sigma_y^2}(x_{t,i}-y_{t,i})^2 + (1-\theta_{t,i}^{(O)})(-\log q_0). \quad (6)$$

Intuitively, $\theta_{t,i}^{(O)}$ here can be considered as a switch that controls the connection between $x_{t,i}$ and $y_{t,i}$, which can be detached with cost $(-\log q_0)$. The switch itself is generated from a Bernoulli distribution with $P(\theta_{t,i}^{(O)}=1) = c$.

Though motivated differently, these variant models share an important aspect in common, namely using switches to control the graph structure (e.g. the presence of edges). Hence, we call them *switchable Markov random fields*.

3.1. Exponential Family of Switchable MRFs

In general, the probabilistic formulation of switchable MRFs can be expressed in the following form

$$p(\mathbf{x}|\boldsymbol{\theta},\mathbf{y}) = \frac{1}{Z}\exp\left(\sum_{k=1}^{M}\sum_{i=1}^{n_k}\theta_{k,i}\phi_{k,i}(\mathbf{x}_{C_k},\mathbf{y})\right). \quad (7)$$

Here, $\mathbf{x}$, $\mathbf{y}$, and $\boldsymbol{\theta}$ respectively denote the variables to be inferred, the given observations, and the structure switches. All potentials are divided into M modules. Each module corresponds to a specific aspect of the model and is controlled by a group of switches where $\mathbf{x}_{C_k}$ denotes those variables involved in the k-th module. Clearly, this is an exponential family model, and the natural parameters are precisely the switches. Assuming that the switches for different modules are independently generated from respective prior distributions, then the joint distribution is given by

$$p(\mathbf{x},\boldsymbol{\theta}|\mathbf{y};\boldsymbol{\tau}) = p(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})\prod_{k=1}^{M}p_k(\boldsymbol{\theta}_k|\boldsymbol{\tau}_k). \quad (8)$$

Here $\boldsymbol{\theta}_k$ refers to the vector of switches for the k-th module with associated hyper-parameter $\boldsymbol{\tau}_k$. Note that choosing priors in the following form leads to analytic inference updates as we will see later:

$$p_k(\boldsymbol{\theta}_k|\boldsymbol{\tau}_k) = \exp(\boldsymbol{\tau}_k^T\boldsymbol{\theta}_k - A_k(\boldsymbol{\tau}_k)). \quad (9)$$

Binomial, multinomial and many other well-known distributions are in this form.

3.2. Variational Inference Algorithm

Based on the formulation presented above, the goal of inference is to solve the marginal distributions of $\mathbf{x}$ and $\boldsymbol{\theta}$ from the joint posterior given by Eq.(8). If we choose the priors as in Eq.(9), the distribution can be written as

$$\exp\left(\sum_{k=1}^{M} \boldsymbol{\theta}_k^T(\boldsymbol{\tau}_k + \boldsymbol{\phi}_k(\mathbf{x}_{C_k}, \mathbf{y})) - A(\boldsymbol{\tau}, \mathbf{y})\right). \quad (10)$$

Here, $\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,n_k})$ is the potential vector for the k-th module. In general, obtaining the exact marginals can be intractable due to the nonlinearity $\boldsymbol{\phi}_k$. Leveraging the exponential family form, we derive a variational algorithm via the mean field approximation. Specifically, we consider the product distributions over $\mathbf{x}$ and $\boldsymbol{\theta}$ as follows

$$q(\mathbf{x}, \boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{\zeta}) = q_x(\mathbf{x}|\boldsymbol{\eta}) \prod q_{\theta_k}(\boldsymbol{\theta}_k|\boldsymbol{\zeta}_k). \quad (11)$$

Here, $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ are variational parameters. Our objective here is to seek a distribution $\hat{q}$ from this family that optimally approximates the posterior distribution $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}; \boldsymbol{\tau})$ in terms of minimizing the K-L divergence. With p given, this is equivalent to maximizing the following objective function:

$$L(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \sum_{k=1}^{M}\sum_{i=1}^{n_k} \mathrm{E}_{\boldsymbol{\zeta}_k}[\boldsymbol{\theta}_k]^T \mathrm{E}_{\boldsymbol{\eta}}[\boldsymbol{\phi}_k(\mathbf{x}_{C_k}, \mathbf{y})] + H_{q_x}(\boldsymbol{\eta}) + \sum_{k=1}^{M} H_{q_{\theta_k}}(\boldsymbol{\zeta}_k). \quad (12)$$

Here, H_{q_x} and $H_{q_{\theta_k}}$ are the entropies of q_x and q_{θ_k} as functions of $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}_k$ respectively. This problem can be solved by alternatively updating $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ leaving the other fixed. Suppose we choose $q_{\theta_k}(\boldsymbol{\theta}_k|\boldsymbol{\zeta}_k)$ as below

$$q_{\theta_k}(\boldsymbol{\theta}_k|\boldsymbol{\zeta}_k) = \exp(\boldsymbol{\zeta}_k^T\boldsymbol{\theta}_k - A_k(\boldsymbol{\zeta}_k)). \quad (13)$$

With $\boldsymbol{\eta}$ fixed, there is a close-form formula for updating $\boldsymbol{\zeta}_k$:

$$\hat{\boldsymbol{\zeta}}_k = \boldsymbol{\tau}_k + \mathrm{E}_{\boldsymbol{\eta}}[\boldsymbol{\phi}_k(\mathbf{x}_{C_k}, \mathbf{y})]. \quad (14)$$

On the other hand, when $\boldsymbol{\zeta}$ is fixed, the optimal $\boldsymbol{\eta}$ can be obtained by solving the following problem:

$$\hat{\boldsymbol{\eta}} = \operatorname*{argmax}_{\boldsymbol{\eta}} \sum_{k=1}^{M}\sum_{i=1}^{n_k} \bar{\boldsymbol{\theta}}_k^T \mathrm{E}_{\boldsymbol{\eta}}[\boldsymbol{\phi}_k(\mathbf{x}_{C_k}, \mathbf{y})] + H_{q_x}(\boldsymbol{\eta}). \quad (15)$$

Here, $\bar{\boldsymbol{\theta}}_k = \mathrm{E}_{\boldsymbol{\zeta}_k}[\boldsymbol{\theta}_k]$. This is equivalent to performing variational inference of $\mathbf{x}$ over a "mean MRF" with each potential term weighted by $\bar{\boldsymbol{\theta}}_k$. Both Eq.(14) and (15) involve $\mathrm{E}_{\boldsymbol{\eta}}[\boldsymbol{\phi}_k(\mathbf{x}_{C_k}, \mathbf{y})]$. Hence, it is advisable to choose the form of q_x such that these expectation terms are easy to work with. For instance, if $\boldsymbol{\phi}_k$ is quadratic, it is convenient to let q_x be a Gaussian. Depending on the form of base MRF, one can apply various inference techniques to solve this problem, and resort to further approximation when necessary.

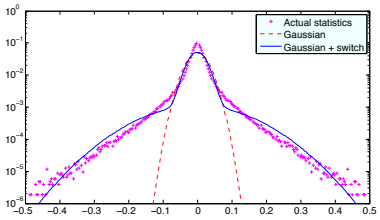


Figure 1. This figure compares the pdf yielded by standard Gaussian model and that by a switchable model with the actual distribution of neighbor pixel difference obtained from natural images. Note that the y-axis is in log scale.

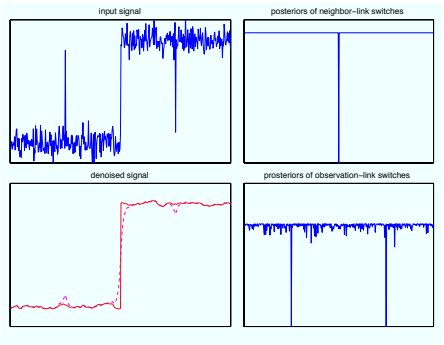


Figure 2. The top-left of this figure shows the input signal, while the bottom-left compares the results obtained from standard Gaussian MRF and switchable MRF, respectively depicted with solid line and dashed line. On the right column are the posterior probabilities of switches for observation-links and neighbor-links.

3.3. Analysis of the Model

We examine several important aspects of the switchable MRFs. Again, take denoising for example. Consider two pixels x_i and x_j jointly generated from a Gaussian distribution with zero mean, marginal variance σ^2 and correlation coefficient ρ. Then, the distribution of $\Delta x = x_i - x_j$ is also Gaussian, with variance $2\sigma^2(1-\rho)$. With a switch θ with prior $P(\theta = 1) = p_0$ added to the link between them, the marginal of Δx becomes a Gaussian scale mixture, as

$$p_0\mathcal{N}(\Delta x|0, 2\sigma^2(1-\rho)) + (1-p_0)\mathcal{N}(\Delta x|0, 2\sigma^2). \quad (16)$$

Figure 1 shows that with such change, the heavy-tail characteristics in natural images can be effectively captured.

Next, with the purpose of testing its response to sharp changes and outlier observations, we construct a signal with a swift jump, and superimpose on it with white noises and sparse pulses. Then, we apply our method to recover it. The results are shown in figure 2. We can see that standard Gaussian MRF leads to over-smoothing across the big jump, while the switchable MRF preserves the sharpness of the change by breaking the neighbor-link there. Moreover, the effect of the pulses that could otherwise affect the recovered signal has been successfully filtered out by detaching the corresponding observation links.

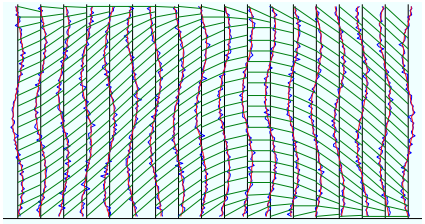


Figure 3. This figure shows the direction of temporal switches inferred from a switchable model constructed on a dynamic 1D signal. Each switch selects a link connecting a source point to the corresponding target point in next time. Note that only a downsampled subset of such links are depicted here.

Finally, we apply the model in a dynamic context, which uses switches to steer the temporal links across time. In particular, we construct a dynamic 1D signal which moves forward and backward periodically. The inferred temporal links are illustrated in figure 3. We can see that they are successfully adapted to the signal motion.

The analysis above clearly demonstrates that with switches incorporated to control the graph structure, our model effectively addresses the difficulties faced by traditional Gaussian MRFs. Compared to the models relying on high order interactions, this way is much more efficient.

4. Integrated Low-Level Vision System

We develop a low-level vision system that integrates denoising, inpainting, segmentation, and motion estimation based on switchable MRFs.

4.1. System Overview

As illustrated in figure 4, the method comprises a graph at its core underlying various vision tasks. Each node of the graph is connected to a pixel via an *observation link*, and to other nodes in the same frame within a distance to it via *spatial links*. Each observation link and spatial link is controlled by an independent switch with a binomial prior. For videos, we introduce additional *temporal links* connecting each node to possible destination nodes in the next frame. Temporal links with the same source node are controlled by a switch with a multinomial prior selecting which particular link to turn on.

Each vision task of our framework is accomplished through the inference over an MRF model constructed based on the common graph. While these tasks share the same underlying graph, the forms of potentials can be different for different tasks. Specifically, each node is associated with an intrinsic pixel value x_i and a region label z_i. For denoising or inpainting, we use Gaussian MRFs as a base model with pairwise energies of the form $w(x_i - x_j)^2/2$; while for segmentation, we use a Potts model with pairwise energies of the form $w\mathbb{I}(z_i \neq z_j)$.

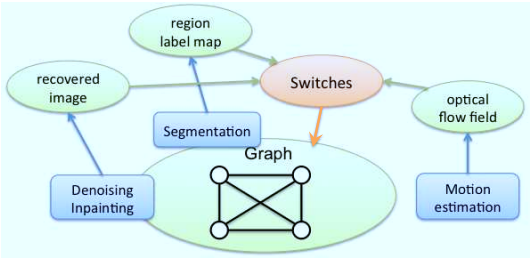


Figure 4. The high-level picture of our low level vision system. Various low level vision tasks share the same underlying graph that captures the appearance and motion structure. In this framework, the graph is controlled by a collection of switches, which are estimated based on both appearance and motion information, and in turn guide the inference in solving the low level vision tasks.

When processing videos, we also incorporate motion information. Currently, we use a state-of-the-art motion estimator [19] to derive optical flow measurements and use them to parameterize a switching prior over temporal links. Concretely, we define temporal link energies, as

$$\sum_{j \in R_{t+1}(i)} \theta_{t,ij}^{(T)} \left(\frac{w_T}{2}(x_{t+1,j} - x_{t,i})^2 + \frac{w_M}{2}||s_j - u_{t+1,i}||^2 \right).$$

Here, s_j is the coordinate of the j-th site, and $u_{t+1,i}$ denotes the predicted location of the i-th pixel in time $t+1$ according to the reference motion field. This is equivalent to treating the flow vectors as a noisy observation and encouraging linking to nodes with similar pixel values and that are close to the predicted destination.

4.2. Inference over Large-Scale MRF

The overall inference procedure is as presented in section 3.2. In each iteration, we sum up the potentials from all tasks to update the posterior of switches using Eq.(14), and then perform task-specific inference based on the updated graph using Eq.(15). We note that the task of denoising and inpainting entails the inference over a large-scale Gaussian MRF that could contain over millions of nodes.

Given a Gaussian MRF, and thus its information matrix $\mathbf{J}$ and linear potential vector $\mathbf{h}$, the goal of inference is to derive the means, variances, and covariances. Here, the mean $\mathbf{J}^{-1}\mathbf{h}$ can be solved efficiently using a sparse equation solver. However, direct computation of the covariance matrix $\mathbf{J}^{-1}$ can be infeasible for large-scale model. An approximate method is loopy belief propagation (LBP), which is not guaranteed to converge and its estimation of the marginal variances are known to be optimistic. Moreover, we note that $\mathrm{E}[(x_i - x_j)^2] = (\mu_i - \mu_j)^2 + (\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij})$. Here, μ_i, μ_j are the means of x_i and x_j, σ_i^2, σ_j^2 are their variances, and σ_{ij} is their covariance. This implies that the computation of the expected potential associated to an edge requires not only the marginal means and variances, but also the covariance, which LBP is not able to provide.

We derive an efficient inference algorithm based on the tree-reweighted approximation [25]. We only present the key results; the detailed derivation is provided in the supplement. In this algorithm, we solve for the variances as well as the covariances between directly connected nodes, by maximizing a variational objective, as

$$-\frac{1}{2}\mathrm{E}_q[\mathbf{x}^T\mathbf{J}\mathbf{x}] + \sum_{v\in V} H_v(\sigma_v^2) - \sum_{\{i,j\}\in E} I_{ij}(\rho_{ij}). \quad (17)$$

Here, V and E are the sets of nodes and edges. $H_v(\sigma_v^2) = \log(\sigma_v^2)/2 + C$ is the marginal entropy at node v that depends only on the marginal variance and $I_{ij}(\rho_{ij}) = -\log(1-\rho_{ij}^2)/2$ is the mutual information between node i and j, depending only on the correlation coefficient ρ_{ij}. The solution is via iterative updates as follows

$$\sigma_v \leftarrow (2J_{vv})^{-1}(\sqrt{b_v^2 + 4J_{vv}} - b_v), \quad (18)$$

$$\rho_{ij} \leftarrow (2a_{ij})^{-1}(\sqrt{\beta_{ij}^2 + 4a_{ij}^2} - \beta_{ij}). \quad (19)$$

Here, $a_{ij} = J_{ij}\sigma_i\sigma_j$ and $b_v = \sum_{u\in N(v)} J_{vu}\rho_{vu}\sigma_u$. Differing from previous work on tree-reweighted inference, which uses a fixed-point message passing scheme and is restricted to discrete MRFs, ours is an alternate optimzation method tailored for Gaussian MRFs with guaranteed convergence.

5. Experiments

We conducted experiments to test the proposed approach on real data, and compare it with other methods qualitatively and quantitatively on denoising. We also demonstrate its capability in image inpainting and segmentation.

Given an image or video, we first normalize it such that the dynamic range of the pixel values is in $[0, 1]$. Then, we obtain an initial guess of the true pixel values through median filtering with 5×5 window. Using the filtered result, we calculate the median of neighbor-pixel-difference (denoted by δ_{nb}), and roughly estimate the noise variance (denoted by σ_0^2). The base weights of the neighbor links and the observation links are respectively set to $1/(2\delta_{nb}^2)$ and $1/(2\sigma_0^2)$. The prior confidence of each link is set to 0.999. This setting works well in most cases under consideration and was used through our experiments.

5.1. Image Denoising and Inpainting

We test the image denoising performance on a set of about 600 images collected from internet, which comprises images of various categories. The data set is constructed with an aim to provide a unified testbed for various vision tasks. We compare our method with standard MRF-based denoising, non-local mean filtering [6], and field of experts [13][1]. The results are shown in figure 5. We can see that our method works well across different conditions, producing clean and sharp images. The results yielded by non-local mean filtering are relatively blurry. FOE works well in moderate SNR conditions (e.g. PSNR¿20) which are consistent with the original paper. However, its performance degrades noticeably as the noise variance continues to increase. This issue has been observed by other authors. There are two reasons: (1) the standard FOE model uses 25 filters with small kernels (5x5) and do not yield robust performance in lower SNR conditions, and (2) the complexity of FOE models in some cases may also contribute to its vulnerability to low SNR. When the images are corrupted by shot noise, non-local mean filtering performs very poorly by propagating of the effects of outlier pixels. More results are provided in the supplement.

We also evaluate the performance quantitatively in terms of PSNR, with different levels of noise. In particular, for Gaussian white noise, we respectively test on noise variances from 0.005 to 0.1; while for shot noise, we respectively test on noise densities from 0.02 to 0.3. Figure 6 shows the average performance obtained on all images in our testing set. For the cases with Gaussian noise, our method performs consistently better than standard MRF model and non-local mean filtering. Whereas FoE yields very good performance comparable to ours when noise variance is small (e.g. 0.005), it breaks down quickly as the noise variance increases. This is partly due to the vulnerability of high order models to noise. For the cases with shot noise, both standard MRF and non-local mean filtering performs poorly; while our method works drastically better and produces near perfect results. The switches of the observation links that would automatically detach the outlier pixels play a key role in its success.

Next, we apply our method to image inpainting. In particular, we infer the pixel values occluded by the text as shown in figure 7. We see that our algorithm recovers the scene nearly perfectly. The superior performance of our switchable models is mainly due to two reasons: (1) By incorporating switches to adapt the graphical structure, it preserves local image structures like edges. Moreover, the iterative inference procedure refines the graph gradually, leading to further improvement. (2) The simplicity of the model structure makes it resilient to large noise variance.

Compared to high order models, our method runs much more efficiently (5 times faster than FoEs in testing stage), and does not require a pre-training stage, which in itself is time-consuming and difficult for most high order models.

5.2. Video Denoising

For video denoising. a graph with temporal links is established over the entire video. The input video as well as the denoised results are provided in the supplement package. Figure 8 demonstrates the adapted temporal link struc-

[1]We use the MATLAB codes as well as the pre-trained models (with 24 filters on 5×5 cliques) publicly available in Roth's website.

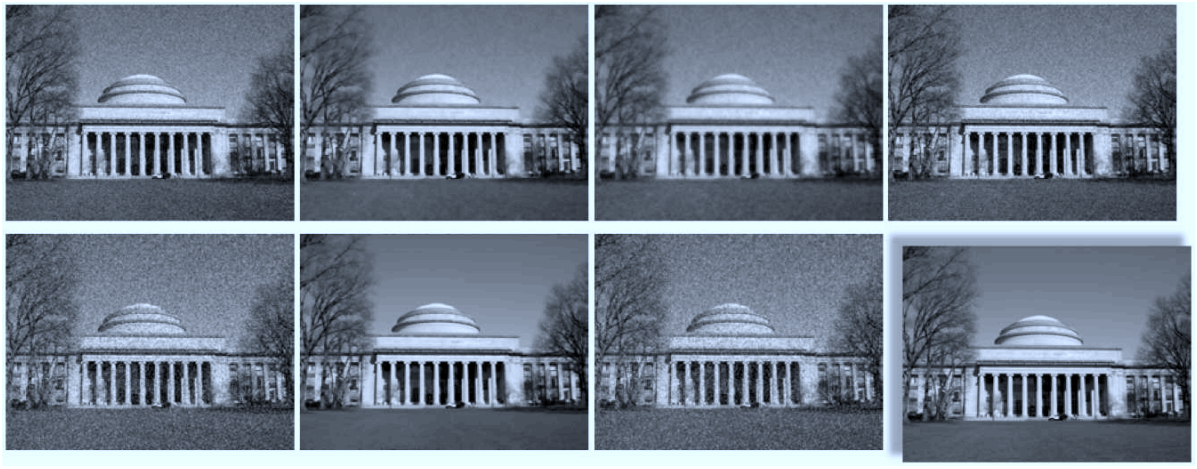
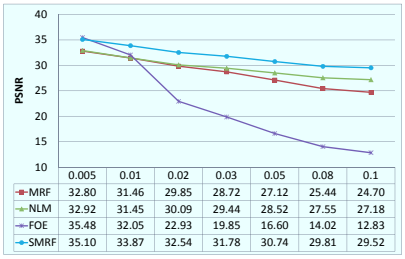
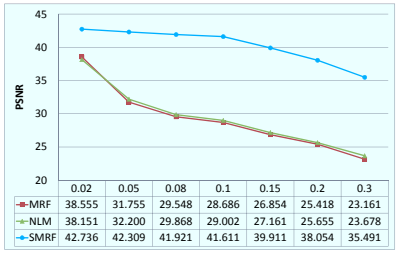


Figure 5. The results of image denoising. **Rows:** The first row are the results on an image with Gaussian noise with variances 0.02. The second row are the result with shot noise of density 0.2. **Columns:** From left to right are the noisy images, the results of switchable MRF, non-local mean, and field of experts. Note that the last image in the second row is the ground-truth.



	0.005	0.01	0.02	0.03	0.05	0.08	0.1
MRF	32.80	31.46	29.85	28.72	27.12	25.44	24.70
NLM	32.92	31.45	30.09	29.44	28.52	27.55	27.18
FOE	35.48	32.05	22.93	19.85	16.60	14.02	12.83
SMRF	35.10	33.87	32.54	31.78	30.74	29.81	29.52

(a) Results on images with gaussian white noise



	0.02	0.05	0.08	0.1	0.15	0.2	0.3
MRF	38.555	31.755	29.548	28.686	26.854	25.418	23.161
NLM	38.151	32.200	29.868	29.002	27.161	25.655	23.678
SMRF	42.736	42.309	41.921	41.611	39.911	38.054	35.491

(b) Results on images with shot noise

Figure 6. Quantitative comparison of the denoising performance yielded by different methods with different levels of noise. The curves reflect the results averaged over the 600 testing images.

Figure 7. The left is an image corrupted by description text, and the right is the recovered image produced our inpainting method.

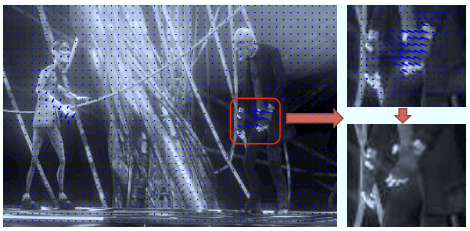


Figure 8. On the left is a map of "average temporal link directions" overlain on top of the frame, which are the difference vectors between the posterior mean of the destination node coordinates and the source coordinates. The two pictures on the right focus on a local part that captures a moving hand.

ture by visualizing the "mean temporal link directions", which shows that our inference algorithm can effectively steer the links towards the motion direction.

5.3. Segmentation

As we have discussed above, our inference algorithm would lead to a graph that preserves the image structure, such as object boundaries. To test this, we perform segmentation using graph cut [2] based on the posterior mean graph yielded by our method. In doing so, we manually specify the initial seeds of the foreground and background regions. The result is shown in figure 9. We see that the bridge is properly separated from the background. This result in a sense confirms our belief that the graph structure derived by our approach is consistent with the image structure.

6. Conclusion

In this paper, we presented a new framework for low level vision, based on switchable MRFs. By introducing

Figure 9. On the left is the input image with strokes indicating the initial seeds of foreground and background. On the right is the segmentation result. We reduced the brightness of the background part so as to distinguish it from the foreground.

switches to control the underlying graphical structure, we formulated an exponential family of MRFs, and thereupon derived a variational inference algorithm. Then we developed a low level vision system that couples different tasks via a shared graph. The inference over this system not only leads to improved task-specific solutions, but also provides a graph that adapts to the image structure. The results of comparative experiments on real data clearly showed that our approach effectively overcomes the issues suffered by traditional methods, and exhibits substantially better robustness than high order models. In addition, it runs efficiently and does not require a time-consuming pre-training stage.

Acknowledgement

D. Lin was partially supported by the Office of Naval Research Multidisciplinary Research Initiative (MURI) program, award N000141110688. J. Fisher was partially supported by the Defense Advanced Research Projects Agency, award FA8650-11-1-7154.

References

[1] M. Black and A. Rangarajan. The outlier process: Unifying line processes and robust statistics. In *Proc. of CVPR*, 1994. 2

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1233, 2001. 1, 7

[3] a. Buades, B. Coll, and J.-M. Morel. A Non-local Algorithm for Image Denoising. In *proc. of CVPR'05*, 2005. 1, 2

[4] a. Buades, B. Coll, and J. M. Morel. A Review of Image Denoising Algorithms, with a New One. *Multiscale Modeling & Simulation*, 4(2):490, 2005. 1, 2

[5] G. Gilboa, J. Darbon, S. Osher, and T. Chan. Nonlocal Convex Functionals for Image Regularization, 2006. 1

[6] G. Gilboa and S. Osher. Nonlocal Linear Image Regularization and Supervised Segmentation. *Multiscale Modeling and Simulation*, 6(2):595, 2007. 1, 2, 6

[7] S. Kindermann, S. Osher, and P. W. Jones. Deblurring and Denoising of Images by Nonlocal Functionals. *Multiscale Modeling & Simulation*, 4(4):1091, 2005. 1

[8] P. Kohli and M. Kumar. Energy Minimization for Linear Envelope MRFs. In *Proc. of CVPR'10*, 2010. 1

[9] Y. Li and D. Huttenlocher. Learning for Optical Flow Using Stochastic Optimization. In *Proc. of ECCV'08*, 2008. 1, 2

[10] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990. 1, 2

[11] J. Polzehl and V. G. Spokoiny. Adaptive weights smoothing with applications to image restoration. *J. R. Statist. Soc. B*, 62:335–354, 2000. 1

[12] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image Denoising using Scale Mixtures of Gaussians in the Wavelet Domain. *IEEE Transactions on Image Processing*, 12(11):1338–51, 2003. 1, 2

[13] S. Roth and M. J. Black. Fields of Experts: A Framework for Learning Image Priors. In *Proc. of CVPR'05*, 2005. 1, 2, 6

[14] S. Roth and M. J. Black. On the Spatial Statistics of Optical Flow. In *Proc. of ICCV'05*, 2005. 1

[15] S. Roth and M. J. Black. Steerable Random Fields. In *Proc. of ICCV'07*, 2007. 1, 2

[16] K. Samuel and M. Tappen. Learning Optimized MAP Estimates in Continuously-valued MRF Models. In *Proc. of CVPR'09*, 2009. 1

[17] H. Scharr, M. J. Black, and H. W. Haussecker. Image statistics and anisotropic diffusion. In *Proc. of ICCV'03*, 2003. 2

[18] U. Schmidt, Q. Gao, and S. Roth. A Generative Perspective on MRFs in Low-Level Vision. In *Proc. of CVPR'10*, 2010. 1, 2

[19] D. Sun, S. Roth, and M. J. Black. Secrets of Optical Flow Estimation and Their Principles. In *Proc. of CVPR'10*, 2010. 5

[20] M. Tanaka and M. Okutomi. Locally Adaptive Learning for Translation-Variant MRF Image Priors. In *Proc. of CVPR'08*, 2008. 1

[21] M. F. Tappen. Utilizing Variational Optimization to Learn Markov Random Fields. In *Proc. of ICCV'07*, 2007. 1, 2

[22] M. F. Tappen, C. Liu, E. H. Adelson, and W. T. Freeman. Learning Gaussian Conditional Random Fields for Low-Level Vision. In *Proc. of CVPR'07*, 2007. 1, 2

[23] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. of ICCV'98*, 1998. 1, 2

[24] S. Vicente, V. Kolmogorov, and C. Rother. Joint Optimization of Segmentation and Appearance Models. In *Proc. of ICCV'09*, 2009. 1

[25] M. J. Wainwright, T. Jaakkola, and A. Willsky. A New Class of Upper Bounds on the Log Partition Function. *IEEE Transactions on Information Theory*, 51(7):2313–2335, 2005. 2, 6

[26] Y. Weiss and W. T. Freeman. What Makes a Good Model of Natural Images? In *Proc. of CVPR'07*, 2007. 1, 2

[27] S.-C. Zhu, Y.-N. Wu, and D. Mumford. Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998. 1