Whitened Expectation Propagation: Non-Lambertian Shape from Shading and Shadow

Brian Potetz
Google Inc.
potetz@google.com

Mohammadreza Hajiarbabi
University of Kansas
mhajiarb@ittc.ku.edu

Abstract

For problems over continuous random variables, MRFs with large cliques pose a challenge in probabilistic inference. Difficulties in performing optimization efficiently have limited the probabilistic models explored in computer vision and other fields. One inference technique that handles large cliques well is Expectation Propagation. EP offers run times independent of clique size, which instead depend only on the rank, or intrinsic dimensionality, of potentials. This property would be highly advantageous in computer vision. Unfortunately, for grid-shaped models common in vision, traditional Gaussian EP requires quadratic space and cubic time in the number of pixels.

Here, we propose a variation of EP that exploits regularities in natural scene statistics to achieve run times that are linear in both number of pixels and clique size. We test these methods on shape from shading, and we demonstrate strong performance not only for Lambertian surfaces, but also on arbitrary surface reflectance and lighting arrangements, which requires highly non-Gaussian potentials. Finally, we use large, non-local cliques to exploit cast shadow, which is traditionally ignored in shape from shading.

1. Introduction

Probabilistic inference for large loopy graphical models has become an important subfield with a growing body of applications, including many in computer vision. One common method of optimization is belief propagation. BP estimates marginals by minimizing the Bethe free energy, which is an approximate distance measure between a factorized distribution and a set of marginals, based on KL-divergence. The run time of BP is exponential in the clique size C: each potential requires $\mathcal{O}(CM^C)$ operations, where M is the number of states for each variable. This has limited the space of probabilistic models that can be explored by the computer vision community, especially for problems over continuous-valued variables. Several methods have been proposed which reduce the run time of BP [17]. Others have advanced methods of inference which can be applied to probabilistic models over discrete variables with large cliques[10, 24], or large numbers of small cliques [12]. These methods have resulted in significant progress for several applications. Nevertheless, efficient inference for large cliques remains limited to certain forms of potentials, and remains quadratic or worse in clique size.

In 2001, Minka proposed a generalization of BP known as Expectation Propagation [14]. EP estimates moments such as the mean of the distribution (which is also its minimum mean-squared error point estimate). EP works by approximating a factorized distribution with a simpler, tractable distribution from a family of distributions whose moments can be readily computed. When the approximating family is a product of independent univariate marginals, EP is equivalent to BP [14]. Thus, EP is a generalization of BP that permits the inference to account for correlation and dependencies between variables.

For continuous variables, the approximating family used by EP is nearly always Gaussian, due to computational constraints. The principal difference between BP and Gaussian EP can thus be summarized by a trade-off in their respective approximating families: BP favors flexible non-Gaussian marginals, while Gaussian EP favors a flexible covariance structure. Note that the success of either method is not solely dependent on the quality of the approximating family. For example, tree-shaped graphical models can have strong covariance structure, and so the approximating family of BP may be very poor for such models. Nevertheless, BP performs exact inference on trees. In contrast, Gaussian EP can fail even on univariate graphs if the potentials are sufficiently non-Gaussian. In a complex graph, however, accurate covariance models can improve performance because updates to one variable immediately affect distant variables known to be correlated.

Despite its success in a variety of applications [15], Gaussian EP is not common in computer vision. Many computer vision problems are defined over a grid, such as stereo, shape from shading, image super-resolution, and others.

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.219

BP has produced highly successful approaches to each of these, but EP is rarely applied to such problems. One possible reason is that problems in vision are often highly non-Gaussian. For example, properties of images tend to have highly kurtotic distributions [8, 20]. Another possible explanation is that for a grid-based graphical model with D pixels, Gaussian EP requires $\mathcal{O}(D^2)$ space and a run time of $\mathcal{O}(D^3)$. Variations of EP have been proposed to reduce the run time and apply EP to problems in computer vision [16, 22]. However, run times remain quadratic or more in the number of pixels for these variations.

Still, Gaussian EP has properties that would be highly desirable in computer vision. Its running time is independent of clique size, and instead depends polynomially on the *rank* (or intrinsic dimensionality) of each potential (defined below). In this paper, we propose an efficient inference method that retains the computational advantages of EP, reducing run time and space requirements to linear in the number of pixels, while remaining linear in clique size. This is achieved by limiting EP to efficient families of covariance structures chosen based on the statistics of natural scenes. We then test this approach on a problem with highly non-Gaussian potentials: non-Lambertian shape from shading (SfS). Our algorithm demonstrates competitive results on Lambertian SfS, and extends successfully to arbitrary reflectances, which is a novel result in SfS. Finally, we use the method to efficiently perform inference over large cliques produced by cast shadows and by global spatial priors.

2. Expectation Propagation

The aim of expectation propagation (EP) is to approximate a factorized probability distribution

$$P(\vec{x}) = \frac{1}{Z}\prod_{i=1}^{N}\phi_i(\vec{x}_i) \qquad \vec{x}_i \subseteq \vec{x} \tag{1}$$

using an exponential family distribution

$$\tilde{P}(\vec{x}|\vec{\theta}) \propto \exp\Big(\sum\nolimits_j \theta_j \tau_j(\vec{x})\Big) \tag{2}$$

In particular, EP seeks to minimize the KL-divergence $\mathbf{D}(P||\tilde{P})$, which, if achieved, would imply that the expected value of $\tau_j(\vec{x})$ was equivalent under the two distributions: $\mathbf{E}_{\tilde{P}}[\tau_j(\vec{x})] = \mathbf{E}_P[\tau_j(\vec{x})]$. The family $\tilde{P}$ is chosen so that $\mathbf{E}_{\tilde{P}}[\tau_j(\vec{x})]$ can be estimated easily. For example, if $\tau_j(\vec{x}) = x_i$ for some j, then EP can be used to estimate $\mathbf{E}_P[x_i]$, which provides the minimum mean-squared error point estimate of P. EP achieves this goal by approximating each potential function $\phi_i(\vec{x})$ with an exponential family distribution $\tilde{P}_i(\vec{x}_i|\vec{\theta}^{(i)})$. Rather than fitting each $\tilde{P}_i$ to approximate ϕ_i directly, EP iteratively chooses $\theta^{(i)}$ to minimize the KL-divergence $\mathbf{D}(\phi_i(\vec{x}_i)\prod_{j\neq i}^{N}\tilde{P}_j(\vec{x}_j|\vec{\theta}^{(j)}) \mid\mid \tilde{P}(\vec{x}|\vec{\theta}))$, where $\tilde{P}(\vec{x}|\vec{\theta}) \propto \prod_{i=1}^{N}\tilde{P}_i(\vec{x}_i|\vec{\theta}^{(i)})$ and $\vec{\theta} = \sum_i \vec{\theta}^{(i)}$. The process is repeated until the distribution hopefully converges. Minka showed that when $\vec{x}$ is discrete-valued and the approximating exponential family is a product of independent univariate discrete distributions, then EP is equivalent to classical belief propagation (BP) [14]. Thus, EP differs from BP in that it can account for covariance and interdependencies between variables, which may improve performance and require fewer iterations.

When the elements of the vector $\vec{x}$ are real-valued, the approximating exponential family is nearly always chosen to be Gaussian: $G(\vec{x}) \propto \exp(-\frac{1}{2}(\vec{x}-\vec{\mu})'S^{-1}(\vec{x}-\vec{\mu}))$. For each factor i, Gaussian EP performs three steps. In step 1, the parameters of $G_{\setminus i}(\vec{x}) \propto \prod_{j\neq i}^{N} G_j(\vec{x}_j)$ are computed:

$$S_{\setminus i} = (S^{-1} - S_i^{-1})^{-1} \tag{3}$$

$$\mu_{\setminus i} = S_{\setminus i}(S^{-1}\mu - S_i^{-1}\mu_i) \tag{4}$$

In step 2, the posterior distribution parameters S and μ are updated to match the mean and variance of $\phi_i(\vec{x}_i)G_{\setminus i}(\vec{x})$. If $\phi_i(\vec{x}_i)G_{\setminus i}(\vec{x})$ has a simple analytic form, its mean and variance can by found by differentiating the log partition function. Alternatively, sampling may be used. Finally, in step 3, the parameters S_i and μ_i are updated:

$$S_i = (S^{-1} - S_{\setminus i}^{-1})^{-1} \tag{5}$$

$$\mu_i = S_i(S^{-1}\mu - S_{\setminus i}^{-1}\mu_{\setminus i}) \tag{6}$$

These computations can be made more efficient when ϕ_i is of reduced rank. We define ϕ_i to have rank K if ϕ_i can be expressed as $\phi_i(V_i\vec{x})$ for a $K\times D$ matrix V_i, where D is the dimensionality of $\vec{x}$. It can be shown that it is sufficient to store $V_i\vec{\mu}_i$ and $V_iS_iV_i'$ rather than the full $D\times D$ matrix S_i [15]. This allows the step 1 and 3 updates to be completed in $\mathcal{O}(K^3 + KD^2)$ time, rather than $\mathcal{O}(D^3)$.

Regardless of the rank of each potential, the covariance matrix of the posterior S remains full-rank, and must be stored as a $D\times D$ matrix. For large problems with tens of thousands of variables or more, this becomes limiting. For problems that seek to infer an image, such as denoising, stereo, matting, novel view synthesis, or shape from shading, the number of variables is equal to the number of pixels. Thus, even for a small 256×256 image, D is over 65 thousand and S could require over 30 gigabytes to store. This may be one reason why EP has not become commonplace for these or similar computer vision problems.

If the graphical model underlying equation 1 is sparsely connected, it may alleviate memory requirements to store the inverse covariance matrix S^{-1} rather than S. It can be shown that S^{-1} contains non-zero entries only between variable nodes that share a potential [22]. Suppose that in step 2 of EP, V_iSV_i' and $V_i\mu$ are found through sampling or by differentiating the log partition function. The update to

S^{-1} and μ can be written:

$$S^{-1} = S_{\setminus i}^{-1} + V_i'((V_i S V_i')^{-1} - (V_i S_{\setminus i} V_i')^{-1})V_i \quad (7)$$

$$\mu = \mu_{\setminus i} + S_{\setminus i} V_i'(V_i S_{\setminus i} V_i')^{-1}(V_i \mu - V_i \mu_{\setminus i}) \quad (8)$$

The bottleneck of this approach is the multiplication by $S_{\setminus i}$ in equation 8, because we have stored $S_{\setminus i}^{-1}$ rather than $S_{\setminus i}$. This requires solving a sparse system of linear equations. The complexity of this depends on the sparsity pattern of $S_{\setminus i}^{-1}$. When the underlying graphical model is highly sparse, such as a nearest-neighbor pairwise-connected MRFs, each iteration can be performed in time $\mathcal{O}(D^{1.5})$ [2]. As the graph becomes more dense, run time approaches $\mathcal{O}(D^3)$. Thus, using this approach, EP requires $\mathcal{O}(K^3 + K^2 D + K D^{\rho(\mathcal{G})})$ for $1.5 \leq \rho(\mathcal{G}) \leq 3$ for each potential.

3. Whitened EP

For many problems of computer vision, both the number of variables D and the number of potentials N grow linearly with the number of pixels. In these cases, classical EP is $\mathcal{O}(D^3)$, since operations at each pixel require access to a $D \times D$ covariance matrix. The run time when storing the inverse covariance matrix [22] is at least $\mathcal{O}(D^{2.5})$. It is preferable for a computer vision algorithm to scale linearly with the number of pixels (as achieved, typically, by BP). One desirable property of EP, however, is that the run time is independent of the size of the cliques; only the rank of the potentials affects the run time. Low-rank potentials of large clique size have a wide array of promising applications in computer vision [17, 10]. One example is the use of strong spatial priors that can capture complex and long-range dependencies, such as Fields of Experts [20], or priors over higher-order derivatives [24]. Also, in a multi-scale setting, potentials at coarse scales require large cliques, but rank remains the same at any scale. Binocular occlusion [11], shadow, and unknown global properties all introduce large cliques that may be formulated with low rank. Difficulty in performing inference over large-clique potentials has limited the probabilistic models used in computer vision.

In this section, we propose an algorithm that achieves both of these goals: run time that is linear in the number of pixels *and* in clique size. To begin, observe that the bulk of the computational expense of EP is involved in computing the the covariance structure S. When S describes the covariance structure of an image, a high degree of regularity may be expected. The second order statistics of natural images [21] and range images [8, 18] are well studied; one of the most robust statistical trends of natural scenes is they tend to have power spectra that obeys a power law: $|\mathcal{F}[I]|^2 = \frac{A}{f^\beta}$ for constants A and β. Variations of EP designed for computer vision should make use of such regularities to reduce the computational demands of EP.

Expectation propagation can be made more efficient by limiting the forms of covariance structure expressible by S. In order for moment matching to correspond to minimizing KL-divergence, the approximating family $\tilde{P}$ must be an exponential family distribution (Eq. 2). Thus, S^{-1} must be parameterizable by $\vec{\theta}$ and expressible as $S^{-1} = \sum \theta_k B_k$ for some set of symmetric $D \times D$ matrices B_k. Our goal is to select B_k according to three principles. First, expressible covariance structure must include the covariance matrix for natural scenes. Additionally, since scene statistics are typically stationary, we prefer that local covariance structure achievable in one region of an image is also achievable in any region. Thus, for any B_k, there must exist some B_l that is equivalent to B_k after cyclic permutation by offset (d_x, d_y), for any such offset. This means that the number of B_k must be some multiple of D, and when $\theta_k = 1$ for all k, S^{-1} must be circulant. Finally, we seek to choose B_k that permit EP to run in linear time with respect to image size.

Let $\mathcal{S}$ denote the covariance matrix for natural scenes. Because images are stationary, $\mathcal{S}$ must form a circulant matrix, which means that its eigenvectors form a Fourier basis. Let $\mathcal{S} = \mathcal{F}\mathcal{A}\mathcal{F}'$, where $\mathcal{F}$ is a Fourier basis, and $\mathcal{A}$ is a diagonal matrix whose diagonal is the power spectrum of natural scenes. Also, let $\mathcal{W} = \mathcal{F}\mathcal{A}^{-\frac{1}{2}}\mathcal{F}'$. $\mathcal{W}$ is a whitening matrix for natural scenes: convolving an image with a column of $\mathcal{W}$ will, on average, remove covariance between pixels. One approach that satisfies our three goals is to constrain S to have the form $\mathcal{W}^{-1} D_S \mathcal{W}^{-1}$, where only D_S (a diagonal matrix) is permitted to vary. To perform EP efficiently with this constraint, it is preferable to maintain D_S and $\mathcal{W}\mu$ instead of S and μ. Under this transformation, the EP update equations for a potential i become:

$$D_{\setminus i} = (D_S^{-1} - D_i^{-1})^{-1} \quad (9)$$

$$\mathcal{W}\mu_{\setminus i} = D_{\setminus i}(D_S^{-1}\mathcal{W}\mu - D_i^{-1}\mathcal{W}\mu_i) \quad (10)$$

$$\begin{aligned} D_S = & D_{\setminus i} - D_{\setminus i} \operatorname{diag}[V_w'(I - (V_i S V_i')(V_w D_{\setminus i} V_w')^{-1}) \\ & (V_w D_{\setminus i} V_w')^{-1} V_w] D_{\setminus i} \end{aligned} \quad (11)$$

$$\mathcal{W}\mu = \mathcal{W}\mu_{\setminus i} + D_{\setminus i} V_w'(V_w D_{\setminus i} V_w')^{-1}(V_i \mu - V_i \mu_i) \quad (12)$$

$$D_i = (D_S^{-1} - D_{\setminus i}^{-1})^{-1} \quad (13)$$

$$\mathcal{W}\mu_i = D_i(D_S^{-1}\mathcal{W}\mu - D_{\setminus i}^{-1}\mathcal{W}\mu_{\setminus i}) \quad (14)$$

where $V_w = V_i \mathcal{W}^{-1}$, and diag denotes the operation of discarding off-diagonal elements. The largest bottleneck above is computing V_w. However, note that V_i is only non-zero in C columns, where C is the clique size of the potential. Also, $\mathcal{W}^{-1}$ corresponds to convolution by a spatially-limited dewhitening filter, which for natural scenes decays exponentially from its center [21]. Thus, $\mathcal{W}^{-1} V_i'$ can be computed in time linear in C. Moreover, in equations 11 and 12, D_S and $\mathcal{W}\mu$ only require an update in rows and columns where $\mathcal{W}^{-1} V_i'$ is non-zero. Similarly, D_i and $\mathcal{W}\mu_i$ are only non-zero at those locations. Thus, each update equation can be performed in time $\mathcal{O}(K^3 + K^2C)$, giving the whitened EP technique a total run time of $\mathcal{O}(NK^2C)$ per iteration.

Recall that when S is constrained to be diagonal, EP is equivalent to belief propagation [14]. Thus, the proposal that Gaussian EP might still work effectively if S^{-1} was constrained to $\mathcal{W}D_S^{-1}\mathcal{W}$ is equivalent to the proposal that BP might work effectively if messages were approximated by Gaussians *as long as the variables were whitened beforehand to reduce correlation*.

In order to achieve linear time EP with respect to image size, we are not limited solely to diagonal covariance structure in whitened image space. If we constrain S to have the form $\sum_i \mathcal{F}\mathcal{A}_i^{\frac{1}{2}}\mathcal{F}' D_i \mathcal{F}\mathcal{A}_i^{\frac{1}{2}}\mathcal{F}'$, then EP remains linear in the number of pixels as long as there exists some whitening transform W such that $\mathcal{F}\mathcal{A}_i^{\frac{1}{2}}\mathcal{F}'$ and $\mathcal{F}\mathcal{A}_i^{-\frac{1}{2}}\mathcal{F}'$ both correspond to convolution with spatially limited filters.

Note that there are alternative strategies to selecting $\mathcal{W}$ besides choosing the whitening filter of the prior $P(\vec{x})$. The covariance structure of the posterior distribution may differ from that of the prior. One alternative strategy would be to select $\mathcal{W}$ so that $\mathcal{W}\mathcal{W}$ was equal to the average covariance found using sparse EP (which estimates covariance). In the shape from shading application of section 4, we found empirically that the benefit of this approach was small.

4. Shape from Shading

Whitened EP permits inference over images in linear time with respect to both pixels and clique size. To achieve this, it constrains the approximating distribution to be Gaussian with a covariance matrix $\mathcal{W}D_S^{-1}\mathcal{W}$ for some diagonal D_S. In this section, we test whether performance remains competitive using this approach. In particular, we are interested in whether Gaussian message approximation will be effective when the potentials ϕ_i are highly non-Gaussian. One highly non-Gaussian problem in computer vision is shape from shading (SfS). The goal of SfS is to estimate 3D shape from a single image, under the assumption that albedo is uniform, lighting originates from a single point from a known direction, and the surface reflectance function is both uniform and known. If we define $p = \frac{\partial z}{\partial x}$ and $q = \frac{\partial z}{\partial y}$, then $i(x, y) = R(p, q)$, where i is the input image, and z is a range image corresponding to the true 3D shape.

In recent years, several methods have been developed that solve the classical SfS problem well as long as surface reflectance R is assumed to be Lambertian [19, 17, 6, 3, 7]. Still, humans are able to exploit shading cues under far more general scenarios. In order to compete with human performance, SfS algorithms may need to accommodate non-Lambertian reflectance (which is highly nonlinear and non-Gaussian), shadow cues (which are highly non-local and may produce large cliques), unknown lighting and albedo properties (which are global properties, possibly requiring fully connected potentials), and the integration of multiple depth cues. For this reason, more flexible optimization methods for SfS are desirable. Our hope is that whitened EP, by permitting efficient inference over large cliques, will enable new MRF models capable of tackling generalized depth inference problems. In this section, we demonstrate how whitened EP handles several of these issues.

MRF Data Likelihood In the past, MRF models for SfS have inferred surface normals rather than depth [17]. In such models, two variable nodes are used for each pixel: one for $p(x, y)$ and one for $q(x, y)$. For each pixel, one potential $\phi_R(p, q|i)$ enforces the surface normal to be consistent with the known pixel intensity $i(x, y)$. However, not all surface normal maps correspond to a valid surface z. A depth map z only exists if p and q are *integrable*, or exhibit zero curl, so that $\frac{\partial p}{\partial y} = \frac{\partial q}{\partial x}$. Methods that infer surface normals must include additional MRF potentials that encourage p and q to obey this relationship. Enforcing integrability is often the largest computational bottleneck of probabilistic inference because it requires a clique size of at least four variables [17]. Also, because the integrability constraint is usually not perfectly satisfied, error may result. Finally, once p and q are inferred, computing the surface z requires an additional post-processing operation which is sometimes costly. Past methods have used a sparse $D\times D$ matrix pseudoinverse [17].

Alternatively, inferring surface depth z directly avoids these problems. This has been difficult to do using belief propagation because it requires a clique size for ϕ_R of at least three. Belief propagation is exponential in clique size, and ϕ_R is not eligible for computational shortcuts such as the linear constraint node simplification. The SfS solution of [17] used p and q variables with 300 bins, and would thus sacrifice a 300-fold speed decrease to infer depth z directly.

In contrast, whitened EP can either infer surface normals or infer depth directly, and the two objectives require similar run times. To infer depth, whitened EP operates over a MRF whose variable nodes correspond to the whitened surface depth. Let $z_w(x, y) = \mathcal{W}z$ refer to the whitened surface depth, where $\mathcal{W}$ is the linear whitening transform. Surface derivatives p and q can be recovered from z_w by convolving with the derivatives of inverse whitening filter. Then, for each pixel (x, y), we can enforce that the surface normal at that point is consistent with the known pixel intensity $i(x, y)$ with the potential $\phi_R(v_p \cdot z,\ v_q \cdot z \,|\, i)$, where v_p and v_q are the derivatives of inverse whitening filter centered at point (x, y). The clique size of this potential is the size of the support of v_p and v_q, and the rank of the potential is two. Because whitened EP is linear in both clique size and rank, inference over this potential is efficient. Alternatively, if whitened EP is used to infer surface normals p and q, the clique size would be twice the support of the inverse whitened filter.

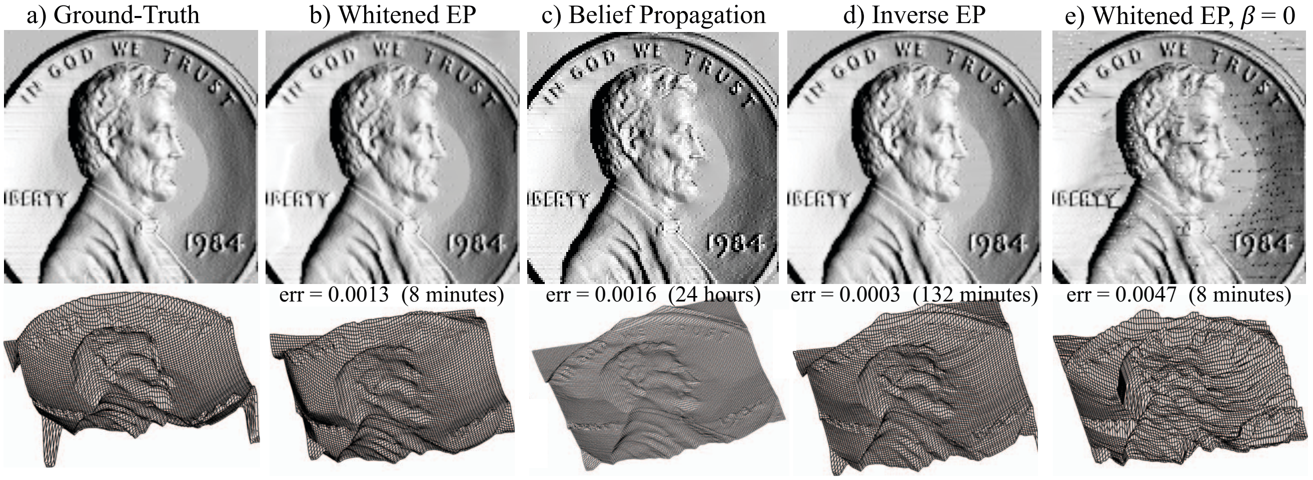

Figure 1. Results of the Whitened EP algorithm for Lambertian SfS, in comparison with other methods. Subfigure **c)** shows the results of linear constraint node BP [17]. Subfigure **d)** shows the results of EP using a full covariance matrix. Each of these methods is able to satisfy the Lambertian constraint accurately. Whitened EP is able to perform comparably in a substantially less time than other methods and with greater flexibility towards dense factor graphs or large cliques. Subfigure **e)** shows the results of diagonal EP without whitening.

In our experiments, we will use whitened EP to infer depth directly. We use a Laplace distribution for ϕ_R to penalize depth maps z that are not consistent with the known pixel intensity:

$$\phi_R(v_p \cdot z,\, v_q \cdot z|i) = \phi_R(p, q|i) = e^{-|R(p,q)-i)|/b} \tag{15}$$

where $R(p,q) = i$ is the reflectance map given by the known surface BRDF and lighting. Example ϕ_R are shown in the left column of figure 2.

MRF Spatial Prior The SfS problem is highly ambiguous: even when lighting direction and albedo are known, one image is consistent with large families of possible 3D surfaces which all render identically [6]. A strong spatial prior $P(z)$ is required to identify a 3D surface that is realistic out of all possibilities that are consistent with the image. Historically, MRFs over continuous variables have focused on pairwise-connected spatial priors (cliques of at most two) due to the high computational costs of large cliques. Methods that allow spatial priors with larger cliques have produced substantial performance gains [24, 17]. However, these methods are limited in the size and form of cliques achievable by the method.

Whitened EP provides two benefits for spatial priors. First, inference is linear in clique size, which could allow the use of large clique spatial priors such as Fields of Experts [20], which consists of 5×5 potentials of rank one. Additionally, as with any Gaussian EP method, any Gaussian potential requires no computational cost, regardless of rank or clique size. This is because for a Gaussian potential $\phi_i(\vec{x})$, in every iteration of EP the term approximation $G_i(\vec{x})$ is set equal to $\phi_i(\vec{x})$ [15]. This redundant operation can be avoided by initializing the posterior approximation $G(\vec{x})$ to the product of all Gaussian potentials.

In all following SfS experiments, we use a spatial prior that is implemented as a Gaussian with zero mean and covariance matrix equal to the covariance structure of natural range images $\mathcal{S}$. This prior has full rank and clique size D, making it impractical to implement using BP. Because the prior is Gaussian, it requires only $\mathcal{O}(1)$ in each iteration to implement using EP. In addition to its efficient run time, unifying many pairwise potentials into one large potential increases the fidelity of the Bethe approximation implicit in message passing algorithms [25]. Finally, this approach allows us to match the full covariance structure of natural scenes, including distant non-local covariances. As mentioned earlier, a pairwise-connected MRF produces a restricted covariance structure whose inverse matrix S^{-1} only contains elements along three unique diagonals [15]. One consequence of this restriction is that pairwise MRFs capture the statistics of first-order derivatives, but not of second or higher-order derivatives. In depth inference algorithms, this causes pairwise MRFs to overemphasize frontoparallel planes, and has been regarded as a chief limitation of the approach [4, 24]. Gaussian potentials permit EP to capture the second order statistics of all higher-order derivatives and any other linear feature. The disadvantage of this spatial prior is that higher-order moments of these features are not captured, such as the high kurtosis of derivatives. Empirically, we found that Gaussian priors significantly outperformed pairwise kurtotic priors when performing SfS using EP, possibly because SfS is typically performed on single surfaces rather than cluttered scenes. Note, though, that EP would remain efficient if both forms of priors were used simultaneously. Future research is needed to train MRF models that make use of both forms of spatial prior simultaneously.

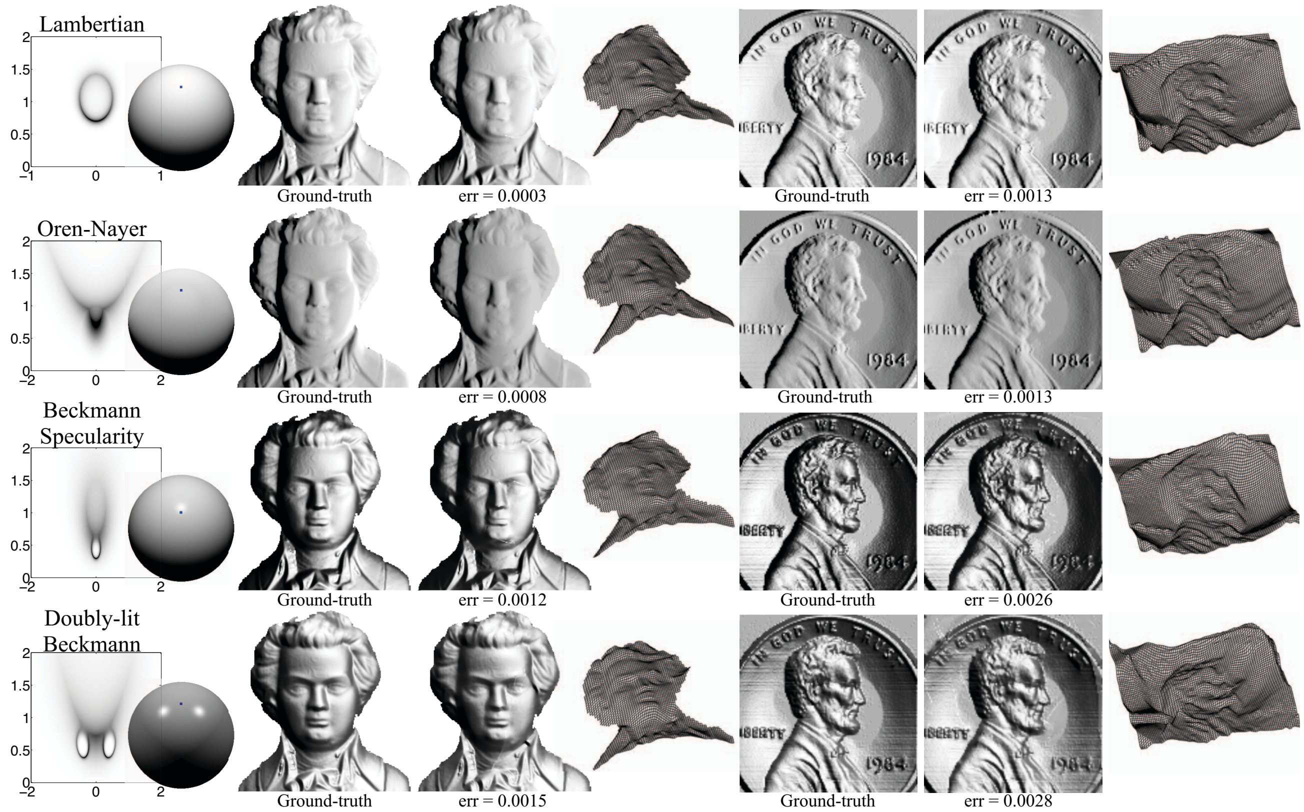

Figure 2. Results of whitened EP under several reflectances and lighting conditions. The left column shows example potential functions $\phi_R(p, q|i)$. In each case, potentials are highly non-Gaussian. The potential $\phi_R(p, q|i)$ differs at each pixel depending on intensity; here the chosen intensity is given by the blue dot on each sphere. For each reflectance, inferred surfaces are shown for benchmark SfS images. Mean-squared image error is reported for each case.

Lambertian SfS We first test our approach on Lambertian SfS, where it can be compared to past Lambertian SfS algorithms. The full MRF model in each experiment was

$$P(z) \propto \exp(-\frac{1}{2} z' \mathcal{S}^{-1} z) \prod_{x,y} \phi_R(p_{x,y}, q_{x,y} | i_{x,y}) \qquad (16)$$

where $\mathcal{S}$ is given by an $\frac{A}{f^\beta}$ power spectrum power law for range images. The power spectra of range images has been measured between $\beta = 2$ and $\beta = 2.88$, depending partly on the degree of occlusion present within the scene [8, 18] We chose the compromise value of $\beta = 2.5$ for all experiments; performance varied little for values of β between 2 and 3. The strength of the spatial prior, A, was fixed at 100.

In all experiments, whitened EP was run for 10 iterations, which is typically near to convergence. Convergence is not guaranteed for EP, but no divergence was encountered for this application. $V_i S V_i'$ was found numerically at each potential by sampling over a 26×26 discrete reflectance map (resembling fig. 2, left column). Importance sampling may have produced faster and more accurate results. For potentials with simpler forms, $V_i S V_i'$ may be found analytically by differentiating the log partition function.

Figure 1 shows the results of Whitened EP on a canonical benchmark image. For comparison, results are also shown for BP in 1c (reproduced from [17]). We also implemented classical EP with a sparse inverse covariance matrix; results are shown in 1d. The image error is reported below each result, which gives the mean squared error between the original image and the rerendered inferred surface; light intensities were chosen so that the reflectance map ranged from 0 to 1. The performance of all three methods is similar, and each method is able to infer a 3D surface that is closely consistent with the input image. In each case, improving the quality of the inferred surface is more likely to require an improvement to the MRF model (i.e. stronger spatial priors) than an improvement to the inference method. Among these methods, whitened EP is fastest and admits a wider class of MRF models. For the 128×128 penny image, whitened EP required 8 minutes on a 2.8GHz Xeon, and run-time grows linearly with the number of pixels, linearly with the clique size of the potentials, and linearly in their rank. The BP result required 24 hours, grows linearly with the number of pixels, quadratically in the clique size, and exponentially in rank. Sparse EP required 132 minutes, and grows more than

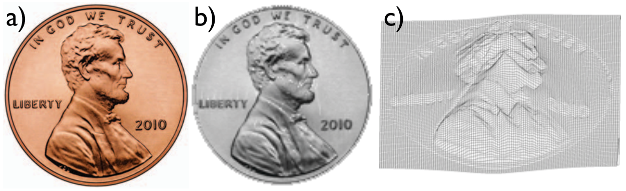

Figure 3. Example output for a natural image. **a)** A penny image was taken from Wikipedia. **b)** The output of the Whitened EP algorithm, rendered under the assumed lighting conditions. Mean squared image error was 0.0017. **c)** The inferred surface.

quadratically in the number of pixels. Other Lambertian SfS algorithms have reported image errors for the penny image of 0.0071 [9] and 0.0517 [13].

We also tested performance for $\mathcal{W} = I$, or equivalently, $\beta = 0$. This approach is equivalent to EP with diagonal covariance, or BP with Gaussian approximation of messages. Results are shown in figure 1e. This method struggled to identify a 3D surface that was consistent with the input image, even when pairwise priors were reintroduced to the model. This matches findings from other applications that if the posterior has correlated variables and S is constrained to be diagonal regardless, EP can produce poor results [5]. Our results suggests that the constrained covariance family used by whitened EP provided a sufficient approximation of the full covariance structure inferred by standard EP.

Non-Lambertian SfS Many of the leading SfS techniques rely explicitly on the exact form of the Lambertian equation, and do not generalize automatically to alternative reflectance types [19, 6, 7]. While there has been some success in applying methods such as Lax-Friedrichs and fast-marching to non-Lambertian reflectance [1, 23], these generalizations must proceed on a case-by-case basis for each class of reflectance functions. Past SfS methods based on BP were applicable to arbitrary reflectance functions, but relied on the simple form of the Lambertian equation to improve speed (24 hours), and was not tested on other reflectances [17].

One concern regarding Gaussian EP methods is that many computer vision applications require probability distributions that are highly non-Gaussian. SfS is one example, and non-Lambertian SfS produces especially non-Gaussian potentials. Example potentials ϕ_R are shown along the left column of figure 2, and are highly non-Gaussian. In spite of the non-Gaussianity of the posterior, whitened EP is able to infer 3D surfaces consistent with the input image under a wide variety of reflectances, including Oren-Nayar reflectance for rough surfaces, surfaces with Beckmann specularities, and surfaces lit from multiple lighting sources (see results in fig. 2). This is the first method we are aware of that demonstrates strong performance under arbitrary reflectance and lighting. Given the results in figure 2, we expect that in the future SfS will gain more from richer probabilistic models than by higher fidelity with the input image.

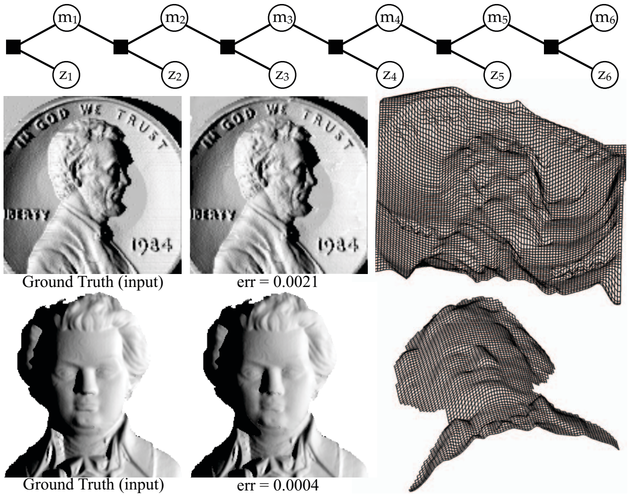

Figure 4. Shape from shading with cast shadows. At top is the factor graph that enforces all lit pixels to be unshadowed.

Figure 3 shows the algorithm performance on a natural image, taken from the "Penny" entry in Wikipedia. The lighting direction was manually estimated at $(0, 0.6, 1)$, and we used a Beckmann specularity with $m = 0.4$.

SfS with Cast Shadows Traditionally, past SfS methods have assumed that images are free of any cast shadows, due to the additional complexity this adds to the inference process. When cast shadows are present and lighting originates from a single point source, we must enforce two rules. First, pixels lying in shadow must be occluded from the lighting direction. Because we are inferring depth directly (as opposed to surface normals), this can be enforced simply by a pairwise potential of rank one. Suppose that lighting comes from the left, and suppose z_{unlit} is the depth of a shadowed pixel, and z_{lit} is the depth at the nearest unshadowed pixel to its left. Then $z_{lit} - z_{unlit} \geq \Delta x L_z / L_x$, where Δx is the horizontal distance between the two pixels, and L is the lighting vector. Computing $V_i S V_i'$ here does not require sampling, and can be found analytically [14].

Secondly, we must also enforce that pixels that are lit within the image are unshadowed in the inferred shape. This requires that $z_{left} - z_{lit} < \Delta x L_z / L_x$ for all pixels z_{left} to the left of z_{lit}. A straightforward implementation would require a pairwise potential between all pairs of pixels within the same row, and the run time would be quadratic in the number of pixels. Alternatively, we can combine all such constraints into a single potential. EP requires that we find the mean and variance of $G_{\setminus u}(z)\phi_u(z)$, where ϕ_u is the potential enforcing that all lit pixels are unoccluded from the light. We can solve for this mean and variance with a single pass of EP by noticing that ϕ_u can be decomposed into a tree-shaped factor graph. Intuitively, for each lit pixel

$z(x,y)$, we temporarily infer latent variables $m(x,y)$:

$$m(x,y) = \min_{x' \leq x}(z(x',y) + (x-x')L_z/L_x) \quad (17)$$

$$= \min(z(x,y),\, m(x-1,y) + L_z/L_x) \quad (18)$$

The factor graph for computing $G_{\setminus u}(z)\phi_u(z)$ is shown at the top of figure 4. The factor $\phi_L(z_i, m_i, m_{i-1})$ shown in this graph is an indicator function that is one only if equation 18 is satisfied, *and* $z_i < m_{i-1} + L_z/L_x$ (indicating that the pixel is unshadowed). Latent variables m_i can be discarded after the potential is updated.

Note that this approach to enforcing shadow cues would be expensive using BP because the potential ϕ_L is real-valued with a clique size of three, and is not eligible for LCN computational shortcuts. Traditional Gaussian EP becomes inefficient whenever shadow cues are enforced because non-local connectivity produces an inverse covariance matrix that is no longer sparse.

Example results for scenes with cast shadows are shown in figure 4. Importantly, the shadow constraint is satisfied completely by the inferred surface: all pixels that are lit in the input image are unshadowed in the inferred surface, and all black pixels in the input are shadowed in the output.

5. Conclusions

The methods in this paper reduce the run time of EP from cubic to linear in the number of pixels for visual inference, while retaining a run time that is linear in clique size. This is a substantial improvement over BP, which is exponential in clique size. The computational expense of inference for large cliques has prohibited the investigation of complex probabilistic models for vision. Our hope is that whitened EP will facilitate further research in these directions.

Results for whitened EP on SfS shows that the sacrifice in performance for this approach is small, even in problems with highly non-Gaussian potentials. Performance remained strong for surfaces with arbitrary reflectance and arbitrary lighting, which is a novel finding in SfS. We expect that efficient inference with large cliques will be especially beneficial for depth inference, where multi-scale representations, complex spatial priors, shadows, occlusions, and the simultaneous inference of unknown global scene attributes all necessitate potentials with large cliques.

References

[1] A. H. Ahmed and A. A. Farag. A new formulation for shape from shading for non-lambertian surfaces. In *CVPR*, 2006. 7

[2] N. Alon and R. Yuster. Solving linear systems through nested dissection. In *IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 225–234, 2010. 3

[3] J. T. Barron and J. Malik. High-frequency shape and albedo from shading using natural image statistics. In *CVPR*, 2011. 4

[4] P. N. Belhumeur. A Bayesian approach to binocular steropsis. *International Journal of Computer Vision*, 19(3):237–260, 1996. 5

[5] A. Birlutiu and T. Heskes. Expectation propagation for rating players in sports competitions. In *KDD*, 2007. 7

[6] A. Ecker and A. D. Jepson. Polynomial shape from shading. In *CVPR*, pages 145–152, June 2010. 4, 5, 7

[7] D. A. Forsyth. Variable-source shading analysis. *Int. J. of Computer Vision*, 91(3):280–302, Feb. 2011. 4, 7

[8] J. Huang, A. B. Lee, and D. Mumford. Statistics of range images. In *CVPR*, pages 1324–1331, 2000. 2, 3, 6

[9] N. Khan, L. Tram, and M. Tappen. Training many-parameter shape-from-shading models using a surface database. In *3DIM 2009 Workshop at ICCV*, pages 1433–1440, 2009. 7

[10] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. 1, 3

[11] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, 2002. 3

[12] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NIPS*, pages 109–117. 2011. 1

[13] K. Lee and C. Kuo. Shape from shading with a linear triangular element surface model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(8):815–822, 1993. 7

[14] T. P. Minka. Expectation propagation for approximate bayesian inference. In *UAI 2001*, page 362, 2001. 1, 2, 4, 7

[15] T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, January 2001. 1, 2, 5

[16] T. P. Minka and Y. Qi. Tree-structured approximations by expectation propagation. In *NIPS*, 2004. 2

[17] B. Potetz. Efficient belief propagation for vision using linear constraint nodes. In *CVPR*. 2007. 1, 3, 4, 5, 6, 7

[18] B. Potetz and T. S. Lee. Scaling laws in natural scenes and the inference of 3D shape. In *NIPS*. 2006. 3, 6

[19] E. Prados and O. Faugeras. Perspective shape from shading and viscosity solutions. In *ICCV*, page 826, 2003. 4, 7

[20] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *CVPR*, page 860, 2005. 2, 3, 5

[21] D. L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37:3385–3398, 1997. 3

[22] M. A. J. van Gerven, B. Cseke, F. P. de Lange, and T. Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50(1):150–161, 2010. 2, 3

[23] O. Vogel, M. Breuß, T. Leichtweis, and J. Weickert. Fast shape from shading for phong-type surfaces. In *Proceedings of the Second International Conference on Scale Space and Variational Methods in Computer Vision*, 2009. 7

[24] O. J. Woodford, P. H. S. Torr, I. D. Reid, and A. W. Fitzgibbon. Global stereo reconstruction under second order smoothness priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(12):2115–2128, 2009. 1, 3, 5

[25] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005. 5