Author's manuscript, published in *J. Syst. Sci. Complex.*, vol. 23, 2010.

BAYESIAN IMAGE SUPERRESOLUTION AND HIDDEN VARIABLE MODELING

Atsunori KANEMURA · Shin-ichi MAEDA · Wataru FUKUDA · Shin ISHII

Received: 30 June 2008 / Revised: 9 August 2008

Abstract Superresolution is an image processing technique that estimates an original high-resolution image from its low-resolution and degraded observations. In superresolution tasks, there have been problems regarding the computational cost for the estimation of high-dimensional variables. These problems are now being overcome by the recent development of fast computers and the development of powerful computational techniques such as variational Bayesian approximation. In this article, we review a Bayesian treatment of the superresolution problem and present its extensions based on hierarchical modeling by employing hidden variables.

Key words Bayesian estimation, hidden variables, image superresolution, Markov random fields, variational estimation.

1 Introduction

Suppose we have T images $\mathbf{y}_t$ $(t = 1, \ldots, T)$ of size $M_{\mathrm{O}} \times N_{\mathrm{O}} = P_{\mathrm{O}}$ that all show observations of the same scene. If the observed images contain different information as compared to each other, we can attempt to estimate the underlying high-resolution image $\mathbf{x}$ of magnified size $M_{\mathrm{H}} \times N_{\mathrm{H}} = P_{\mathrm{H}}$, where $M_{\mathrm{H}} \times N_{\mathrm{H}} = rM_{\mathrm{O}} \times rN_{\mathrm{O}}$, $P_{\mathrm{H}} = r^2 P_{\mathrm{O}}$, and r is the magnification factor (Fig. 1). We call this estimation problem *superresolution* [1–4].* A characteristic feature of the superresolution problem is that we need to estimate registration parameters that define the relative motions between images. These relative motions enable superresolution; without them, the information contained in the observations would not increase even if the number of observations increases. Let $\boldsymbol{\theta}_t$ denote the registration parameters for the tth observation. For the sake of convenience, we denote $\mathcal{D} = \{\mathbf{y}_t \mid t = 1, \ldots, T\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t \mid t = 1, \ldots, T\}$. From the viewpoint of Bayesian statistics, the estimation problem is translated into the computation of the posterior probability $p(\mathbf{x}|\mathcal{D}, \boldsymbol{\theta})$, which is derived by the Bayes theorem:

$$p(\mathbf{x}|\mathcal{D}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}) \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}, \boldsymbol{\theta}_t)}{\int p(\mathbf{x}) \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}, \boldsymbol{\theta}_t)\, d\mathbf{x}}, \tag{1}$$

where the prior probability $p(\mathbf{x})$ and the likelihood $p(\mathbf{y}_t|\mathbf{x}, \boldsymbol{\theta}_t)$ must be provided in advance. In this article, we will first review the Bayesian superresolution method proposed by Tipping

Atsunori KANEMURA · Shin-ichi MAEDA · Wataru FUKUDA · Shin ISHII
Graduate School of Informatics, Kyoto University, Kyoto 611-0011*, Japan.*
Email: {atsu-kan,ichi,fukuda}@sys.i.kyoto-u.ac.jp; ishii@i.kyoto-u.ac.jp.

*This type of superresolution is in particular called *multiframe* superresolution or *reconstruction-based* superresolution. There is another type of superresolution called *example-based* superresolution [5], which estimates a high-resolution image from only one image rather than from multiple images, based on a database developed in advance. Example-based methods constitute another large class but they are beyond the scope of this article.

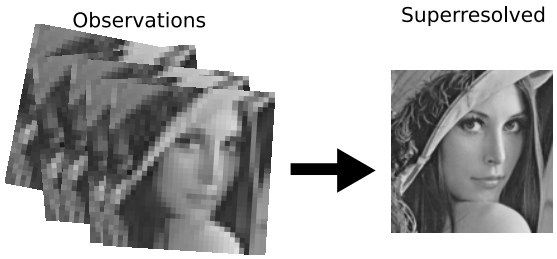


Figure 1 Superresolution is the problem of estimating a high-resolution image from multiple low-resolution, degraded observations of the same scene.

& Bishop [6] and then describe its extensions based on hierarchical modeling [7–9] with hidden variables.

The concept of superresolution, making use of multiple images to estimate a high-resolution image, was first proposed by Tsai & Huang [10]. They assumed a motion model restricted only to global translation and performed all computations in the Fourier domain. Other methods for multiframe superresolution include IBP (iterative back projection) methods [11], POCS (projection onto convex sets) methods [12], and probabilistic methods [13] using MRFs (Markov random fields) [14,15]. All the abovementioned methods can be considered to be an optimization problem of a certain cost function subject to certain constraints. We can obtain an appropriate cost function that has a natural interpretation reflecting our prior knowledge by utilizing the statistical estimation framework.

In Section 2, we review the mechanism of Bayesian superresolution and describe a basic Bayesian superresolution method employing single-layer Gaussian distributions. In Section 3, we introduce a compound Gaussian MRF having an edge layer in addition to the high-resolution image layer and show that the compound model is superior to the single-layer model. In Section 4, we present a hierarchical likelihood model in which hidden variables represent possible occlusions in observed images so that occlusion removal is successfully achieved. In Section 5, we conclude the article and discuss possible future directions.

2 Bayesian Superresolution

Tipping & Bishop [6] proposed a Bayesian treatment of the superresolution problem where *Bayesian marginalization* of hidden (unobservable) variables plays an important role. Bayesian superresolution has extended the joint MAP (maximum *a posteriori*) superresolution method proposed by Hardie *et al.* [16], who used the naive posterior probability (1) as the cost function both for the registration parameters and the high-resolution image.

2.1 Mechanism of Bayesian Superresolution

We begin by defining the prior and the likelihood. The prior $p(\mathbf{x})$ represents our *a priori* knowledge or expectation of the high-resolution image and it is often selected to be an MRF that imposes smoothness constraints on the image, reflecting our prior knowledge that neighboring pixels are likely to have similar values [2, 4]. The prior can be understood as a regularizer of the superresolution problem, which is definitely ill-posed due to the downscale and irreversible noise processes. On the other hand, the likelihood $p(\mathbf{y}_t|\mathbf{x}, \boldsymbol{\theta}_t)$ represents the observation process from the high-resolution image $\mathbf{x}$ to an observed image $\mathbf{y}_t$. The registration parameters $\boldsymbol{\theta}_t$ characterize the likelihood. The likelihood is a model of a physical process in the real world

and therefore there is less possibility for arguments as compared to the prior.

Since every quantity relevant to the estimation is computed from the prior and the likelihood, the performance of a superresolution algorithm strongly depends on the choice of prior and likelihood models. In the standard formulation, the prior and the likelihood are both selected as Gaussian distributions. As the prior, a simple Gaussian distribution can impose smoothness constraints over pixel values, and it is known that the Gaussian distribution is not appropriate for natural images because it overly smoothens edges and hence the estimated images may often be blurred. This disadvantage is not limited to the superresolution problem; in fact, it is shared by various image processing problems, and considerable efforts have been devoted to develop probability distributions that can preserve edges [2, 3, 16–18].

Bayesian superresolution estimates the registration parameters and the high-resolution image based on a prescribed prior and likelihood. Although the design of hierarchical models for the prior and likelihood using hidden variables is the main focus of this article, we first review the mechanism of the Bayesian superresolution method without specifying concrete prior and likelihood models. The registration parameters $\boldsymbol{\theta}$ are estimated by maximizing the marginalized likelihood:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}), \tag{2}$$

where the marginalized likelihood L is derived by marginalizing (integrating) out $\mathbf{x}$ from the joint distribution:

$$L(\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathcal{D}|\boldsymbol{\theta})\, d\mathbf{x} = \int p(\mathbf{x}) p(\mathcal{D}|\mathbf{x}, \boldsymbol{\theta})\, d\mathbf{x}. = \int p(\mathbf{x}) \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}, \boldsymbol{\theta}_t)\, d\mathbf{x}. \tag{3}$$

After obtaining the estimates $\hat{\boldsymbol{\theta}}$ of the registration parameters, the high-resolution image is estimated by the mean of the posterior distribution (1):

$$\hat{\mathbf{x}} = \mathbb{E}(\mathbf{x}) = \int \mathbf{x} p(\mathbf{x}|\mathcal{D}, \hat{\boldsymbol{\theta}})\, d\mathbf{x}. \tag{4}$$

Marginalization is the most important process in Bayesian superresolution; however, at the same time, it is also the main problem because the integration operation in (3) is not necessarily tractable for arbitrary prior $p(\mathbf{x})$ or likelihood $p(\mathbf{y}_t|\mathbf{x}, \boldsymbol{\theta}_t)$. Therefore, these distributions are restricted to those which the integration in (3) is tractable with. Existing Bayesian superresolution methods [6, 19] use single-layer Gaussian distributions because the marginalized likelihood can be analytically evaluated in such cases. Consequently, their estimation results fail to retain sharp edges and the estimated high-resolution images tend to be overly blurred. It is difficult to directly use edge-preserving prior distributions because they cannot be readily marginalized.

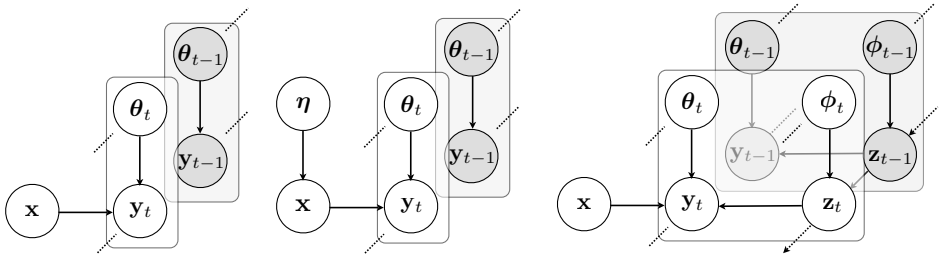
2.2 Single-Layer Bayesian Superresolution

In this subsection, the prior and likelihood models are specified to be single-layer Gaussian distributions, and the properties of the estimators under them are presented. A graphical model depicting the statistical dependency structure of the single-layer model is shown in Fig. 2(a).

The prior represents the smoothness constraints, and it is given by

$$p(\mathbf{x}) \propto \exp\Big\{-\frac{\rho}{2} \sum_{i \sim j} (x_i - x_j)^2\Big\}, \tag{5}$$

where ρ is a precision parameter that determines the strength of the prior belief and $i \sim j$ implies "pixels i and j are adjacent," and summation $\sum_{i \sim j}$ is taken over all pairs of neighboring



(a) Single-layer model (b) Compound prior model (c) Hierarchical likelihood model

Figure 2 Graphical models for the single-layer, compound, and hierarchical models.

pixels. Let $\mathcal{N}(i)$ be the set of the four immediate neighbors of a pixel i on the high-resolution plane, $\mathcal{N}(i) = \{j \mid j \sim i\}$. Since the exponent of (5) is always nonpositive and a quadratic function of $\mathbf{x}$, $p(\mathbf{x})$ becomes a Gaussian distribution. Since the distribution (5) only has a local dependency structure, $p(x_i|\mathbf{x}_{\backslash i}) = p(x_i|\mathbf{x}_{\mathcal{N}(i)})$ holds (we denote $\mathbf{x}_{\backslash i} = \{x_j \mid \forall j \neq i\}$ and $\mathbf{x}_{\mathcal{N}(i)} = \{x_j \mid j \in \mathcal{N}(i)\}$). Therefore, $p(\mathbf{x})$ is a Gaussian MRF and its potentials are given by $V_{ij}(x_i, x_j) = (x_i - x_j)^2$; therefore, the complete energy is $E(\mathbf{x}) = \sum_{i\sim j} V_{ij}(x_i, x_j)$. Note that the squared potential has a shape shown in Fig. 4(a), which has no robustness, i.e., estimation under this potential will result in overly smoothed images because it strongly penalizes abrupt changes (edges) in the image, although edges are very important for recognition by the human visual system [20]. The explicit form of $p(\mathbf{x})$ is given by the following single-layer Gaussian distribution[†]:

$$p(\mathbf{x}) = \mathrm{Gauss}(\mathbf{x}|\mathbf{0}, \rho^{-1}A^{-1}) = \frac{|A|^{1/2}}{(2\pi/\rho)^{P_\mathrm{H}/2}} \exp\Big\{-\frac{\rho}{2}\mathbf{x}^\mathrm{T}A\mathbf{x}\Big\}, \tag{6}$$

where A is a symmetric matrix that is derived as follows. Noting the fact $\sum_{i\sim j} = \frac{1}{2}\sum_{i=1}^{P_\mathrm{H}}\sum_{j\in\mathcal{N}(i)}$, we obtain

$$\sum_{i\sim j}(x_i - x_j)^2 = 2\sum_{i\sim j}(x_i^2 - x_i x_j) = \sum_{i=1}^{P_\mathrm{H}} x_i^2 \sum_{j\in\mathcal{N}(i)} 1 - \sum_{i=1}^{P_\mathrm{H}}\sum_{j\in\mathcal{N}(i)} x_i x_j \tag{7}$$

$$= \sum_{i=1}^{P_\mathrm{H}} x_i^2 A_{ii} + \sum_{i=1}^{P_\mathrm{H}}\sum_{j=1}^{P_\mathrm{H}} x_i x_j A_{ij} = \mathbf{x}^\mathrm{T}A\mathbf{x}, \tag{8}$$

where

$$A_{ij} = \begin{cases} |\mathcal{N}(i)| & (i = j) \\ -1 & (i \sim j) \\ 0 & (\text{otherwise}) \end{cases}. \tag{9}$$

This matrix works as a spatial difference filter. Generally, matrices used in the prior have been selected as high-pass filters such as the difference filter, the Laplacian filter, or filters with wider spatial ranges [4].

The likelihood is defined according to our assumption of the observation process. We assume that the original high-resolution image $\mathbf{x}$ is (i) geometrically transformed, (ii) blurred with a

[†]The symbol $\mathrm{Gauss}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denotes a Gaussian distribution of $\mathbf{x}$ whose mean is $\boldsymbol{\mu}$ and covariance is Σ, i.e., $\mathrm{Gauss}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = (|2\pi\Sigma|)^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathrm{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$.

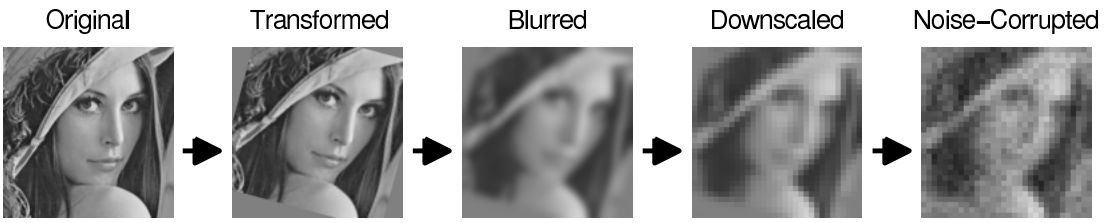


Figure 3 Observation process.

PSF (point spread function), (iii) downscaled, and (iv) corrupted by Gaussian noise (Fig. 3). This process can be represented by the following equation:

$$\mathbf{y}_t = W(\boldsymbol{\theta}_t)\mathbf{x} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \text{Gauss}(\mathbf{0}, \beta^{-1}I), \tag{10}$$

where $W(\boldsymbol{\theta}_t)$ is a non-square matrix responsible for (i)–(iii) and $\boldsymbol{\varepsilon}_t$ is Gaussian noise with uniform precision (inverse variance) β. The ij element of the observation matrix W is the contribution from the jth pixel of the high-resolution image to the ith pixel of the observed low-resolution image, and it is defined by

$$W_{ij}(\boldsymbol{\theta}_t) = \frac{1}{Z_i} \exp\left\{-\frac{d_{ij}^2(\boldsymbol{\theta}_t)}{2\gamma^2}\right\}, \tag{11}$$

where d_{ij} is the Euclidean distance between the low-resolution pixel i projected onto the high-resolution plane and the high-resolution pixel j, γ is the width (standard deviation) parameter of the Gaussian PSF, and Z_i is determined by constraint $\sum_j W_{ij} = 1$. If we allow translational and rotational motions, the distance between pixels i and j is given by

$$d_{ij}(\mathbf{s}_t, \psi_t) = \|R(\psi_t)(\mathbf{r}_i - \bar{\mathbf{r}}) + \mathbf{s}_t - \mathbf{r}_j\|, \tag{12}$$

where $\mathbf{s}_t$ is the amount of translational motion, $\mathbf{r}_i$ and $\mathbf{r}_j$ are position vectors of the pixels i and j on the high-resolution plane, respectively, $R(\psi_t)$ is the rotation matrix of ψ_t radian, and $\bar{\mathbf{r}}$ is the center of rotation. Under this motion model, the registration parameters are $\boldsymbol{\theta}_t = \{\mathbf{s}_t, \psi_t\}$. Although we use this three-dimensional motion model in the experiments, we can use a more general model such as projection transformation, which has eight-dimensional parameters [21]. We sometimes write $W_t = W(\boldsymbol{\theta}_t)$ for the sake of simplicity. The single-layer likelihood is given by

$$p(\mathbf{y}_t|\mathbf{x}, \boldsymbol{\theta}_t) = \text{Gauss}(\mathbf{y}_t|W(\boldsymbol{\theta}_t)\mathbf{x}, \beta^{-1}I) = \frac{1}{(2\pi/\beta)^{P_O/2}} \exp\left\{-\frac{\beta}{2}\|\mathbf{y}_t - W(\boldsymbol{\theta}_t)\mathbf{x}\|^2\right\}. \tag{13}$$

The EM (expectation-maximization) algorithm [22, 23] is used to find the registration parameters that maximize the marginal likelihood. Here, we present a general formulation of the EM algorithm, which will be revisited later when hidden variables are introduced to the model. Let $\boldsymbol{\tau}$ be hidden (unobservable) variables. In this section, $\boldsymbol{\tau} = \mathbf{x}$; however, in later sections, we expand $\boldsymbol{\tau}$ to include additional hidden variables. The EM algorithm can be formulated as a minimization procedure of the following variational free energy [23, 24]

$$F(q, \boldsymbol{\theta}) = -\int q(\boldsymbol{\tau}) \ln \frac{p(\boldsymbol{\tau}, \mathcal{D}|\boldsymbol{\theta})}{q(\boldsymbol{\tau})}\, d\boldsymbol{\tau} = -\left\langle \ln \frac{p(\boldsymbol{\tau}, \mathcal{D}|\boldsymbol{\theta})}{q(\boldsymbol{\tau})} \right\rangle_{\boldsymbol{\tau}}, \tag{14}$$

where q is an arbitrary probability distribution called a trial distribution, and the brackets $\langle\cdot\rangle_{\boldsymbol{\tau}}$ denote the expectation operator with respect to q. Subscripts such as $\langle\cdot\rangle_{\boldsymbol{\tau}}$ are omitted when there is no ambiguity. The free energy F is a functional of the function q and also a function of the parameters $\boldsymbol{\theta}$. Since $p(\boldsymbol{\tau}, \mathcal{D}|\boldsymbol{\theta}) = p(\boldsymbol{\tau}|\mathcal{D}, \boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})$, the free energy can be decomposed as

$$F(q, \boldsymbol{\theta}) = -\ln L(\boldsymbol{\theta}) + D_{\mathrm{KL}}(q(\boldsymbol{\tau})\|p(\boldsymbol{\tau}|\mathcal{D}, \boldsymbol{\theta})). \tag{15}$$

Here, L is the marginalized likelihood defined by (3), and D_{KL} is the KL (Kullback-Leibler) divergence between the trial distribution $q(\boldsymbol{\tau})$ and the true posterior $p(\boldsymbol{\tau}|\mathcal{D}, \boldsymbol{\theta})$ defined by

$$D_{\mathrm{KL}}(q(\boldsymbol{\tau})\|p(\boldsymbol{\tau}|\mathcal{D}, \boldsymbol{\theta})) = -\left\langle \ln \frac{p(\boldsymbol{\tau}|\mathcal{D}, \boldsymbol{\theta})}{q(\boldsymbol{\tau})} \right\rangle. \tag{16}$$

The KL divergence satisfies $D_{\mathrm{KL}}(q\|p) \geq 0$ for any q and p and $D_{\mathrm{KL}}(q\|p) = 0$ if and only if q and p are identical distributions [24, 25]. From (15), we see that minimizing F with respect to q is equivalent to minimizing D_{KL}. Therefore, the optimal trial distribution q^* that minimizes the free energy F coincides with the true posterior and then, $D_{\mathrm{KL}} = 0$. Furthermore, if we minimize F with respect to $\boldsymbol{\theta}$, it reduces to the minimization of $-L$, i.e., maximization of L, and we obtain the parameter estimates $\hat{\boldsymbol{\theta}}$ defined by (2). That is,

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \min_{q} F(q, \boldsymbol{\theta}). \tag{17}$$

In practice, we employ the coordinate descent optimization of $F(q, \boldsymbol{\theta})$; the variational optimization with respect to q is called the E step, and the optimization with respect to $\boldsymbol{\theta}$ the M step. Therefore, we iterate the following two steps until convergence is achieved:

$$\text{E step:} \quad q^* = \operatorname*{argmin}_{q} F(q, \boldsymbol{\theta}), \tag{18}$$

$$\text{M step:} \quad \boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} F(q, \boldsymbol{\theta}). \tag{19}$$

Under the single-layer Gaussian model specified by (6) and (13), the posterior distribution of the high-resolution image is computed according to (1), and we obtain

$$q^* = p(\mathbf{x}|\mathcal{D}, \boldsymbol{\theta}) = \mathrm{Gauss}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{P_{\mathrm{H}}/2}|\Sigma|^{1/2}} \exp\Big\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\Big\} \tag{20}$$

where

$$\Sigma = \Big(\rho A + \beta \sum_{t=1}^{T} W_t^{\mathrm{T}} W_t\Big)^{-1}, \tag{21}$$

$$\boldsymbol{\mu} = \beta\Sigma\Big(\sum_{t=1}^{T} W_t^{\mathrm{T}} \mathbf{y}_t\Big). \tag{22}$$

In the E step, the evaluation of the posterior distribution is reduced to the computation of its sufficient statistics $\boldsymbol{\mu}$ and Σ.

In the M step, the free energy is optimized with respect to $\boldsymbol{\theta}$. Omitting terms independent of $\boldsymbol{\theta}$, we obtain

$$F = -\left\langle \ln \frac{p(\mathbf{x}, \mathcal{D}|\boldsymbol{\theta})}{q(\mathbf{x})} \right\rangle = -\sum_{t=1}^{T} \langle \ln p(\mathbf{y}_t|\mathbf{x}, \boldsymbol{\theta}_t)\rangle_{\mathbf{x}} + \text{const.} \tag{23}$$

Since $-\ln p(\mathbf{y}_t|\mathbf{x},\boldsymbol{\theta}_t) = \frac{\beta}{2}\|\mathbf{y}_t - W(\boldsymbol{\theta}_t)\mathbf{x}\|^2 + \text{const.}$, the optimization problem is reduced to the minimization of the following expected squared error

$$\sum_{t=1}^{T}\langle\|\mathbf{y}_t - W(\boldsymbol{\theta}_t)\mathbf{x}\|^2\rangle = \sum_{t=1}^{T}\{\|\mathbf{y}_t - W(\boldsymbol{\theta}_t)\boldsymbol{\mu}\|^2 + \mathrm{tr}(\Sigma W(\boldsymbol{\theta}_t)^{\mathrm{T}}W(\boldsymbol{\theta}_t))\}, \tag{24}$$

where the first term on the right-hand side is the squared error between the observed images and the high-resolution image transformed by the observation matrices, and the second term represents the uncertainty regarding the high-resolution image through its posterior covariance. The second term is the advantage of marginalization over the MAP method [16], where the cost function only comprises the squared error term. Tipping & Bishop [6] have shown that due to the uncertainty term, accurate estimation of the parameters is achieved by avoiding overfitting.

After the EM algorithm converges, we have the estimates for the registration parameters. At the same time, an estimate $\hat{\mathbf{x}}$ for the high-resolution image is already computed in the E step as the mean value $\boldsymbol{\mu}$ of the posterior distribution (22). From (22), $\boldsymbol{\mu}$ is a linear transformation of the observed images and $\beta\Sigma W_t^{\mathrm{T}}$ is the inverse filtering kernel. This $\boldsymbol{\mu}$ also provides the maximum posterior probability:

$$\hat{\mathbf{x}} = \boldsymbol{\mu} = \operatorname*{argmax}_{\mathbf{x}} p(\mathbf{x}|\mathcal{D},\boldsymbol{\theta}) = \operatorname*{argmin}_{\mathbf{x}}\Big(\rho\|A^{1/2}\mathbf{x}\|^2 + \beta\sum_{t=1}^{T}\|\mathbf{y}_t - W_t\mathbf{x}\|^2\Big). \tag{25}$$

The first term is the norm of the high-frequency components of $\mathbf{x}$, and the second term is the squared error in the observation space. Therefore, the estimator $\hat{\mathbf{x}}$ is the result of regularized least squares estimation with employing the high-frequency components of $\mathbf{x}$ as the regularizer.

3 Edge-Preserving Superresolution by Introducing Hidden Variables to Prior

In this section, we introduce a hierarchical model to a prior, which is called a compound prior, by employing hidden variables representing edges in the high-resolution image; we show the advantages of this model in high-resolution image estimation. The likelihood is the same as that of the Gaussian distribution described in the previous section. A graphical model for the compound prior model is shown in Fig. 2(b). The introduction of the edge variables makes exact estimation difficult and thus we utilize the variational EM algorithm to derive a computationally efficient estimation procedure.

3.1 Compound Prior

We introduce binary hidden variables $\eta_{ij} \in \{0,1\}$ representing edges between neighboring pixels i and j, and denote in total $\boldsymbol{\eta} = \{\eta_{ij} \mid i \sim j\}$. The concept of placing binary variables between neighboring pixels is the same as the line process proposed by Geman & Geman [26]. The marginalized prior of the high-resolution image $\mathbf{x}$ is

$$p(\mathbf{x}) = \sum_{\boldsymbol{\eta}} p(\boldsymbol{\eta},\mathbf{x}) = \sum_{\boldsymbol{\eta}} p(\boldsymbol{\eta})p(\mathbf{x}|\boldsymbol{\eta}), \tag{26}$$

which is a mixture distribution. It is suggested that the uncertainty regarding edge positions is considered in a soft way via marginalization with respect to $\boldsymbol{\eta}$, rather than hard switching between $\eta_{ij} = 0$ and 1 by making a point estimate of $\boldsymbol{\eta}$. Note, however, that the computational complexity of $\sum_{\boldsymbol{\eta}}$ increases exponentially with the number of edge variables.

The joint prior of the edges $\boldsymbol{\eta}$ and the high-resolution image $\mathbf{x}$ is defined as the Boltzmann distribution:

$$p(\boldsymbol{\eta}, \mathbf{x}) = \frac{1}{Z} \exp\Big\{-\frac{\rho}{2} E(\boldsymbol{\eta}, \mathbf{x})\Big\}. \tag{27}$$

The energy function is defined as[‡]

$$E(\boldsymbol{\eta}, \mathbf{x}) = \sum_{i \sim j} \big(\eta_{ij}(x_i - x_j)^2 + (1 - \eta_{ij})\lambda\big), \tag{28}$$

where λ is a constant. The lower is the energy, the higher is the probability. When $\eta_{ij} = 1$, the squared error $(x_i - x_j)^2$ between neighboring pixels i and j is left, and when $\eta_{ij} = 0$, the constant λ remains. Therefore, the prior probability of $\mathbf{x}$ imposes smoothness constraints between neighboring pixel values x_i and x_j when $\eta_{ij} = 1$, and there is no smoothing when $\eta_{ij} = 0$. For a fixed $\mathbf{x}$ but variable $\boldsymbol{\eta}$, we observe that if $(x_i - x_j)^2 > \lambda$, $\eta_{ij} = 0$ is chosen and if $(x_i - x_j)^2 < \lambda$, $\eta_{ij} = 1$ gives a higher probability. This implies that λ is the threshold for judging the presence of edge. This distribution also possesses a local dependency structure, i.e., it is an MRF, and this type of compound model is called compound Gaussian MRF [15,27,28]. The local potential function of the compound MRF is $V_{ij}(\eta_{ij}, x_i, x_j) = \eta_{ij}(x_i - x_j)^2 + (1 - \eta_{ij})\lambda$, which is called Hampel's loss function in robust statistics [15], and its shape is shown in Fig. 4(b). Prior (27) can be equivalently represented as

$$p(\boldsymbol{\eta}, \mathbf{x}) = \mathrm{Ber}(\boldsymbol{\eta}|\nu)\, \mathrm{Gauss}(\mathbf{x}|\mathbf{0}, (\rho A_{\boldsymbol{\eta}})^{-1}). \tag{29}$$

Here, $\mathrm{Ber}(\boldsymbol{\eta}|\nu)$ is the Bernoulli distribution:

$$\mathrm{Ber}(\boldsymbol{\eta}|\nu) = \prod_{i \sim j} \nu^{\eta_{ij}} (1 - \nu)^{1 - \eta_{ij}}, \tag{30}$$

where $\nu = \mathrm{sig}(\lambda\rho/2) \triangleq 1/(1 + \exp\{-\lambda\rho/2\})$, and the precision matrix is

$$[A_{\boldsymbol{\eta}}]_{ij} = \begin{cases} \sum_{k \in \mathcal{N}(i)} \eta_{ik} & (i = j) \\ -\eta_{ij} & (i \sim j) \\ 0 & (\text{otherwise}) \end{cases}. \tag{31}$$

As compared to matrix A defined by (9), $A_{\boldsymbol{\eta}}$ is dependent on $\boldsymbol{\eta}$ and therefore the strength of smoothing is controlled by the edge configuration $\boldsymbol{\eta}$. Note that if we assume $\boldsymbol{\eta} = \mathbf{1}$ ($\eta_{ij} = 1$ for all $i \sim j$), this model would coincide with the classical single-layer Gaussian model (5). Therefore, the compound model is a generalization of the single-layer model.

The exact posterior distribution under the compound prior is a Gaussian mixture

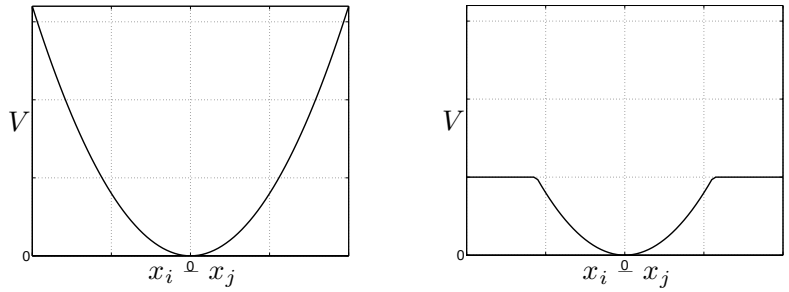
$$p(\mathbf{x}|\mathcal{D}, \boldsymbol{\theta}) = \sum_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\mathcal{D}) p(\mathbf{x}|\boldsymbol{\eta}, \mathcal{D}, \boldsymbol{\theta}) = \sum_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\mathcal{D})\, \mathrm{Gauss}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\eta}}, \Sigma_{\boldsymbol{\eta}}), \tag{32}$$

where the parameters for the conditional posterior are

$$\Sigma_{\boldsymbol{\eta}} = \Big(\rho A_{\boldsymbol{\eta}} + \beta \sum_{t=1}^{T} W_t^{\mathrm{T}} W_t\Big)^{-1}, \tag{33}$$

$$\boldsymbol{\mu}_{\boldsymbol{\eta}} = \beta \Sigma_{\boldsymbol{\eta}} \Big(\sum_{t=1}^{T} W_t^{\mathrm{T}} \mathbf{y}_t\Big). \tag{34}$$

[‡]When $\boldsymbol{\eta} = \mathbf{0}$, the distribution becomes improper with respect to $\mathbf{x}$; however, this problem can be avoided by adding $\varepsilon\|\mathbf{x}\|^2$ ($\varepsilon > 0$) to E and setting ε to a sufficiently small value.



(a) Squared potential. (b) Hampel's robust potential.

Figure 4 Potential functions.

Thus, the exact estimate for the high-resolution image is given by

$$\hat{\mathbf{x}}_{\mathrm{ExactC}} = \sum_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\mathcal{D})\boldsymbol{\mu}_{\boldsymbol{\eta}}. \tag{35}$$

This is a weighted average of high-resolution image estimates under every possible edge configuration $\boldsymbol{\eta}$ whose weights are given by its posterior probability $p(\boldsymbol{\eta}|\mathcal{D})$. This computation is difficult in practice because the sum $\sum_{\boldsymbol{\eta}}$ has a computational complexity that grows exponentially with the number of pixels. Moreover, since the marginalized likelihood L also has an exponential complexity, it is difficult to obtain exact estimates $\hat{\boldsymbol{\theta}}$ for the parameters. This is an example of the difficulty in marginalization regarding Bayesian superresolution. Then, we use the variational approximation method described in the following subsection.

3.2 Variational EM Estimation

In this subsection, we derive a variational EM algorithm for the compound model for efficient Bayesian superresolution.

In Section 2.2, we presented the general formulation of the EM algorithm under hidden variables $\boldsymbol{\tau}$. When we use the compound model, the hidden variables are $\boldsymbol{\tau} = \{\boldsymbol{\eta}, \mathbf{x}\}$. In the previous subsection, we saw that the optimal trial distribution that minimizes the free energy is the true posterior distribution; however, it is computationally intractable for the compound model. This intractability arises because the optimal q is searched in the space of arbitrary distributions. Then, we restrict the functional form of q to be the following factorized distribution

$$q(\boldsymbol{\eta}, \mathbf{x}) = \prod_{i \sim j} q(\eta_{ij}) q(\mathbf{x}), \tag{36}$$

and optimize $F(q, \boldsymbol{\theta})$ with respect to q and $\boldsymbol{\theta}$. Since q is optimized in the restricted space, the exact relation of (17) no longer holds. However, F can be regarded as an upper bound of $-L$ and the variational EM algorithm can be understood as a bound optimization method.

In the E step, each factor of the trial distribution is variationally optimized. It is known that the optimal factor can be analytically derived if the other factors are fixed [24, 29], and the optimal factors are

$$\ln q^*(\eta_{ij}) = \langle \ln p(\boldsymbol{\eta}, \mathbf{x}, \mathcal{D}|\boldsymbol{\theta}) \rangle_{\boldsymbol{\eta}_{\setminus ij}, \mathbf{x}} + \text{const.}, \tag{37}$$

$$\ln q^*(\mathbf{x}) = \langle \ln p(\boldsymbol{\eta}, \mathbf{x}, \mathcal{D}|\boldsymbol{\theta}) \rangle_{\boldsymbol{\eta}} + \text{const.}, \tag{38}$$

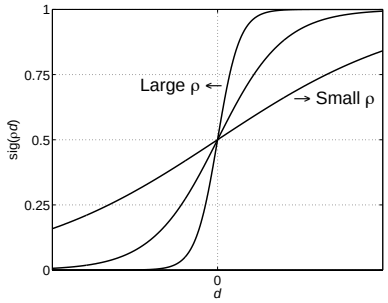


Figure 5 Sigmoid function with various gradients.

where the bracket pairs denote the expectations with respect to q, that is, $\langle\cdot\rangle_{\boldsymbol{\eta}_{\setminus ij},\mathbf{x}}$ and $\langle\cdot\rangle_{\boldsymbol{\eta}}$ denote the expectations with respect to $q(\boldsymbol{\eta}_{\setminus ij},\mathbf{x})$ and $q(\boldsymbol{\eta})$, respectively. The notation $\boldsymbol{\eta}_{\setminus ij}$ is a set of variables $\boldsymbol{\eta}$ except for η_{ij}. Here, we ignore the term $\langle\ln|A_{\boldsymbol{\eta}}|\rangle$ that appears when rearranging the right-hand side of (37) since it is empirically known that the effect of this term is small. The optimal trial distribution can be computed by iterating

$$q^*(\boldsymbol{\eta}) = \mathrm{Ber}(\boldsymbol{\eta}|\bar{\boldsymbol{\nu}}) = \prod_{ij} \bar{\nu}_{ij}^{\eta_{ij}}(1-\bar{\nu}_{ij})^{1-\eta_{ij}}, \tag{39}$$

$$q^*(\mathbf{x}) = \mathrm{Gauss}(\mathbf{x}|\boldsymbol{\mu}_{\mathrm{C}},\Sigma_{\mathrm{C}}) = \frac{1}{(2\pi)^{P_{\mathrm{H}}/2}|\Sigma_{\mathrm{C}}|^{1/2}} \exp\Big\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{C}})^{\mathrm{T}}\Sigma_{\mathrm{C}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{C}})\Big\}, \tag{40}$$

where

$$\bar{\nu}_{ij} = \mathrm{sig}\Big(\frac{\rho}{2}(\lambda - \langle(x_i-x_j)^2\rangle)\Big) \triangleq \frac{1}{1+\exp\Big\{-\frac{\rho}{2}(\lambda-\langle(x_i-x_j)^2\rangle)\Big\}}, \tag{41}$$

$$\Sigma_{\mathrm{C}} = \Big(\rho\langle A_{\boldsymbol{\eta}}\rangle + \beta\sum_{t=1}^{T} W_t^{\mathrm{T}}W_t\Big)^{-1}, \tag{42}$$

$$\boldsymbol{\mu}_{\mathrm{C}} = \beta\Sigma_{\mathrm{C}}\Big(\sum_{t=1}^{T} W_t^{\mathrm{T}}\mathbf{y}_t\Big). \tag{43}$$

The parameter $\bar{\nu}_{ij}$ is the expectation of the edge variable η_{ij}. Equation (41) suggests that soft identification of the edge existence is achieved, instead of the classical hard switching between 0 and 1; this is done by using the sigmoid function. As we change the precision parameter ρ of the prior, the gradient of the sigmoid function changes and therefore the edge sensitivity can be controlled (Fig. 5). The covariance Σ_{C} of $\mathbf{x}$ contains the expected matrix $\langle A_{\boldsymbol{\eta}}\rangle$ whose elements are given by the edge expectations $\bar{\boldsymbol{\nu}}$. The mean $\boldsymbol{\mu}_{\mathrm{C}}$ is again a linear transformation of the observed images; however, in this case, the inverse kernel $\beta\Sigma_{\mathrm{C}}W_t^{\mathrm{T}}$ incorporates the estimated edge pattern $\bar{\boldsymbol{\nu}}$. According to the edge estimate, the strength of the regularization varies spatially such that the edges are less regularized whereas non-edge regions are more regularized, so that edge preservation is achieved.

In the M step, we optimize F with respect to $\boldsymbol{\theta}$. We find that the terms dependent on $\boldsymbol{\theta}$ are

$$\sum_{t=1}^{T}\langle\|\mathbf{y}_t - W(\boldsymbol{\theta}_t)\mathbf{x}\|^2\rangle_{\mathbf{x}} = \sum_{t=1}^{T}\{\|\mathbf{y}_t - W(\boldsymbol{\theta}_t)\boldsymbol{\mu}_{\mathrm{C}}\|^2 + \mathrm{tr}(\Sigma_{\mathrm{C}}W(\boldsymbol{\theta}_t)^{\mathrm{T}}W(\boldsymbol{\theta}_t))\}, \tag{44}$$

Figure 6 Images used in the experiments: Cameraman, Lenna, Girl, Beads, and License.

Table 1 Mean ISNRs with standard deviations in 12 experiments for five images.

	CGMRF [dB]	SGMRF [dB]
Cameraman	$\mathbf{4.88} \pm 0.14$	$4.48 \pm \mathbf{0.05}$
Lenna	$\mathbf{10.04} \pm 0.23$	$9.52 \pm \mathbf{0.17}$
Girl	$\mathbf{7.75} \pm 0.07$	$7.42 \pm \mathbf{0.06}$
Beads	$\mathbf{10.11 \pm 0.23}$	9.58 ± 0.34
License	$\mathbf{7.80} \pm 0.30$	$7.35 \pm \mathbf{0.09}$
Total	$\mathbf{8.11} \pm 1.93$	$7.67 \pm \mathbf{1.88}$

which is the expected squared error; this is the same as (24), except that it employs different Σ_{C} and $\boldsymbol{\mu}_{\mathrm{C}}$.

The estimate for the high-resolution image is given by the mean $\boldsymbol{\mu}_{\mathrm{C}}$ of the trial distribution:

$$\hat{\mathbf{x}}_{\mathrm{VarC}} = \boldsymbol{\mu}_{\mathrm{C}}. \tag{45}$$

This variational estimate $\hat{\mathbf{x}}_{\mathrm{VarC}}$ is the regularized least square solution with the regularization matrix being $\langle A_{\boldsymbol{\eta}} \rangle$ (see (42)). The expected matrix $\langle A_{\boldsymbol{\eta}} \rangle$ takes care of the edge probability for each pixel and the strength of smoothing is controlled by it. In contrast, the exact estimate $\hat{\mathbf{x}}_{\mathrm{ExactC}}$ is a weighted average of the conditional means $\boldsymbol{\mu}_{\boldsymbol{\eta}}$, which are regularized least square solutions whose regularization matrices are $A_{\boldsymbol{\eta}}$ (see (33)). Therefore, the difference between the variational estimate and the exact estimate arises from the manner in which the expectation is obtained, and then, replacing $A_{\boldsymbol{\eta}}$ with $\langle A_{\boldsymbol{\eta}} \rangle$ is effective in reducing the complexity. On the other hand, the difference between $\hat{\mathbf{x}}_{\mathrm{VarC}}$ and $\hat{\mathbf{x}}$ is that the regularization is space *variant*. That is, in the single-layer model, the high-pass filtering matrix A is fixed at every pixel; in contrast, in the variational estimation of the compound model, $\langle A_{\boldsymbol{\eta}} \rangle$ is different for each pixel and thus it realizes a pixel-wise smoothing effect.

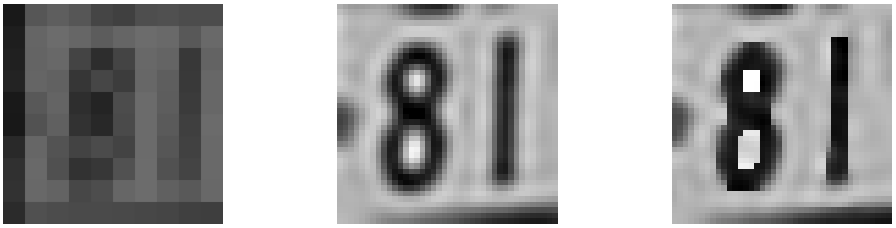
3.3 Experiments

We conducted experiments to compare the single-layer Gaussian MRF model (SGMRF) and the compound Gaussian MRF model (CGMRF). The reconstruction errors were measured based on the PSNR (peak signal-to-noise ratio), defined by

$$PSNR(\hat{\mathbf{x}}) = 10 \log_{10} \frac{M^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|^2 / P_{\mathrm{H}}} \quad [\mathrm{dB}], \tag{46}$$

where M is the maximum pixel value. The higher is the PSNR, the better is the estimate.

The five images shown in Fig. 6 were used as original images. We generated $T = 16$ observed images by synthetically applying geometrical transformation, blurring, downscaling, and noise corruption. The geometrical transformations consisted of translational and rotational motions, and the amounts of translation were drawn from a uniform distribution between -2 and 2 pixels and the angles of rotation were drawn from a uniform distribution between $-4\pi/180$



(a) Observed image (1/16) (b) SGMRF's estimate (c) CGMRF's estimate

(d) CGMRF: Horizontal edge (e) CGMRF: Vertical edge

Figure 7 Comparison of superresolution of License by SGMRF and our CGMRF.

and $4\pi/180$ radian. A Gaussian PSF with a width of 2 was used as the blurring kernel. The decimation factor was $r = 4$. Finally, the SNR of the noise was set at 30 dB. Such a generation procedure was repeated 12 times for each original image, forming 60 sets of $T = 16$ observation images in total. The pixel values were represented as floating point numbers normalized within $[0, 1]$.

The following are the specifications of the variational EM algorithm: initial values for all the parameters were set at 0. The scaled conjugate gradient method [30] was used to optimize $\boldsymbol{\theta}$. The algorithm was terminated at the lth iteration if the following conditions were satisfied: $F^l - F^{l-1} < 10^{-4}$, $\|\boldsymbol{\mu}^l - \boldsymbol{\mu}^{l-1}\|/\|\boldsymbol{\mu}^{l-1}\| < 10^{-4}$, and $\|\boldsymbol{\theta}^l - \boldsymbol{\theta}^{l-1}\|/\|\boldsymbol{\theta}^{l-1}\| < 10^{-4}$, where the superscripts indicate the iteration step. The following hyperparameters were used: $\lambda = 0.025$, $\rho = 30$, and $\beta = 4500$.

Since the absolute values of PSNR are different for images, we used the PSNR of the mean image of the observations as the baseline and the improvement from it, ISNR (improvement in SNR), was computed for the estimated high-resolution images. Table 1 shows the mean ISNRs with standard deviation for high-resolution images estimated by the single-layer model (SGMRF) and the compound model (CGMRF). We observe that the compound method was superior for all the images, and there was an average improvement of more than 0.4 dB. The estimation results for the License image are shown in Fig. 7. We observe that the extraction of edges (d), (e) is successfully done, and the image estimation of the compound model (c) is definitely sharper than that of the single-layer model (b).

4 Superresolution with Occlusion Removal by Introducing Hidden Variables for Likelihood

In this section, we describe a Bayesian superresolution method [8,9] that assumes occlusions

in the observations. The single-layer Gaussian MRF is used as the prior; however, hidden variables indicating occlusions in observations are introduced to the likelihood, making the likelihood a hierarchical model. A graphical model under the hierarchical likelihood is shown in Fig. 2(c). The exact computation under this model is again computationally intractable and we use the variational EM algorithm to derive an efficient superresolution algorithm.

4.1 Hierarchical Likelihood

For an observed image $\mathbf{y}_t$, we introduce binary random variables $\mathbf{z}_t$ that indicate the existence of occlusions. Since we cannot directly observe $\mathbf{z}_t$, they are hidden variables. Whether or not occlusion is present in the ith pixel of the tth observed image is indicated by $z_{ti} \in \{+1, -1\}$. If $z_{ti} = -1$, the pixel is occluded, whereas if $z_{ti} = +1$, the pixel is intact. We make the following assumptions:

1) We regard occluded pixels as a region with large observation noise.

2) Obstacles have a spatial continuity.

3) Obstacles move along time.

We write $\mathbf{z} = \{\mathbf{z}_t \mid t = 1, \ldots, T\}$. The marginal likelihood with the introduction of the occlusion pattern $\mathbf{z}$ is given by

$$p(\mathcal{D}|\mathbf{x}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z}|\boldsymbol{\phi}) p(\mathcal{D}|\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}), \tag{47}$$

where $\boldsymbol{\phi}$ denotes the amount of occlusion movement and defines the occlusion prior, and the sum $\sum_{\mathbf{z}}$ is taken over all possible occlusion patterns.

The conditional likelihood is given by

$$p(\mathbf{y}_t|\mathbf{x}, \mathbf{z}_t, \boldsymbol{\theta}_t) = \mathrm{Gauss}(\mathbf{y}|W(\boldsymbol{\theta}_t)\mathbf{x}, B(\mathbf{z}_t)^{-1}), \tag{48}$$

where the precision matrix B is a diagonal matrix $B(\mathbf{z}_t) = \mathrm{diag}(\beta_1(z_1), \ldots, \beta_{P_\mathrm{O}}(z_{tP_\mathrm{O}}))$ whose elements are dependent on the occlusion pattern $\mathbf{z}_t$. Reflecting Assumption 1), we set the precision (inverse variance) of the observation noise to

$$\beta_i(z_{ti}) = \begin{cases} \beta_\mathrm{H} & (z_{ti} = +1) \\ \beta_\mathrm{L} & (z_{ti} = -1) \end{cases}. \tag{49}$$

Here, we take $\beta_\mathrm{H} > \beta_\mathrm{L}$ to satisfy the condition that when $z_{ti} = +1$, there is no occlusion (low noise = high precision), whereas when $z_{ti} = -1$, there is occlusion (low precision).

Assumptions 2) and 3) are represented as the hierarchical prior for the occlusion patterns $\mathbf{z}$. Assumption 3) is fulfilled by the following Markov property:

$$p(\mathbf{z}|\boldsymbol{\phi}) = p(\mathbf{z}_1) p(\mathbf{z}_2|\mathbf{z}_1, \boldsymbol{\phi}_1) \cdots p(\mathbf{z}_T|\mathbf{z}_{T-1}, \boldsymbol{\phi}_{T-1}), \tag{50}$$

where the moving parameters for $\mathbf{z}_t$ are $\boldsymbol{\phi}_t$. Each prior distribution is given by the Boltzmann distribution:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}, \boldsymbol{\phi}_{t-1}) = \frac{1}{Z} \exp\{-E(\mathbf{z}_t, \mathbf{z}_{t-1})\}, \tag{51}$$

where the energy function is defined as

$$E(\mathbf{z}_t, \mathbf{z}_{t-1}) = -J_\mathrm{self} \sum_{i=1}^{P_\mathrm{O}} z_{ti} - J_\mathrm{inter} \sum_{i \sim j} z_{ti} z_{tj} - J_\mathrm{move} \mathbf{z}_t^\mathrm{T} \tilde{\mathbf{z}}_t. \tag{52}$$

Here, $J_\bullet$ are scalar constants, and $\tilde{\mathbf{z}}_t$ is a predicted occlusion pattern at time t that is obtained by moving the previous pattern $\mathbf{z}_{t-1}$ by a transition matrix $G(\boldsymbol{\phi}_{t-1})$, where $\boldsymbol{\phi}_{t-1}$ denotes the amounts of movement. As the matrix G, we use the bilinear interpolator to cope with subpixel amounts of motion. According to these prior settings, (51) and (52), an occlusion pattern $\mathbf{z}_t$ with a lower energy occurs with a higher probability. Here, we observe the effect of each term of (52). The self-connection coefficient J_{self} of the first term represents the strength of the bias; if $J_{\text{self}} > 0$, z_{ti} is likely to be $+1$, and if $J_{\text{self}} < 0$, z_{ti} is likely to be -1. The interconnection coefficient J_{inter} of the second term defines the degree of correlation within the single occlusion pattern. When $J_{\text{inter}} > 0$, $\mathbf{z}_t$ is likely to take the same values within neighboring pixels, whereas when $J_{\text{inter}} < 0$, the values tend to be different. The third term measures the similarity between $\mathbf{z}_t$ and $\tilde{\mathbf{z}}_t$, and the coefficient J_{move} represents how close the occlusion pattern $\mathbf{z}_t$ is to the predicted pattern $\tilde{\mathbf{z}}_t$, which is determined by the previous pattern $\mathbf{z}_{t-1}$ and the moving amounts $\boldsymbol{\phi}_{t-1}$. When modeling occlusion patterns, we set $J_{\text{self}} > 0$ to represent that the area of reliable (unoccluded) regions is generally larger than that of unreliable (occluded) regions. From Assumptions 2) and 3), we set $J_{\text{inter}} > 0$ and $J_{\text{move}} > 0$ so as to represent spatial and temporal continuities.

With the hierarchical likelihood, the posterior distribution for the high-resolution image becomes

$$p(\mathbf{x}|\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{\sum_{\mathbf{z}} p(\mathbf{z}|\boldsymbol{\phi}) p(\mathbf{x}) p(\mathcal{D}|\mathbf{x}, \boldsymbol{\theta}, \mathbf{z})}{\int \sum_{\mathbf{z}} p(\mathbf{z}|\boldsymbol{\phi}) p(\mathbf{x}) p(\mathcal{D}|\mathbf{x}, \boldsymbol{\theta}, \mathbf{z})\, d\mathbf{x}}. \tag{53}$$

The summation $\sum_{\mathbf{z}}$ is intractable because it should be taken over all configurations of binary patterns $\mathbf{z}$, whose number increases exponentially with the number of pixels. Thus, we utilize the variational EM algorithm to obtain a computationally efficient computational procedure.

4.2 Variational EM Estimation

The true posterior distribution $p(\mathbf{z}, \mathbf{x}|\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is approximated by the trial distribution $q(\mathbf{z}, \mathbf{x})$ for the hidden variables, $\boldsymbol{\tau} = \{\mathbf{x}, \mathbf{z}\}$. The E step is used to minimize the free energy F with respect to q and the M step is used to minimize F with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Since the unconstrained optimization of q implies the intractable computation of the true posterior distribution, we introduce the following factorization assumption to the trial distribution $q(\mathbf{z}, \mathbf{x})$:

$$q(\mathbf{z}, \mathbf{x}) = \prod_{t=1}^{T} \prod_{i=1}^{P_{\text{O}}} q(z_{ti}) q(\mathbf{x}). \tag{54}$$

The optimal trial distribution that minimizes the free energy is searched by iterating the factor-wise optimal solutions

$$q^*(z_{ti}) = \text{Ber}(z_{ti}|\nu_{ti}) = \nu_{ti}^{\frac{1}{2}(1+z_{ti})} (1-\nu_{ti})^{\frac{1}{2}(1-z_{ti})}, \tag{55}$$

$$q^*(\mathbf{x}) = \text{Gauss}(\mathbf{x}|\boldsymbol{\mu}_{\text{H}}, \Sigma_{\text{H}}) = \frac{1}{(2\pi)^{P_{\text{H}}/2} |\Sigma_{\text{H}}|^{1/2}} \exp\Big\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{\text{H}})^{\text{T}} \Sigma_{\text{H}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\text{H}}) \Big\}. \tag{56}$$

The occlusion patterns are identified in a soft way since the parameter ν_{ti} is the probability of occlusion absence $q(z_{ti} = +1)$, which is calculated using the sigmoid function

$$\nu_{ti} = \text{sig}(2\lambda_{ti}) \triangleq \frac{1}{1 + \exp\{-2\lambda_{ti}\}}, \tag{57}$$

where

$$\lambda_{ti} = J_{\mathrm{self}} + J_{\mathrm{inter}} \sum_{j \in \mathcal{N}(i)} \langle z_{tj} \rangle + J_{\mathrm{move}} [G(\boldsymbol{\phi}_{t-1})(2\boldsymbol{\nu}_{t-1} - 1) + G(\boldsymbol{\phi}_t)^{\mathrm{T}}(2\boldsymbol{\nu}_{t+1} - 1)]_i + \frac{1}{4}\left(\ln \frac{\beta_{\mathrm{H}}}{\beta_{\mathrm{L}}} - (\beta_{\mathrm{H}} - \beta_{\mathrm{L}}) \langle e_{ti}^2 \rangle \right). \tag{58}$$

The first through third terms come from the prior distribution, and each of them respectively arises from the self-connection, inter-connection, and occlusion movements. In the second term, the neighboring occlusion probabilities are summed and thus the spatial continuity is enhanced, as required in Assumption 2. The third term consists of the previous pattern advanced by $G(\boldsymbol{\phi}_{t-1})$ and the pattern at the next step moved back by $G(\boldsymbol{\phi}_t)^{\mathrm{T}}$, which accommodates Assumption 3. The fourth term is determined by the observations such that the baseline $\ln(\beta_{\mathrm{H}}/\beta_{\mathrm{L}})$ is compared with the following expected squared error at the ith pixel of the ith observed image:

$$\langle e_{ti}^2 \rangle = \langle (y_{ti} - \mathbf{w}_{ti}^{\mathrm{T}} \mathbf{x})^2 \rangle \tag{59}$$

$$= (y_{ti} - \mathbf{w}_{ti}^{\mathrm{T}} \boldsymbol{\mu}_{\mathrm{H}})^2 + \mathbf{w}_{ti}^{\mathrm{T}} \Sigma_{\mathrm{H}} \mathbf{w}_{ti}, \tag{60}$$

where $\mathbf{w}_{ti}$ is the ith row of W_t. The first term in (60) is the reconstruction error on the ith pixel of the tth image, and the second term is the degree of uncertainty. In practice, we ignore the second term of (60) because the uncertainty is relatively small, and it is empirically known that good results can be obtained even by ignoring the second term [8, 9]. Therefore, if the reconstruction error at a pixel is large, then that pixel is considered to be occluded.

The parameters for $q^*(\mathbf{x})$ are

$$\Sigma_{\mathrm{H}} = \Big(\rho A + \sum_{t=1}^{T} W_t^{\mathrm{T}} \langle B(\mathbf{z}_t) \rangle W_t \Big)^{-1}, \tag{61}$$

$$\boldsymbol{\mu}_{\mathrm{H}} = \Sigma_{\mathrm{H}} \Big(\sum_{t=1}^{T} W_t^{\mathrm{T}} \langle B(\mathbf{z}_t) \rangle \mathbf{y}_t \Big), \tag{62}$$

where $\langle B(\mathbf{z}_t) \rangle$ is a diagonal matrix whose elements are given by the expectations

$$\langle \beta_i(z_{ti}) \rangle = q(z_{ti} = +1)\beta_{\mathrm{H}} + q(z_{ti} = -1)\beta_{\mathrm{L}}. \tag{63}$$

We observe that under the hierarchical likelihood model, the inverse kernel that transforms the observations into the high-resolution image is $\Sigma_{\mathrm{H}} W_t^{\mathrm{T}} \langle B(\mathbf{z}_t) \rangle$. In this kernel, the occlusion probabilities are considered pixel-wise for each observed image so that the importance of occluded pixels become small, whereas the weights of observations in unoccluded regions are larger.

The optimization of the free energy F with respect to $\boldsymbol{\theta}$ is essentially the same as that in the previous sections and the expectation of the weighted squared error

$$\sum_{t=1}^{T} \langle \| B^{1/2}(\mathbf{z}_t)(\mathbf{y}_t - W(\boldsymbol{\theta}_t)\mathbf{x}) \|^2 \rangle = \sum_{t=1}^{T} \{ \| \langle B^{1/2}(\mathbf{z}_t) \rangle (\mathbf{y}_t - W(\boldsymbol{\theta}_t)\boldsymbol{\mu}_{\mathrm{H}}) \|^2 + \mathrm{tr}(\Sigma_{\mathrm{H}} W(\boldsymbol{\theta}_t)^{\mathrm{T}} \langle B(\mathbf{z}_t) \rangle W(\boldsymbol{\theta}_t)) \} \tag{64}$$

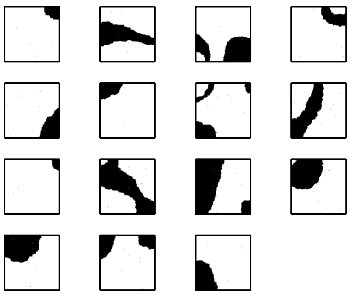


Figure 8 True binary patterns representing occlusions (black: 10 dB noise, white: 40 dB noise).

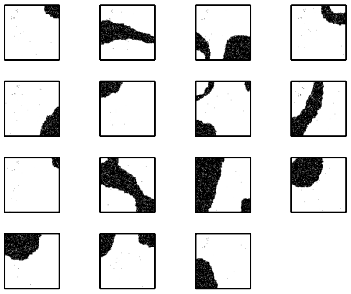


Figure 9 Estimated occlusion patterns. The colors are in grayscale so that $q(z_{ti} = +1) = 0$ corresponds to black (occluded) and $q(z_{ti} = +1) = 1$ to white (not occluded).

is minimized. When minimizing the free energy F with respect to $\boldsymbol{\phi}$, only the following terms related to the expected correlations are dependent on $\boldsymbol{\phi}$ and they should be considered:

$$-\sum_{t=2}^{T}\langle \mathbf{z}_t^{\mathrm{T}}\tilde{\mathbf{z}}_t\rangle = -\sum_{t=2}^{T}\langle \mathbf{z}_t\rangle^{\mathrm{T}} G(\boldsymbol{\phi}_{t-1})\langle \mathbf{z}_{t-1}\rangle. \tag{65}$$

In practice, we restricted the movement to be a uniform motion of fixed $\boldsymbol{\phi}_t$.

The estimate for the high-resolution image is given by $\hat{\mathbf{x}}_{\mathrm{VarH}} = \boldsymbol{\mu}_{\mathrm{H}}$. The difference as compared to the previous sections is that the noise precision matrix $\langle B(\mathbf{z}_t)\rangle$ has different element values by taking the expectation with respect to the posterior probability of the occlusion variables. Due to this modification, the relative strength of regularization becomes different among pixels so as to reflect the possible presence of occlusion. That is, in pixels with low precision, pixel values are strongly smoothed by relatively strong regularization, whereas in pixels with high precision, the observed data are more trusted by relatively weak regularization.

4.3 Experiments

We conducted experiments to observe the effects of the hierarchical likelihood with occlusion patterns. The Lenna image was used as an original image, and two datasets were synthetically generated. The procedure used for generating the datasets was almost the same as that described in Section 3.3 except that the noise variance was controlled according to the occlusion patterns $\mathbf{z}_t$. In the first dataset, $\mathbf{z}_t$ were generated independently without considering the movement. In other words, each $\mathbf{z}_t$ was generated by Gibbs sampling from $p(\mathbf{z})$ with parameters $J_{\mathrm{self}} = 0.01$, $J_{\mathrm{inter}} = 1$, and $J_{\mathrm{move}} = 0$ (Fig. 8). The SNR of the noise was set to 10 dB at

pixels where $z_{ti} = -1$, and 40 dB where $z_{ti} = +1$. In the second dataset, we generated $\mathbf{z}_t$ by shifting a single occlusion pattern, as shown in Fig. 12(a). The amount of the shift motion was an integer common for all the frames, making a uniform motion. The noise strength was 10 dB for pixels with $z_{ti} = -1$ and 45 dB for pixels with $z_{ti} = +1$. We assumed that the following parameters are known: noise precisions β_{H} and β_{L}; prior parameters J_{self}, J_{inter}, and J_{move}; and registration parameters $\boldsymbol{\theta}$. These parameters were assumed for the sake of simplicity, although, in theory, they can be estimated from data.

The purpose of the first experiment using the first dataset is to observe the difference between the single-layer model and the hierarchical model without considering occlusion movements. The estimated high-resolution images with magnification factor $r = 4$ are shown in Fig. 10. The close-up views of the region around the right eye are shown in Fig. 11. The estimation by the hierarchical likelihood model, shown at the top-right panel of Fig. 10, exhibits the highest (best) PSNR of 32.45 dB. The hierarchical model effectively changed the strength of regularization on each pixel based on the estimated occlusion patterns shown in Fig. 9. The single-layer model assuming the uniform low precision (10 dB noise) over all pixels estimated an overly smooth image (bottom-left), whose PSNR was 27.80 dB, and the single-layer model's estimation assuming uniform 40 dB noise, i.e., high precision over all pixels, completely failed to reconstruct the high-resolution image (bottom right). The highest PSNR attained by the single-layer model was 29.51 dB when 21 dB uniform noise was assumed; however, this is still approximately 3 dB worse than the estimation by the hierarchical model.

Using the second dataset, we compared the hierarchical model that assumes no movement ($J_{\mathrm{move}} = 0$) and the hierarchical model that estimates movement ($J_{\mathrm{move}} \neq 0$). When estimating the movement of occlusions, we assumed a uniform motion, i.e., $\boldsymbol{\phi}_t$ was common for all the observations, which is the same as the true movement used for generating the dataset. Nelder and Mead's simplex method [31] was used to minimize the cost function (65). Fig. 12 shows the true noise pattern and its estimation results. Due to the assumption of uniform movement, it is sufficient to show the single occlusion pattern common for all the observations. Fig. 13 shows the high-resolution images for a magnification factor of $r = 4$. We observe that the estimation considering the movement yielded an improvement of approximately 3 dB in this case.

We applied the hierarchical algorithm to a real image sequence. For the sake of simplicity, we fixed the camera when capturing the images so that there was no need for estimating the registration parameters. Since there is no relative motion within the observations, enhancing the resolution is beyond our objectives here and thus we set the magnification factor as $r = 1$. The estimated image is shown in Fig. 14 along with one of the observed images, observation mean, and observation median. The observation mean (c) includes ghosts and is very poor. The observation median (d) has no ghost but is blurred as compared to the estimation by the hierarchical method (b). This is because the hierarchical method performed deconvolution owing to the generative model that includes the blurring process.

5 Conclusion

In this article, we reviewed the Bayesian superresolution method and presented its hierarchical extensions by introducing hidden variables into the prior and the likelihood to construct appropriate hierarchical models. Hidden variables represents edges in the prior and occlusion patterns in the likelihood. The most significant benefit of hierarchical modeling is the local adaptability according to the estimation of hidden variables. In other words, in the compound prior, the strength of smoothing is controlled according to the edge probability dependent on each pixel pair; if an edge is identified, the sharpness is retained, whereas if no edge is esti-

Figure 10 Comparison of estimation results when the observed images suffer from the noise patterns shown in Fig. 8.

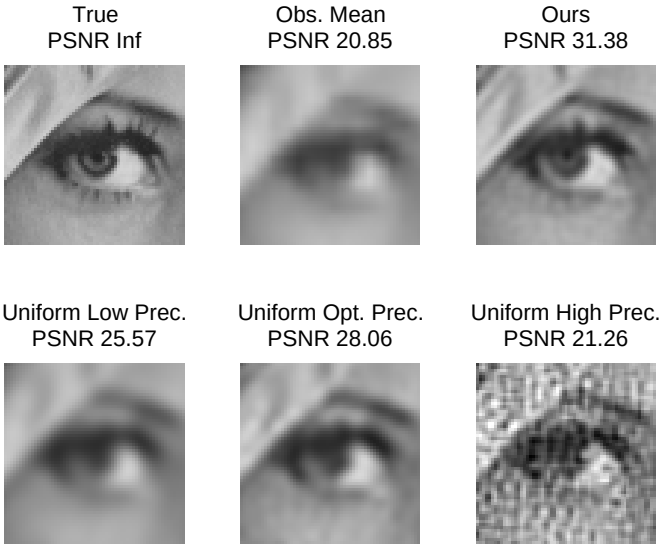
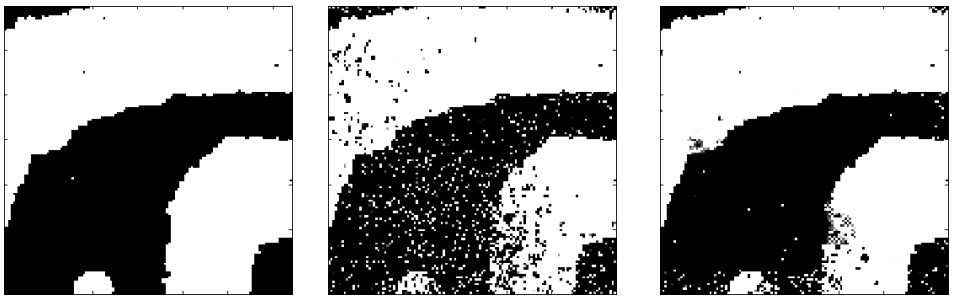


Figure 11 Close-up views of Fig. 10. PSNRs are re-calculated for the shown regions.



(a) True (b) Without move (c) With move

Figure 12 True and estimated occlusion patterns.

(a) Observation mean: PSNR 23.04 (b) Without move: PSNR 28.30 (c) With move: PSNR 31.25

Figure 13 Estimated images with/without estimating occlusions' movement.

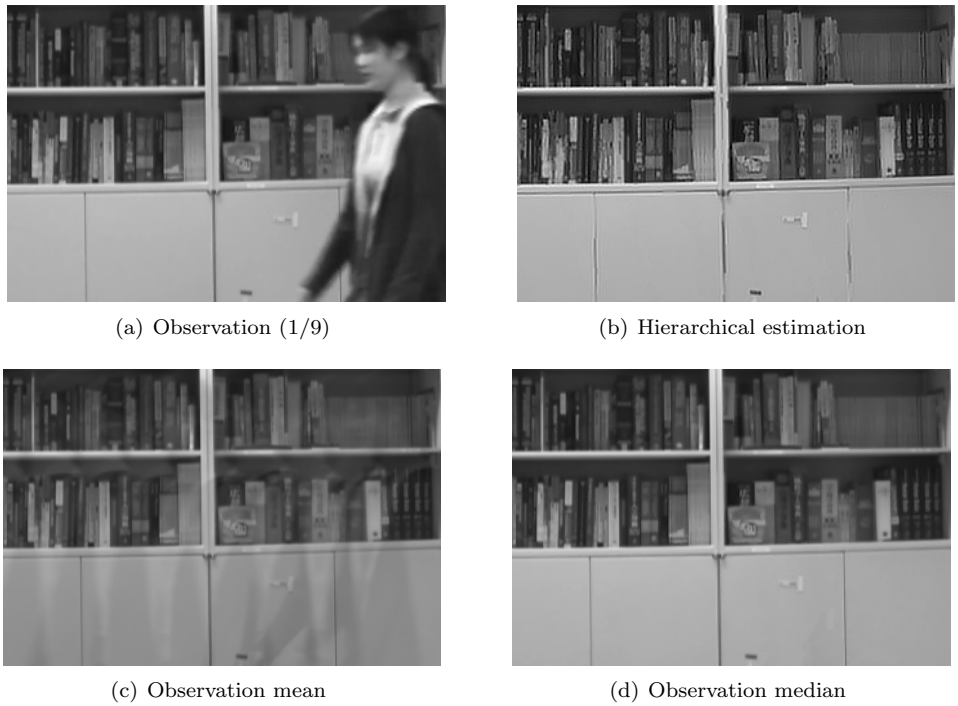
mated, the neighboring pixels are smoothed. In the hierarchical likelihood, each observation pixel is judged if it is occluded or not, and high-precision pixels are effectively used to estimate high-resolution images, whereas low-precision pixels are fickled out.

The key proposal of the Bayesian superresolution method reviewed in this article is the marginalization of the unknown high-resolution image; however, it is also possible to marginalize the registration parameters [21]. In this case, the selection of the prior for the registration parameters would be the most important point to capture the temporal dynamics.

Another way to improve the superresolution methods is to introduce further hierarchical priors on the hidden variables. The hierarchical priors for $\boldsymbol{\eta}$ and $\mathbf{z}$ used in this study were so simple that there was a weak clustering effect. Assuming the spatial regularity of the hidden variables would result in better estimation with more accurately identified edge or occlusion probabilities. We can avoid determining the hyperparameters manually by further introducing hierarchical priors and estimating them by means of Bayesian estimation.

References

[1] S. Borman and R. L. Stevenson, Spatial resolution enhancement of low-resolution image sequences: A comprehensive review with directions for future research, Technical report, Dept. of Electrical Enginerring, University of Notre Dame, 1998.



(a) Observation (1/9) (b) Hierarchical estimation

(c) Observation mean (d) Observation median

Figure 14 Estimation results using a real image sequence.

[2] S. C. Park, M. K. Park, and M. G. Kang, Super-resolution image reconstruction: A technical overview, *IEEE Signal Process. Mag.*, 2003, **20**(3): 21–36.

[3] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, Advances and challenges in super-resolution, *Int. J. Imag. Syst. Tech.*, 2004, **14**(2): 47–57.

[4] A. K. Katsaggelos, R. Molina, and J. Mateos, *Super Resolution of Images and Video*, Morgan & Claypool, San Rafael, CA, 2007.

[5] W. T. Freeman, T. R. Jones, and E. C. Pasztor, Example-based super-resolution, *IEEE Comput. Graphics Appl.*, 2002, **22**(2): 56–65.

[6] M. E. Tipping and C. M. Bishop, Bayesian image super-resolution, in *Advances in Neural Information Processing Systems (NIPS) 15*, (eds. by S. Becker, S. Thrun, and K. Obermayer), MIT Press, Cambridge, MA, 2003, 1279–1286.

[7] A. Kanemura, S. Maeda, and S. Ishii, Edge-preserving Bayesian image superresolution based on compound Markov random fields, in *Proc. International Conference on Artificial Neural Networks (ICANN)*, (ed. by J. Marques de Sá), LNCS 4669, Springer, 2007, II-611–620.

[8] A. Kanemura, S. Maeda, and S. Ishii, Image superresolution under spatially structured noise, *Proc. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2007, 279–284.

[9] W. Fukuda, S. Maeda, A. Kanemura, and S. Ishii, Bayesian image superresolution under moving occlusion (in Japanese), in *IEICE Technical Report*, 2008, **107**(542): 237–242.

[10] R. Y. Tsai and T. S. Huang, Multiframe image restoration and registration, in *Advances in Computer Vision and Image Processing*, JAI Press, Greenwich, CT, 1984, **1**: 317–339.

[11] M. Irani and S. Peleg, Improving resolution by image registration, *CVGIP: Graph. Model. Im.*, 1991, **53**(3): 231–239.

[12] H. Stark and P. Oskoui, High resolution image recovery from image-plane arrays, using convex projections, *J. Opt. Soc. Am. A*, 1989, **6**: 1715–1726.

[13] R. R. Schultz and R. L. Stevenson, Extraction of high-resolution frames from video sequences, *IEEE Trans. Image Process.*, 1996, **5**(6): 996–1011.

[14] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer, Tokyo, 2001.

[15] G. Winkler, *Image Analysis, Random Fields, and Markov Chain Monte Carlo Methods*, Springer, Heidelberg, 2nd ed., 2003.

[16] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, Joint MAP registration and high-resolution image estimation using a sequence of undersampled images, *IEEE Trans. Image Process.*, 1997, **6**(12): 1621–1633.

[17] C. Bouman and K. Sauer, A generalized Gaussian image model for edge-preserving MAP estimation, *IEEE Trans. Image Process.*, 1993, **2**(3): 296–310.

[18] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, Fast and robust multiframe super resolution, *IEEE Trans. Image Process.*, 2004, **13**(10): 1327–1344.

[19] N. A. Woods, N. P. Galatsanos, and A. K. Katsaggelos, Stochastic methods for joint registration, restoration, and interpolation of multiple undersampled images, *IEEE Trans. Image Process.*, 2006, **15**(1): 201–213.

[20] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*, Academic Press, New York, 2nd ed., 1976.

[21] L. C. Pickup, D. P. Capel, S. J. Roberts, and A. Zisserman, Bayesian methods for image super-resolution, *Comput. J.*, 2007, bxm091, Advance Access.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc. B*, 1977, **39**(1): 1–38.

[23] R. M. Neal and G. E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in *Learning in Graphical Models*, (ed. by M. I. Jordan), Kluwer Academic Press, Dordrecht, 1998, 355–368.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[25] S. Kullback, *Information Theory and Statistics*, John Wiley and Sons, New York, 1959, (Reprinted by Dover, 1997).

[26] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1984, **PAMI-6**(6): 721–741.

[27] F.-C. Jeng and J. W. Woods, Compound Gauss-Markov random fields for image estimation, *IEEE Trans. Signal Process.*, 1991, **39**(3): 683–697.

[28] F.-C. Jeng and J. W. Woods, Simulated annealing in compound Gaussian random fields, *IEEE Trans. Inf. Theory*, 1990, **36**(1): 94–107.

[29] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, An introduction to variational methods for graphical models, *Mach. Learn.*, November 1999, **37**(2): 183–233.

[30] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York, 1995.

[31] J. A. Nelder and R. Mead, A simplex method for function minimization, *Comput. J.*, 1965, **7**: 308–313.