Understanding Blind Deconvolution Algorithms

Anat Levin, Yair Weiss, Fredo Durand, and William T. Freeman, *Fellow*, *IEEE*

Abstract—Blind deconvolution is the recovery of a sharp version of a blurred image when the blur kernel is unknown. Recent algorithms have afforded dramatic progress, yet many aspects of the problem remain challenging and hard to understand. The goal of this paper is to analyze and evaluate recent blind deconvolution algorithms both theoretically and experimentally. We explain the previously reported failure of the naive MAP approach by demonstrating that it mostly favors no-blur explanations. We show that, using reasonable image priors, a naive simulations MAP estimation of both latent image and blur kernel is guaranteed to fail even with infinitely large images sampled from the prior. On the other hand, we show that since the kernel size is often smaller than the image size, a MAP estimation of the kernel alone is well constrained and is guaranteed to succeed to recover the true blur. The plethora of recent deconvolution techniques makes an experimental evaluation on ground-truth data important. As a first step toward this experimental evaluation, we have collected blur data with ground truth and compared recent algorithms under equal settings. Additionally, our data demonstrate that the shift-invariant blur assumption made by most algorithms is often violated.

Index Terms—Blind deconvolution, motion deblurring, natrual image statistics, statistical estimation.

1 Introduction

Blind deconvolution is the problem of recovering a sharp version of an input blurry image when the blur kernel is unknown [16]. Mathematically, we wish to decompose a blurred image y as

$$y = k \otimes x, \tag{1}$$

where x is a visually plausible sharp image and k is a nonnegative blur kernel whose support is small compared to the image size. This problem is severely ill-posed and there is an infinite set of pairs (x, k) explaining any observed y. For example, one undesirable solution that perfectly satisfies (1) is the no-blur explanation: k is the delta (identity) kernel and $x = y$. The ill-posed nature of the problem implies that additional assumptions on x or k must be introduced.

Blind deconvolution is the subject of numerous papers in the signal and image processing literature, to name a few consider [1], [14], [29], [21], [23] and the survey in [16]. Despite the exhaustive research in many of these classic papers, results are not shown on real images. Recent algorithms have proposed to address the ill-posedness of blind deconvolution by characterizing x using natural image statistics [22], [5], [18], [11], [13], [3], [27], [4], [32]. While this principle has led to tremendous progress, the results are still far from perfect. Blind deconvolution algorithms exhibit some common building principles and vary in others. The goal of this paper is to analyze the problem and shed new light on recent algorithms. What are the key challenges and what are the important components that make blind deconvolution possible? Additionally, which aspects of the problem should attract further research efforts?

One of the puzzling aspects of blind deconvolution is the failure of the MAP approach. Recent papers emphasize the usage of a sparse derivative prior to favor sharp images. However, a direct application of this principle has not yielded the expected results and all algorithms have required additional components, such as marginalization across all possible images [22], [5], [18], spatially varying terms [13], [26], or solvers that vary their optimization energy over time [26]. In this paper, we analyze the source of the MAP failure. We show that, counterintuitively, the most favorable solution under a sparse prior is usually a blurry image and not a sharp one. Thus, the global optimum of the MAP approach is the no-blur explanation. We discuss solutions to the problem and analyze the answers provided by existing algorithms. We show that one key property making blind deconvolution possible is the strong asymmetry between the dimensionalities of x and k. While the number of unknowns in x increases with image size, the dimensionality of k remains small. Therefore, while a simultaneous MAP estimation of both x and k fails, a MAP estimation of k alone (marginalizing over x) is well constrained and recovers an accurate kernel. We suggest that while the sparse prior is helpful, the key component making blind deconvolution possible is not the choice of prior, but the thoughtful choice of estimator. Furthermore, we show that with a proper estimation rule, blind deconvolution can be performed even with a weak Gaussian prior.

Finally, we collect motion-blurred data with ground truth. This data allow us to make a first step toward a quantitative comparison of recent blind deconvolution

- *A. Levin is with the Department of Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: anat.levin@weizmann.ac.il.*
- *Y. Weiss is with the School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel. E-mail: yweiss@cs.huji.ac.il.*
- *F. Durand is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139. E-mail: fredopdurand@gmail.com.*
- *W.T. Freeman is with the Department of Electrical Engineering and Computer Science and the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139. E-mail: billf@mit.edu.*

Manuscript received 26 Aug. 2010; revised 30 Mar. 2011; accepted 11 May 2011; published online 21 July 2011.
Recommended for acceptance by S.B. Kang and I. Essa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMISI-2010-08-0655.
Digital Object Identifier no. 10.1109/TPAMI.2011.148.

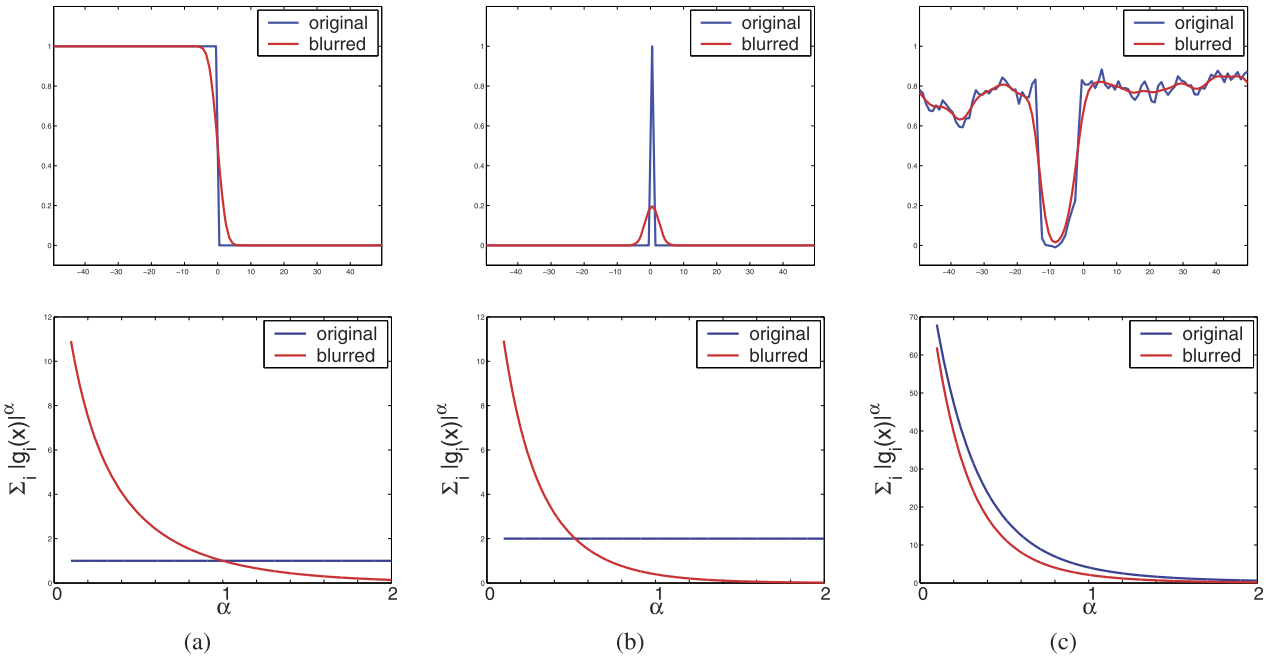


Fig. 1. The $\mathrm{MAP}_{x,k}$ score evaluated on toy 1D signals. Top: Sharp and blurred signals. Bottom: Sum of gradients $-\log p(x) = \sum_i |g_i(x)|^\alpha$ as a function of α.

algorithms. Our evaluation suggests that the variational Bayes approach of [5] outperforms all existing alternatives. This data also show that the shift invariance convolution model involved in most existing algorithms is often violated and that realistic camera shake includes in-plane rotations.

2 $\mathrm{MAP}_{x,k}$ Estimation and Its Limitations

In this paper, y denotes an observed blurry image which is a convolution of an unknown sharp image x with an unknown blur kernel k, plus noise n (this paper assumes i.i.d. Gaussian noise):

$$y = k \otimes x + n. \tag{2}$$

Using capital letters for the Fourier transform of a signal,

$$Y_\omega = K_\omega X_\omega + N_\omega. \tag{3}$$

The goal of blind deconvolution is to infer both k and x given a single input y. Additionally, k is nonnegative, and its support is often small compared to the image size.

The simplest approach is a maximum-a-posteriori ($\mathrm{MAP}_{x,k}$)[1] estimation, seeking a pair $(\hat{x}, \hat{k})$ maximizing

$$p(x, k|y) \propto p(y|x, k)p(x)p(k). \tag{4}$$

For simplicity of exposition, we assume a uniform prior on k. The likelihood term $p(y|x, k)$ is the data fitting term $\log p(y|x, k) = -\lambda \|k \otimes x - y\|^2 + C_1$. The prior $p(x)$ favors natural images, usually based on the observation that their gradient distribution is sparse. A common measure is

$$\log p(x) = -\sum_i |g_{x,i}(x)|^\alpha + |g_{y,i}(x)|^\alpha + C_2, \tag{5}$$

where $g_{x,i}(x)$ and $g_{y,i}(x)$ denote the horizontal and vertical derivatives at pixel i (we use the simple $[-1\ 1]$ filter) and C_1, C_2 are constant normalization terms. Exponent values $\alpha < 1$ lead to sparse priors and natural images usually correspond to α in the range of $[0.5, 0.8]$ [28]. Other choices include a Laplacian prior $\alpha = 1$ and a Gaussian prior $\alpha = 2$. While natural image gradients are very non-Gaussian, we examine this model because it enables an analytical treatment.

Thus, the $\mathrm{MAP}_{x,k}$ approach of (4) translates to seeking $(\hat{x}, \hat{k})$ minimizing

$$(\hat{x}, \hat{k}) = \arg\min_{x,k} \lambda \|k \otimes x - y\|^2 + \sum_i |g_{x,i}(x)|^\alpha + |g_{y,i}(x)|^\alpha. \tag{6}$$

Equation (6) reveals an immediate limitation:

Claim 1. *Let x be an arbitrarily large image sampled from the prior $p(x)$ and $y = k \otimes x$. The image and kernel pair (x, k) maximizing the $\mathrm{MAP}_{x,k}$ score is an image with vanishing norm and a kernel with infinite norm, $\|x\| \to 0$ and $\|k\| \to \infty$.*

Proof. For every pair (x, k) we use a scalar s to define a new pair $x' = s \cdot x, k' = 1/s \cdot k$ with equal data fitting cost $\|k \otimes x - y\|^2 = \|k' \otimes x' - y\|^2$. While the data fitting term is constant, the prior term improves as $s \to 0$. □

This observation is not surprising. The most likely image under the prior in (5) is a flat image with no gradients. One attempt to fix the problem is to assume the mean intensity of the blurred and sharp images should be equal, and constrain the sum of k: $\sum_i k_i = 1$. This eliminates the zero solution, but usually the no-blur solution is still favored.

To understand this, consider the 1D signals x in Fig. 1 that were convolved with a (truncated) Gaussian kernel k^* of standard deviation 4 pixels. We compare two interpretations:

1. We keep estimation variables in subscript to distinguish a MAP estimation of both x and k from a MAP estimation of k alone.

Fig. 2. $\mathrm{MAP}_{x,k}$ failure on real image windows. Windows in which the sharp explanation is favored are marked in red. The percent of windows in which the sharp version is favored decreases with window size.

1) the true kernel: $y = k^* \otimes x^*$, 2) the delta kernel (no blur) $y = \delta \otimes y$. We evaluate the $-\log p(x, k|y)$ score (6) while varying the α parameter in the prior. For the zero noise case, in both of these two interpretations $\|k \otimes x - y\| = 0$, hence the likelihood term vanishes and evaluating $p(x, k|y)$ reduces to evaluating the prior $p(x)$.

For step edges (Fig. 1a), $\mathrm{MAP}_{x,k}$ usually succeeds. The edge is sharper than its blurred version and while the Gaussian prior favors the blurry explanation, appropriate sparse priors ($\alpha < 1$) favor the correct sharp explanation.

In contrast, Fig. 1b presents a narrow peak. Blurring reduces the peak height, and as a result, the Laplacian prior $\alpha = 1$ favors the "no blur" solution ($x = y, k = \delta$) over the correct solution ($x = x^*, k = k^*$) because the absolute sum of gradients is lower. Examining Fig. 1b-bottom suggests that the blurred explanation is winning for smaller α values as well. The sharp explanation is favored only for low alpha values, approaching a binary penalty. However, the sparse models describing natural images are not binary, they are usually in the range $\alpha \in [0.5, 0.8]$ [28].

The last signal considered in Fig. 1c is a row cropped from a natural image, illustrating that natural images contain a lot of medium contrast texture and noise, corresponding to the narrow peak structure. This dominates the statistics more than step edges. As a result, blurring a natural image reduces the overall contrast and, as in Fig. 1b, even sparse priors favor the blurry x explanation.

To confirm the above observation, we blurred the image in Fig. 2 with a Gaussian kernel of standard deviation 3 pixels. We compared the sum of the gradients in the blurred and sharp images using $\alpha = 0.5$. For 15×15 windows the blurred image is favored over 97 percent of the windows, and this phenomenon increases with window size. For 45×45 windows, the blurred version is favored at all windows. Another observation is that if the sharp explanation does win, it happens next to significant edges.

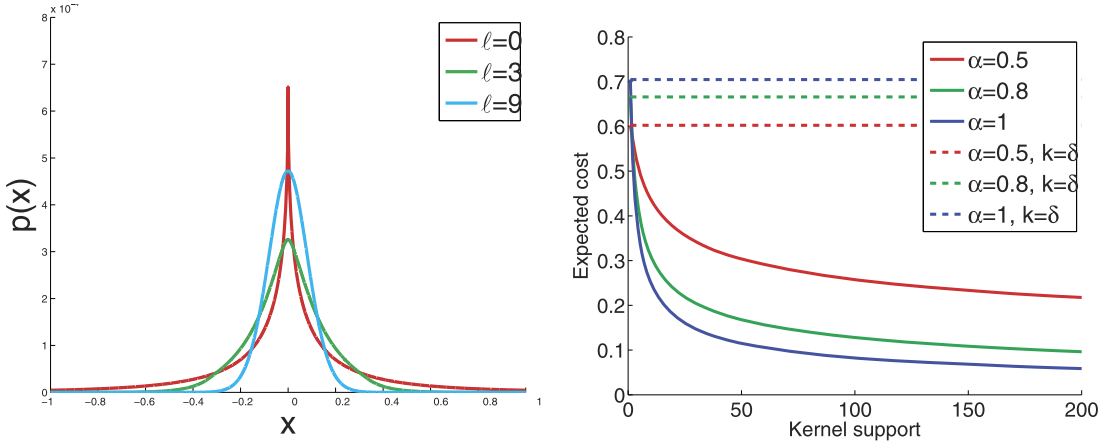
To understand this, note that blur has two opposite effects on the image probability: 1) It makes the signal derivatives less sparse and that reduces the probability. 2) It reduces the variance of the derivatives and that increases its probability. For very specific images, like ideal step edges, the first effect dominates and blur reduces the probability. However, for most natural images, the second effect is stronger and the blur actually increases the probability.

The results of Fig. 2 show that $\mathrm{MAP}_{x,k}$ using a sparsity prior on gradients and a unit sum constraint on k fails to recover the correct blur kernel. One possible reason for this failure is that the prior of (5) is simply a bad prior for natural images, for example, it treats the horizontal and vertical gradients independently and assumes a generalized Gaussian distribution for each. We show that even if the prior is "correct" and the horizontal and vertical gradients are generated independently from a generalized Gaussian distribution, $\mathrm{MAP}_{x,k}$ using this "correct" prior will still fail to recover the true blur kernel, even with infinitely large images.

To illustrate this, let x^0 be a sequence whose gradients are sampled i.i.d. from $p^0(g_i(x^0)) \propto e^{-\gamma|g_i(x^0)|^\alpha}$, x^ℓ a sequence obtained by convolving x^0 with a width ℓ box filter (normalizing the kernel sum to 1), and p^ℓ its probability distribution. The expected negative log probability (effecting the $\mathrm{MAP}_{x,k}$) of x^ℓ under the sharp distribution p^0 is: $E_{p^\ell}[-\log p^0(g(x^\ell))] = -\int p^\ell(g(x)) \log p^0(g(x))dx$. Fig. 3a plots p^ℓ for $\alpha = 0.5$ and Fig. 3b the expected log probability as a function of ℓ. The variance is reduced by convolution, and the negative log probability reduces as well. As a result, a blurred signal is more probable than a sharp one. We note that the observations of Fig. 3 apply not only to box kernels, but also to any kernel k^ℓ with ℓ uniform nonzero entries (that is, the ℓ nonzero entries of k^ℓ equal $1/\ell$). This is simply because the elements of x^0 are sampled i.i.d. and, hence, the statistics are invariant to permutations of kernel entries. For large ℓ values, we can justify this behavior by the central limit theorem. In fact, the same phenomenon is true whenever the L_2 norm of the kernel is small, $\|k\|^2 \ll 1$, because the expected derivative variance in the blurred image decreases as $\|k\|^2$. Examples of kernels whose L_2 norm goes to zero with the support size ℓ include uniform blur kernels, Gaussian blur kernels, and random blur kernels. This idea is formulated in the following claim.

Claim 2. *Let $g_x(x^0), g_y(x^0)$ be infinitely large gradient images*[2] *sampled from the prior*

2. We consider horizontal and vertical gradient fields sampled independently, without enforcing the integrability constraint.



(a) p^l for $\alpha = 0.5$ (b) $E_{p^l}[-\log p^0(x^l)]$

Fig. 3. (a) Comparison of gradient histograms for blurred and unblurred images sampled from $p^0(x)$. Blur kernel width ℓ reduces the average gradient magnitude. (b) Expected negative log probability reduces with blur.

$$p^0(g_{x,i}(x^0), g_{y,i}(x^0)) \propto e^{-|g_{x,i}(x^0)|^\alpha - |g_{y,i}(x^0)|^\alpha}. \quad (7)$$

Let $y = k^* \otimes x^0$ with k^* a unit sum kernel whose L_2 norm is small $\|k^*\|^2 \ll 1$. Then, solving $\mathrm{MAP}_{x,k}$ subject to a unit sum constraint on k will fail to recover the true blur kernel. In particular, if k^* is a unit sum kernel with ℓ uniform nonzero values, $\mathrm{MAP}_{x,k}$ subject to a unit sum constraint is guaranteed to fail for large ℓ values.

Proof. We will show that the no-blur solution $(x = y, k = \delta)$ has higher probability than the correct solution $(x = x^0, k = k^*)$. In both cases, the likelihood term vanishes and the log probability depends only on the histogram of gradients in x. As the image size goes to infinity, the gradient histogram in the true solution will be proportional to the prior (7). Thus, the expected cost of the true solution is simply $C_0 = E_{p^0}[|g(x)|^\alpha]$, where p^0 is a generalized Gaussian with shape parameter α and unit variance. C_0 can be computed analytically

$$C_0 = \frac{\sqrt{\beta}^\alpha}{\alpha}, \quad \text{with} \quad \beta = \Gamma(1/\alpha)/\Gamma(3/\alpha). \quad (8)$$

We now consider unit sum kernels with ℓ uniform nonzero values. The expected cost of the no-blur solution depends on ℓ and is given by $C_\ell = E_{p^\ell}[|g(x)|^\alpha]$. For large ℓ, the central limit theorem implies that $g(x)$, being a sum of ℓ i.i.d. random variables, will be Gaussian with variance $1/\ell$. For a Gaussian p^ℓ, $C_{k^\ell} = E_{p^\ell}[|g(x)|^\alpha]$ can be computed analytically:

$$C_{k^\ell} = \left(\frac{1}{\ell}\right)^{\alpha/2} \frac{\sqrt{2}^{\alpha+1}}{\sqrt{2\pi}} \Gamma(\alpha/2 + 1/2). \quad (9)$$

In other words, for large ℓ, C_{k^ℓ} is proportional to $(1/\ell)^\alpha$ so it is monotonically decreasing and hence becomes smaller than the expected cost of the correct solution as ℓ increases.

For a general kernel, the averaged derivative variance is $\|k^*\|^2$, and

$$C_{k^*} = (\|k^*\|^2)^{\alpha/2} \frac{\sqrt{2}^{\alpha+1}}{\sqrt{2\pi}} \Gamma(\alpha/2 + 1/2). \quad (10)$$

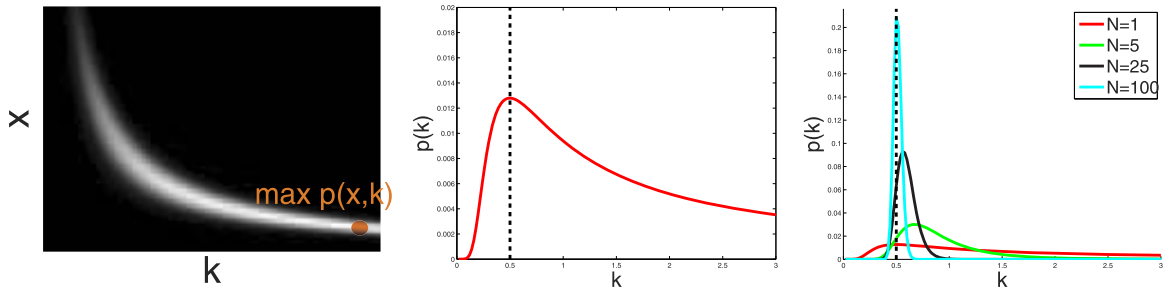
If the L_2 norm of k^* is small $\|k^*\|^2 \ll 1$, the expected cost of the no-blur explanation is smaller than the expected cost of the true solution $C_{k^*} \ll C_0$.

The preceding discussion was about the expected cost of the two solutions. However, note that by the law of large numbers, the histogram of an infinitely large signal approaches the expected one. Thus, for sufficiently long sequences, the claim holds with probability approaching 1. □

The analytic formula for C_ℓ in (9) is accurate only for large ℓ values. While there is no simple close form formula for small ℓ values, we can still compute C_ℓ numerically by averaging over samples, as illustrated in Fig. 3. How large does ℓ have to be for $\mathrm{MAP}_{x,k}$ to fail? This depends on the sparsity of the original image gradients α. For the values of α typically used for modeling natural images (e.g., $\alpha \in [0.5, 1]$), the numerical calculations show that the cost of the no-blur solution will be less than the true solution even for $\ell = 2$. That is, $\mathrm{MAP}_{x,k}$ will fail for any uniform blur kernel. However, for very sparse image priors, $\mathrm{MAP}_{x,k}$ is only guaranteed to fail for blur kernels with larger support (e.g., for $\alpha = 0.3$ the support of the kernel needs to be greater than 7 pixels).

The discussion in this section highlights the fact that the prior alone does not favor the desired result. The source of the problem is that for all α values, the most likely event of the prior in (5) is the fully flat image. This phenomenon is robust to the exact choice of prior, and replacing the model in (5) with higher order derivatives or with more sophisticated natural image priors [24], [30] does not change the result. We also see that the problem is present even if the derivatives signal is sampled exactly from $p(x)$ and the prior is perfectly correct in the generative sense. In the next section, we suggest that to overcome the $\mathrm{MAP}_{x,k}$ limitation, one should reconsider the choice of estimator. We revisit a second group of blind deconvolution algorithms derived from this idea.

$\mathrm{MAP}_{x,k}$ in the literature. Fergus et al. [5] report that their initial attempts to approach blind deconvolution with $\mathrm{MAP}_{x,k}$ failed, resulting in either the original blurred



(a) $p(x,k|y)$ (b) $p(k|y_1)$ (c) $p(k|y_1,..y_N)$

Fig. 4. A toy blind deconvolution problem with one scalar $y = kx + n$ (replotted from [2]). (a) The joint distribution $p(x,k|y)$. A maximum is obtained for $x \to 0, k \to \infty$. (b) The marginalized score $p(k|y)$ produces an optimum closer to the true k^*. (c) The uncertainty of $p(k|y)$ reduces given multiple observations $y_j = kx_j + n_j$.

explanation or a binary two-tone image, depending on parameter tunings.

Algorithms like [13], [11] explicitly detect edges in the image (either manually or automatically), and seek a kernel which transfers these edges into binary ones. This is motivated by the example in Fig. 2, suggesting that $\text{MAP}_{x,k}$ could do the right thing around step edges. Another algorithm which makes usage of this property is [26]. It optimizes a semi-$\text{MAP}_{x,k}$ score, but explicitly detects smooth image regions and reweights their contribution. Thus, the $\text{MAP}_{x,k}$ score is dominated by edges. We discuss this algorithm in detail in [20]. Earlier blind deconvolution papers which exploit a $\text{MAP}_{x,k}$ approach avoid the delta solution using other assumptions which are less applicable for real world images. For example, Ayers and Dainty [1] assume x contains an object on a flat background with a known compact support.

3 MAP_k Estimation

The limitations of MAP estimation in the case of few measurements have been pointed out many times in estimation theory and statistical signal processing [15], [2]. Indeed, in the $\text{MAP}_{x,k}$ problem we can never collect enough measurements because the number of unknowns grows with the image size. In contrast, estimation theory tells us [15] that, given enough measurements, MAP estimators do approach the true solution. Therefore, the key to success is to exploit a special property of blind deconvolution: the strong asymmetry between the dimensionalities of the two unknowns. While the dimensionality of x increases with the image size, the support of the kernel is fixed and small relative to the image size. The image y does provide a large number of measurements for estimating k. As we prove below, for an increasing image size, a MAP_k estimation of k alone (marginalizing over x) can recover the true kernel with an increasing accuracy. This result stands in contrast to Claim 1, which stated that a $\text{MAP}_{x,k}$ estimator continues to fail *even as the number of measurements goes to infinity*. This leads to an alternative blind deconvolution strategy: Use a MAP_k estimator to recover the kernel and, given the kernel, solve for x using a nonblind deconvolution algorithm.

Before providing a formal proof, we attempt to gain an intuition about the difference between MAP_k and $\text{MAP}_{x,k}$ scores. A MAP_k estimator selects $\hat{k} = \arg\max_k p(k|y)$, where $p(k|y) = p(y|k)p(k)/p(y)$, and $p(y|k)$ is obtained by *marginalizing over* x, and evaluating the full volume of possible x interpretations:

$$p(y|k) = \int p(x,y|k)dx. \tag{11}$$

To see the role of marginalization, consider the scalar blind deconvolution problem illustrated in [2]. Suppose a scalar y is observed and should be decomposed as $y = k \cdot x + n$. Assume a zero mean Gaussian prior on the noise and signal, $x \sim N(0,\sigma^2), n \sim N(0,\eta^2)$. Then,

$$P(x,k|y) \propto e^{-\frac{1}{2\eta^2}|kx-y|^2} e^{-\frac{x^2}{2\sigma^2}}. \tag{12}$$

Fig. 4a illustrates the 2D distribution $P(x,k|y)$. Unsurprisingly, it is maximized by $x \to 0, k \to \infty$. On the other hand, $p(y|k)$ is the integral over all x explanations:

$$P(y|k) \propto \int e^{-\frac{1}{2\eta^2}|kx-y|^2} e^{-\frac{x^2}{2\sigma^2}} dx. \tag{13}$$

This integral is not maximized by $k \to \infty$. In fact, if we consider the first term only, $\int e^{-\frac{1}{2\eta^2}|kx-y|^2} dx$, it clearly favors $k \to 0$ values because they allow a larger volume of possible x values. To see that, note that for every k and every $\epsilon > 0$, the size of the set of x values satisfying $|kx - y| < \epsilon$ is $2\epsilon/k$, maximized as $k \to 0$. When the two terms in (13) are taken into account, neither $k \to \infty$ nor $k \to 0$ are favored. We show in Section 3.2.1 that the favored solution is $|k| \approx |y|/\sigma$ and $|x| \approx \sigma$, which make sense because x now behaves like a typical sample from the prior. This is the principle of genericity described in Bayesian terms by Brainard and Freeman [2]. Fig. 4b plots $P(y|k)$, which is essentially summing the columns of Fig. 4a.

Now consider blur in real images: For the delta kernel there is only a single solution $x = y$ satisfying $k \otimes x = y$. However, while the delta spectrum is high everywhere, the true kernel is usually a low pass, and has low-spectrum values. Referring to the notation of (3), if $K_\omega = 0$, an infinite subspace of possible explanations is available as X_ω can be arbitrary (and, with noise, any low $|K_\omega|$ values increase the uncertainty, even if they are not exactly 0). Hence, the true kernel gets an advantage in the $p(y|k)$ score.

We prove that for sufficiently large images, $p(k|y)$ is guaranteed to favor the true kernel.

Claim 3. *Let x be an arbitrarily large image, sampled from a local prior[3] $p(x)$ and $y = k \otimes x + n$. Then, $p(k|y)$ is maximized by the true kernel k^*. Moreover, if $\arg\max_k p(y|k)$ is unique, $p(k|y)$ approaches a delta function.*[4]

Proof. We divide the image into small disjoint windows $\{y^1, \ldots, y^n\}$ and treat them as i.i.d. samples $y^j \sim p(y|k^*)$. We then select $k^{ML} = \arg\max_k \prod_j p(y^j|k)$. Applying the standard consistency theorem for maximum likelihood estimators [15], we know that given enough samples, the ML approaches the true parameters. That is, when $n \to \infty$,

$$p(k^{ML}(\{y^1, \ldots, y^n\}) = k^*) \to 1. \tag{14}$$

Since $p(x)$ is a local prior, taking sufficiently far away disjoint windows will ensure that $p(y|k) \approx \prod_j p(y^j|k)$. Thus, $p(y|k)$ is maximized by k^{ML}. Also, if we select an m times larger image y', $p(y'|k) = p(y|k)^m$. Thus, if $p(y|k) < \max_k p(y|k)$, then $p(y|k) \to 0$. Finally, if $p(k^*) > 0$, then k^{MAP}, k^{ML} are equal on large images since

$$\arg\max_k p(y|k) = \arg\max_k p(y|k)p(k),$$

and, thus, $k^{MAP} \to k^*$. Similarly, if $\max_k p(y|k)$ is unique, $p(k|y)$ approaches a delta function. □

Fig. 4c plots $p(y|k)$ for a scalar blind deconvolution task with N observations $y_j = kx_j + n_j$, illustrating that as N increases, the uncertainty around the solution decreases (compare with Fig. 4b).

3.1 The Loss Function Perspective

As another way to understand the difference between the $\mathrm{MAP}_{x,k}$ and MAP_k estimators, we return to the definition of a Bayesian estimator. A Bayesian estimator involves a loss function $L(\hat{x} - x, \hat{k} - k)$ on both parameters, specifying the price for an estimation error. The expected loss is minimized by

$$(\hat{x}, \hat{k}) = \arg\min \iint p(x, k|y) L(\hat{x} - x, \hat{k} - k) dx dk. \tag{15}$$

One simple choice of loss function yielding the $\mathrm{MAP}_{x,k}$ solution is the Dirac delta loss function: $L(\hat{x} - x, \hat{k} - k) = 1 - \delta((\hat{x}, \hat{k}) - (x, k))$. The limitations of this loss have been pointed out many times [15], [2]. This "all or nothing" loss is too harsh for many signal processing applications, as it completely ignores all information around the mode. Instead, it is common to use loss functions that increase more smoothly with estimation error, such as the mean squared error (MSE) loss: $L(x, k) = |x - \hat{x}|^2 + |k - \hat{k}|^2$, or a robustified loss like the MLM [2].

Claim 4. *If $p(k|y)$ has a unique maxima, then for large images, an MAP_k estimator followed by an MMSE_x image estimation is equivalent to a simultaneous $\mathrm{MMSE}_{x,k}$ estimation of both x and k.*[5]

3. For the accuracy of the proof, we assume that $p(x)$ is a local prior, implying that windows at distance d are independent as $d \to \infty$.

4. Note that Claim 3 does not guarantee that the MAP_k is unique. For example, if the kernel support is not constrained enough, multiple spatial shifts of the kernel provide equally good solutions. The problem can be easily avoided by a weak prior on k (e.g., favoring centered kernels).

5. If multiple solutions with equal probability exist, $\mathrm{MMSE}_{x,k}$ and MAP_k are not fully equivalent, and $\mathrm{MMSE}_{x,k}$ leads to undesired averaging. On the other hand, MAP_k avoids the problem by picking one solution.

Proof. The mean squared error is minimized by the mean, and in our case $\mathrm{MMSE}_{x,k}$ provides

$$\begin{aligned} \hat{x} &= \iint p(x, k|y) x dx dk \\ &= \iint p(k|y) p(x|y, k) x dx dk \\ &= \int p(k|y) \mu^{(k)} dk, \end{aligned} \tag{16}$$

where $\mu^{(k)} = \int p(x|y, k) x dx$, is a "nonblind" MMSE_x estimation of x given k. From Claim 3, $p(k|y)$ is a delta function and, thus, $\hat{x} = \mu^{(k^{MAP})}$. □

3.2 Examples of MAP_k Estimation

Claim 3 reduces to a robust blind deconvolution strategy: Use MAP_k estimator to recover $k^{MAP} = \arg\max_k p(k|y)$, and then use k^{MAP} to solve for x using some nonblind deconvolution algorithm. To illustrate the MAP_k approach, we start with the simple case of a Gaussian prior on $p(x)$, as it permits a derivation in closed form.

3.2.1 The Gaussian Prior

The prior on X in (5) is a convolution and thus diagonal in the frequency domain. If G_x, G_y denote the Fourier transform of the derivatives g_x, g_y, then

$$X \sim N(0, diag(\sigma_\omega^2)) \quad \sigma_\omega^2 = \beta(\|G_{x,\omega}\|^2 + \|G_{y,\omega}\|^2)^{-1}. \tag{17}$$

Note that since a derivative filter is zero at low frequencies and high at higher frequencies, this is similar to the classical $1/f^2$ power spectrum law for images. Denoting noise variance by η^2, we can express $p(X, Y; K) = p(Y|X; K)p(X)$ as

$$p(X, Y; K) \propto e^{-\frac{1}{2\eta^2}\|K_\omega X_\omega - Y_\omega\|^2 - \frac{1}{2\sigma_\omega^2}\|X_\omega\|^2} \tag{18}$$

(see Section 5 for details). Conditioned on k, the mean and mode of a Gaussian are equal

$$X_\omega^{MAP} = \left(|K_\omega|^2 + \frac{\eta^2}{\sigma_\omega^2}\right)^{-1} K_\omega^T Y_\omega. \tag{19}$$

Equation (19) is the classic Wiener filter [9]. One can also integrate X and express $p(Y|K)$ analytically. This is also a diagonal zero mean Gaussian with

$$Y \sim N(0, diag(\phi_\omega^2)), \quad \phi_\omega^2 = \sigma_\omega^2 |K_\omega|^2 + \eta^2. \tag{20}$$

Equation (20) is maximized when $\phi_\omega^2 = |Y_\omega|^2$, and for blind deconvolution, this implies

$$|\hat{K}_\omega|^2 = \max\left(0, \frac{|Y_\omega|^2 - \eta^2}{\sigma_\omega^2}\right). \tag{21}$$

The image estimated using $\hat{K}$ satisfies $|X_\omega|^2 \approx \sigma_\omega^2$. Therefore, MAP_k does not result in a trivial $X = 0$ solution as $\mathrm{MAP}_{x,k}$ would, but in a solution whose variance matches the prior variance σ^2, that is, a solution which looks like a typical sample from the prior $p(X)$.

Another way to interpret the MAP_k, is to note that

$$\log p(Y|K) = \log p(X^{MAP}, Y; K) - \frac{1}{2}\sum_{\omega} \log\left(\frac{|K_\omega|^2}{\eta^2} + \frac{1}{\sigma_\omega^2}\right) + C. \quad (22)$$

Referring to (18), the second term is just the log determinant of the covariance of $p(X|Y;K)$. This second term is optimized when $K_\omega = 0$, i.e., by kernels with more blur. That is, $\log p(Y|K)$ is equal to the $\mathrm{MAP}_{x,k}$ score of the mode plus a term favoring kernels with blur.

The discussion above suggests that the Gaussian MAP_k provides a reasonable solution to blind deconvolution. In the experiment section, we evaluate this algorithm and show that, while weaker than the sparse prior, it can provide acceptable solutions. This stands in contrast to the complete failure of a $\mathrm{MAP}_{x,k}$ approach, even with the seemingly better sparse prior. This demonstrates that a careful choice of estimator is actually more critical than the choice of prior.

Note that (21) is accurate if every frequency is estimated independently. In practice, the solution can be further constrained because the limited spatial support of k implies that the frequency coefficients $\{K_\omega\}$ are linearly dependent. Another important issue is that (21) provides information on the kernel power spectrum alone but leaves uncertainty about the phase. Many variants of Gaussian blind deconvolution algorithms are available in the image processing literature (e.g., [14], [21]), but in most cases only symmetric kernels are considered since their phase is known to be zero. However, realistic camera shake kernels are usually not symmetric. In Section 5, we describe a Gaussian blind deconvolution algorithm which attempts to recover non-symmetric kernels as well.

3.2.2 Approximation Strategies with a Sparse Prior

The challenge with the MAP_k approach is that for a general sparse prior $p(k|y)$, (11) cannot be computed in closed form. Several previous blind deconvolution algorithms can be viewed as approximation strategies for MAP_k, although the authors might not have motivated them in this way.

A simple approximation is proposed by Levin [18] for the 1D blur case. It assumes that the observed derivatives of y are independent (this is usually weaker than assuming independent derivatives of x): $\log p(y|k) = \sum_i \log p(g_{x,i}(y)|k)$. Since $p(g_{x,i}(y)|k)$ is a 1D distribution, it can be expressed as a 1D table or a histogram h^k. The independence assumption implies that instead of summing over image pixels, one can express $p(y|k)$ by summing over histogram bins:

$$\log p(y|k) = \sum_i \log p(g_{x,i}(y)|k) = \sum_j h_j \log\left(h_j^k\right), \quad (23)$$

where h denotes the gradients histogram in the observed image and j is a bin index. Note that maximizing (23) is equivalent to minimizing the histogram distance between the observed and expected histograms h,h^k. This is because the Kullback Leibler divergence is equal to the negative log probability, plus a constant that does not depend on k (the negative entropy):

$$D_{KL}(h, h^k) = \sum_j h_j \log(h_j) - \sum_j h_j \log\left(h_j^k\right). \quad (24)$$

Since the KL divergence is nonnegative, the probability is maximized when the histograms h, h^k are equal. This very simple approach is already able to avoid the delta solution, but, as we demonstrate in Section 4.1, it is not accurately identifying the exact filter width.

A stronger approximation is the variational Bayes mean-field approach taken by Mishkin and Mackay [22] and by Fergus et al. [5]. The idea is to build an approximating distribution with a simpler parametric form:

$$p(x, k|y) \approx q(x, k) = \prod_i q(g_{i,x}(x)) q(g_{i,y}(x)) \prod_j q(k_j). \quad (25)$$

Since q is expressed in the gradient domain, this does not recover x directly. Thus, they also pick the MAP_k kernel from q and then solve for x using nonblind deconvolution.

A third way to approximate the MAP_k is the Laplace approximation [2], which is a generalization of (22),

$$\log p(y|k) \approx \log p(x^{MAP}, y; k) - \frac{1}{2}\log|A| + C, \quad (26)$$

$$A = \frac{\partial^2}{\partial x_i \partial x_j} \log p(x, y; k)|_{x=x^{MAP}}. \quad (27)$$

The Laplace approximation states that $p(y|k)$ can be expressed by the probability of the mode x^{MAP} plus the log determinant of the variance around the mode. As discussed above, higher variance is usually achieved when k contains more zero frequencies, i.e., more blur. Therefore, the Laplace approximation suggests that $p(y|k)$ is the $\mathrm{MAP}_{x,k}$ score plus a term pulling toward kernels with more blur. Unfortunately, in the non-Gaussian case the covariance matrix isn't diagonal and exact inversion is less trivial. Some earlier blind deconvolution approaches [29], [23] can be viewed as simplified forms of a blur favoring term. For example, they bias toward blurry kernels by adding a term penalizing the high frequencies of k or with an explicit prior on the kernel. Another approach was exploit by Bronstein et al. [3]. They note that in the absence of noise and with invertible kernels, $p(k|y)$ can be evaluated exactly for sparse priors as well. This reduces to optimizing the sparsity of the image plus the log determinant of the kernel spectrum.

Blind equalization. Blind deconvolution has been studied extensively in communication engineering under the name *blind equalization*. These interesting algorithms relate to the MAP_k category as they seek equalizer kernels which will match some overall statistics of the deblurred signals to the known statistics of the input. The statistic to be matched could be the average magnitude [8] or the kurtosis [25]. There has been much analysis proving the *global* convergence of gradient algorithms on such criteria to the correct equalizer under a wide range of conditions [12].

4 Evaluating Blind Deconvolution Algorithms

In this section, we compare blind deconvolution strategies on the same data. We start with a synthetic 1D example and in the second part turn to real 2D motion.

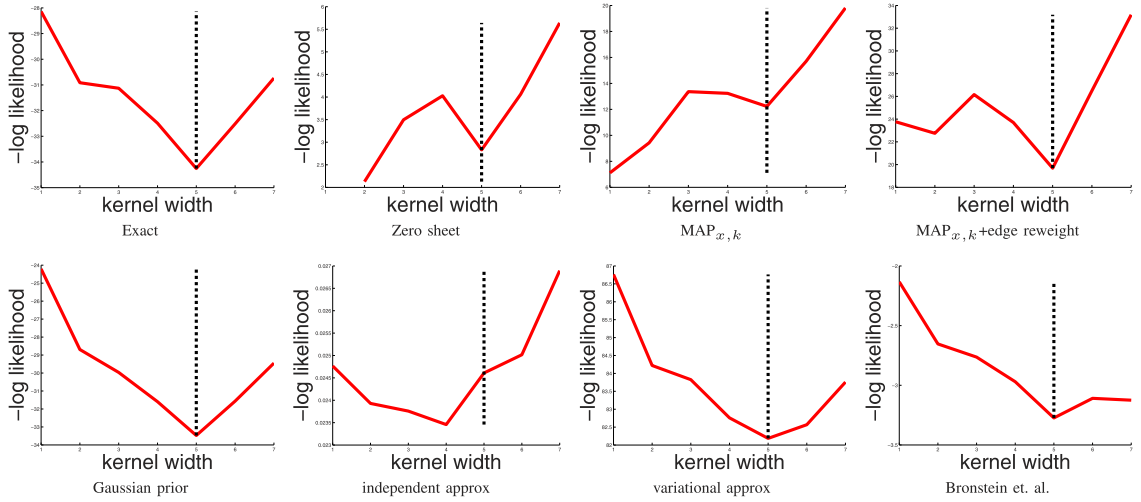


Fig. 5. $\log p(y|k)$ scores using various approximation strategies on 1D image signals. Successful algorithms locate the minimum score at the true kernel width, denoted by the dashed line.

4.1 1D Evaluation

As a first test, we use a set of 1,000 signals of size 10×1 cropped from a natural image. These small 1D signals allow us to evaluate the marginalization integral in (11) exactly, even for a sparse prior. The signals were convolved with a 5-tap box filter (cyclic convolution was used) and an i.i.d. Gaussian noise with standard deviation 0.01 was added. We explicitly search over the explanations of all box filters of size $\ell = 1, \ldots, 7$ taps (all filters normalized to 1). The explicit search allows comparison of the score of different blind deconvolution strategies without folding in optimization errors. (In practice, optimization errors do have a large effect on the successes of blind deconvolution algorithms.)

In Fig. 5, we evaluate various approximations to $p(k|y)$. Fig. 5 plots the approximated negative log probability score for kernel widths $\ell = 1, \ldots, 7$. The exact $-\log p(y|k)$ score is minimized by the true box width $\ell = 5$.

We tested the zero sheet separation (e.g., [17]), an earlier image processing approach with no probabilistic formulation. This algorithm measures the Fourier magnitude of y at the zero frequencies of each box filter k. If the image was indeed convolved with that filter, low Fourier content is expected. However, this approach considers the zero frequencies alone, ignoring all other information, and is known to be noise sensitive. It is also limited to kernel families from a simple parametric form and with a clear zeros structure. In Fig. 5, the zero sheet measure has a local minimum at the right answer but it is not the global minimum.

Supporting the example in Section 2, a pure $\mathrm{MAP}_{x,k}$ approach ($p(y|k) \approx p(x^{MAP}, y|k)$) favors no-blur ($\ell = 1$). Reweighting the derivative penalty around edges [26] can improve the situation, but the delta solution still provides a noticeable local optimum.

The correct minimum is favored with a variational Bayes approximation [5] and with the semi-Laplace approximation of [3]. The independence approximation [18] is able to overcome the delta solution, but does not localize the solution very accurately (minimum at $\ell = 4$ instead of $\ell = 5$). Finally, the correct solution is identified even with the poor image prior provided by a Gaussian model, demonstrating that the choice of estimator ($\mathrm{MAP}_{x,k}$ versus MAP_k) is more critical than the actual prior (Gaussian versus sparse).

Since claim 3 guarantees success only for large images, we attempt to evaluate how large an image should be in practice. Fig. 6 plots the uncertainty in $p(k|y)$ for multiple random samples of $N10 \times 1$ columns. The probability is tightly peaked at the right answer for as little as $N = 20$ columns. The search space in Fig. 6 is limited to the single parameter family of box filters. In real motion deblurring, one searches over a larger family of kernels and a larger uncertainty is expected.

4.2 2D Evaluation

To see whether the predicted difference in performance between MAP_k and $\mathrm{MAP}_{x,k}$ happens in real images, we have collected blurred data with ground truth. We capture a sharp version of a planar scene (Fig. 7a) by mounting the camera on a tripod, as well as a few blurred shots of the same scene. Using the sharp reference, we solve for a nonnegative kernel k minimizing $\|k \otimes x - y\|^2$. The scene in Fig. 7a includes high-frequency noise patterns which helps

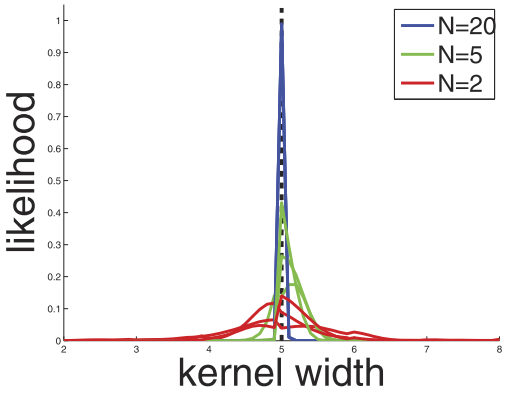
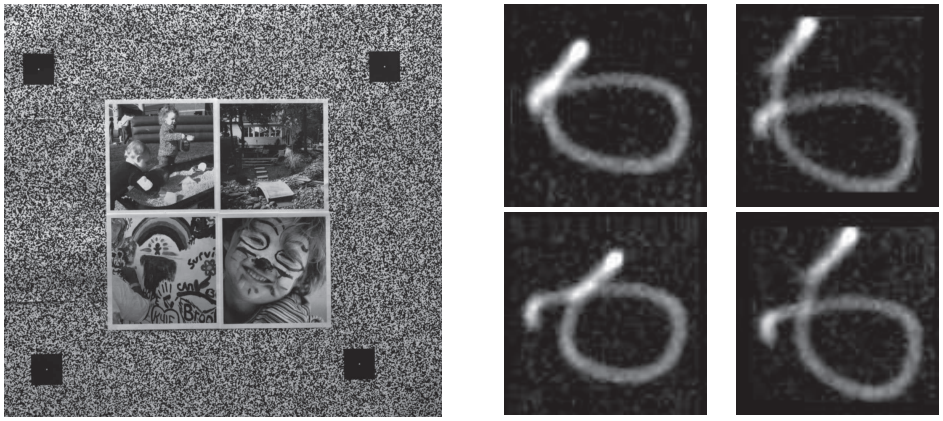


Fig. 6. The uncertainty in kernel estimation decreases with more samples. For as little as $N = 20$ columns it is already tightly peaked at the true answer.



(a) (b)

Fig. 7. Ground truth data acquisition. (a) Calibration image. (b) Smear of points at four corners, each somewhat different, demonstrating that the spatially uniform blur model is violated.

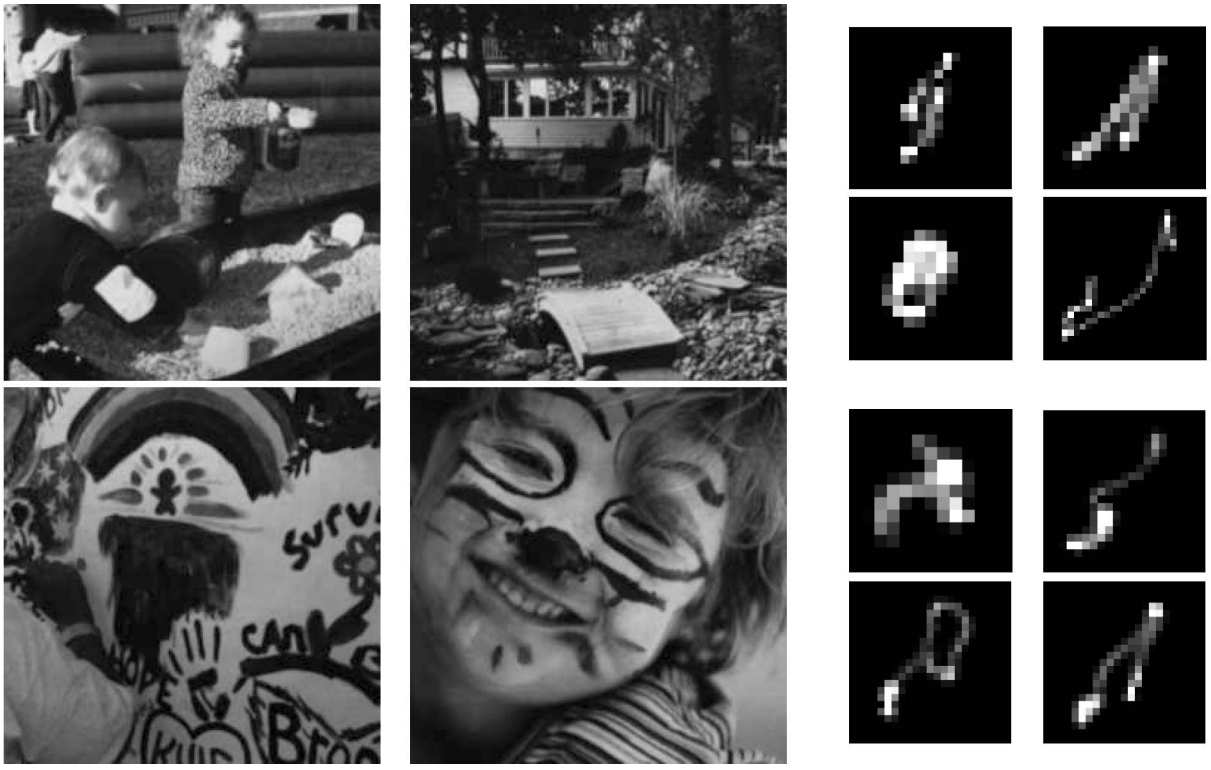


Fig. 8. Ground truth data: Four images and eight blur kernels, resulting in 32 test images.

to stabilize the constraints on k. The central area of the frame includes four real images used as input to the various blind deconvolution algorithms.

We first observed that assuming a uniform blur over the image is not realistic, even for planar scenes. For example, Fig. 7b shows traces of points at four corners of an image captured by a hand-held camera, with a clear variation between the corners. This suggests that an in-plane rotation (rotation around the z-axis) is a significant component of human hand shake. Recently, Whyte et al. [31] proposed a new blind deconvolution approach which can account for in-plane rotations. However, a uniform assumption is made by most existing algorithms we wish to evaluate; hence, we need to evaluate them on data which obey their assumption. To capture images with spatially invariant blur we placed the camera on a tripod, locking the Z-axis rotation handle of the tripod but loosening the X and Y handles. We calibrated the blur of eight such images and cropped four 255×255 windows from each, leading to 32 test images, displayed in Fig. 8 and available online.[6] We used an 85 mm lens and a 0.3 seconds exposure. The kernels' support varied from 10 to 25 pixels.

We can measure the SSD error between a deconvolved output and the ground truth. However, wider kernels result in larger deconvolution error even with the true kernel. To normalize this effect, we measure the ratio between deconvolution error with the estimated kernel and deconvolution with the ground truth kernel. In Fig. 9, we plot the cumulative histogram of error ratios (e.g., bin $r = 3$ counts

6. www.wisdom.weizmann.ac.il/~levina/papers/LevinEtalCVPR09Data.zip.

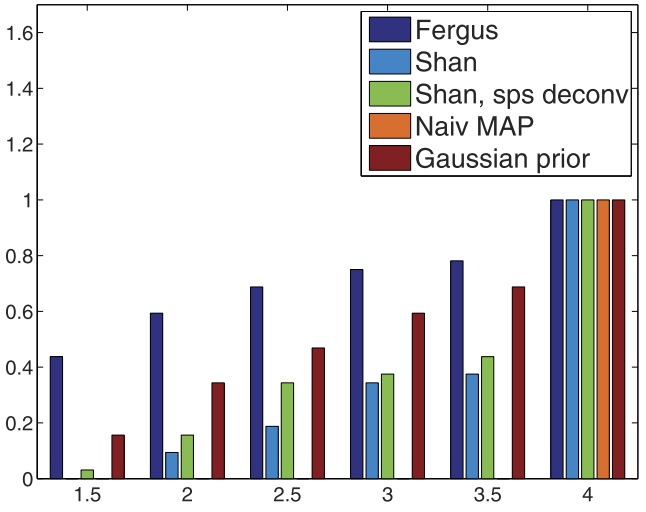


Fig. 9. Evaluation results: Cumulative histogram of the deconvolution error ratio across the 32 test examples. Bar height indicates the percentage of image having error ratio below r; high bars indicate better performance.

the percentage of test examples achieving error ratio below 3). Empirically, we noticed that error ratios above 2 are already visually implausible. One test image is presented in Fig. 10, all others are included in [20].

We have evaluated the algorithms of Fergus et al. [5] and Shan et al. [26] (each using the authors' implementation), as well as MAP_k estimation using a Gaussian prior (described in Section 5), and a simplified $\mathrm{MAP}_{x,k}$ approach constraining $\sum k_i = 1$ (we used coordinate descent, iterating between holding x constant and solving for k, and then holding k constant and solving for x using the sparse deconvolution algorithm of [19]). We note that a thorough evaluation of all blind deconvolution algorithms is outside the scope of this paper, and seems like an important topic for future research. To keep the experimental evaluation focused, we included only algorithms which are truly MAP_k or $\mathrm{MAP}_{x,k}$ approaches. We note that several other good blind deconvolution algorithms were proposed in the literature [4], [32], [18], [13], [3]. However, [4], [32] do not fall into the MAP_k, $\mathrm{MAP}_{x,k}$ categories. The algorithms of [18], [13], [3] were not tested because the first was designed for 1D motion only and the others focus on smaller blur kernels.

We made our best attempt to adjust the parameters of Shan et al. [26], but run all test images with equal parameters. Fergus et al. [5] used Richardson-Lucy nonblind deconvolution in their code. Since this algorithm is a source for ringing artifacts, we improved the results using the kernel estimated by the authors' code with the (nonblind) sparse deconvolution of [19]. Similarly, we used sparse deconvolution with the kernel estimated by Shan et al.

Our experimental evaluation show that the pitfalls of $\mathrm{MAP}_{x,k}$ estimation, which we proved analytically for images sampled from the prior, also happen with real images and indeed these methods favor the no-blur solution. At the same time, the advantages of the MAP_k approach, which we proved analytically for images sampled from the prior, can also be seen with real images and the variational method of Fergus et al. [5] performed significantly better than the competing algorithms which used a similar prior over natural images but a different estimation strategy. It is important to remember that algorithms that are based on different priors, or even ones that are not based on priors at all, were not included in our comparison and it would be interesting to perform a more comprehensive comparison in the future.

These evaluation highlights the fact that the choice of estimator ($\mathrm{MAP}_{x,k}$ versus MSE and MAP_k) is much more critical than the actual prior (Gaussian versus sparse). We note that this does not imply that any sort of alternating minimization will always fail nor that $\mathrm{MAP}_{x,k}$ will fail for any prior. However, for the common reasonable priors, $\mathrm{MAP}_{x,k}$ will fail while MAP_k will succeed, even in cases where the prior is perfectly "correct" in the generative sense.

We note that many of the artifacts in the original results of [5] can be attributed to the Richardson-Lucy nonblind deconvolution artifacts or to nonuniform blur in their test images. Our comparison also suggests that applying sparse deconvolution using the kernels produced by Shan et al. [26] improves their results. As expected, the naive $\mathrm{MAP}_{x,k}$ approach produces small kernels approaching the delta solution.

5 Blind Deconvolution with a Gaussian Prior

To complete Section 3.2.1, we provide a detailed derivation of a MAP_k estimation algorithm using a Gaussian prior. The simple analytic treatment of a Gaussian prior is attractive both from a computational viewpoint and from a research viewpoint, as it affords intuition. While the algorithm is not as powerful as sparse deconvolution algorithms, it approaches the solution using second order statistics alone.

To derive the Gaussian algorithm, we rewrite the generative model explicitly for a Gaussian prior and, to simplify notation, use the frequency domain.

1. $\mathbf{p}(\mathbf{Y}|\mathbf{X};\mathbf{K})$: The spatial i.i.d. Gaussian observation noise is invariant to the frequency basis change. Therefore,

$$(Y_\omega|X_\omega;K_\omega) \sim N(K_\omega X_\omega, \eta^2), \tag{28}$$

where η^2 denotes the noise variance.

2. $\mathbf{p}(\mathbf{X})$: The prior on X uses a convolution and is diagonal in the frequency domain. If G_x, G_y denote the Fourier transform of the derivative filters g_x, g_y, the convolution and Parseval's theorems result in $\sum_i |g_{x,i}(x)|^2 + |g_{y,i}(x)|^2 = \sum_\omega |G_{x,\omega}X_\omega|^2 + |G_{y,\omega}X_\omega|^2$. Therefore, X follows a zero mean Gaussian distribution with diagonal covariance:

$$X \sim N(0, diag(\sigma_\omega^2)) \quad \sigma_\omega^2 = \beta(\|G_{x,\omega}\|^2 + \|G_{y,\omega}\|^2)^{-1} \tag{29}$$

(the scale β can be fitted based on the derivative histogram in a natural image). Note that since a derivative filter is zero at the low frequencies and high at the higher frequencies, this is very similar to the classical $1/f^2$ power spectrum law (and our algorithm produced very similar results with an explicit $1/f^2$ prior).

3. MAP_x estimation:

$$X^{MAP} = \arg\max p(X,Y;K) = \arg\max p(Y|X;K)p(X). \tag{30}$$

Therefore, solving for the MAP_x (using (28) and (29)) is a least square minimization:

$$X_\omega^{MAP} = \arg\min \frac{1}{\eta^2}\|K_\omega X_\omega - Y_\omega\|^2 + \frac{1}{\sigma_\omega^2}\|X_\omega\|^2, \tag{31}$$

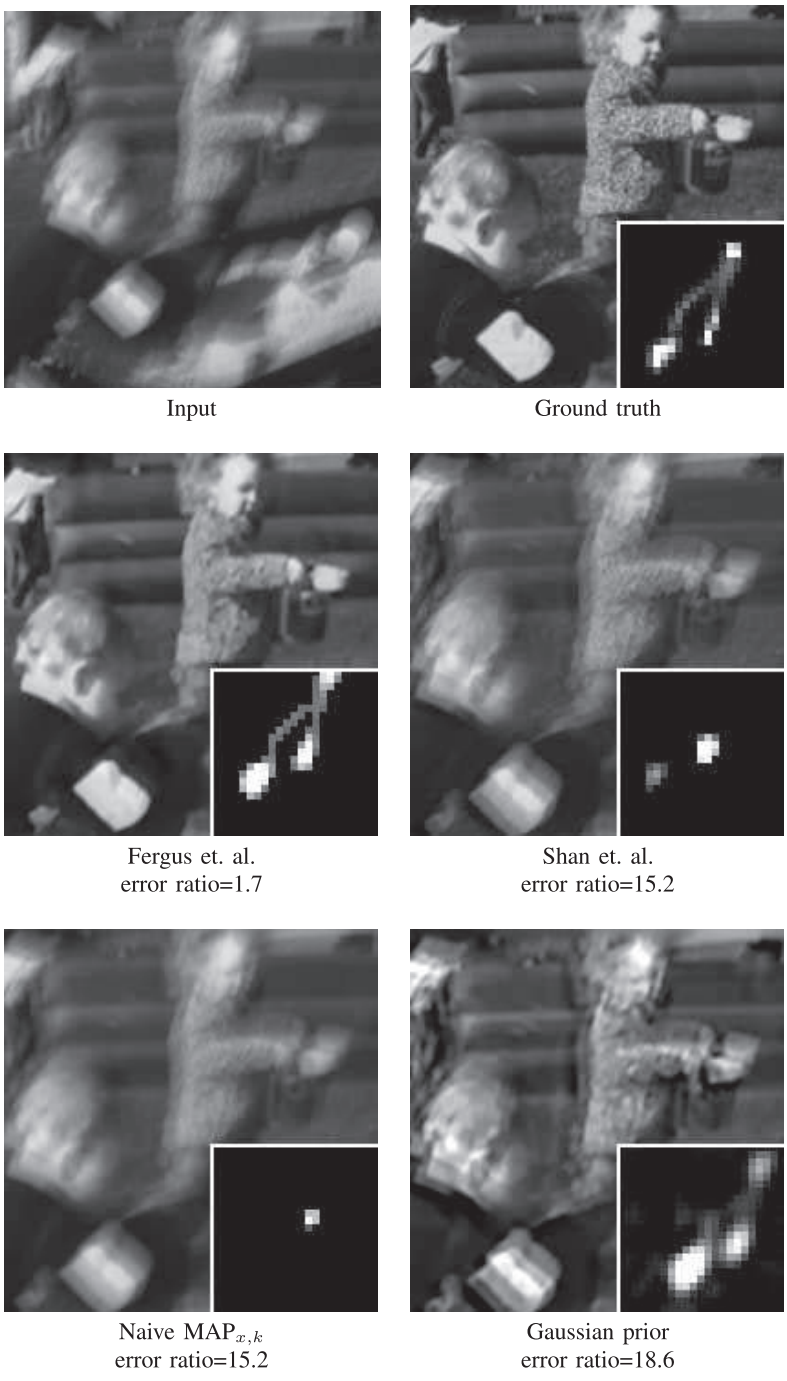


Fig. 10. Visual deconvolution results by various deconvolution algorithms. See [20] for more examples.

$$X_\omega^{MAP} = \left(|K_\omega|^2 + \frac{\eta^2}{\sigma_\omega^2}\right)^{-1} K_\omega^T Y_\omega. \tag{32}$$

Equation (32) is essentially the famous Wiener filter [9]. The prior term in (32) pulls the estimation toward zero, pulling stronger at high frequencies where the expected signal magnitude is small ($\sigma_\omega \to 0$) and noise contribution is higher. When the filter value $K_\omega = 0$, the signal value cannot be recovered and the prior leads the estimation to $X_\omega = 0$.

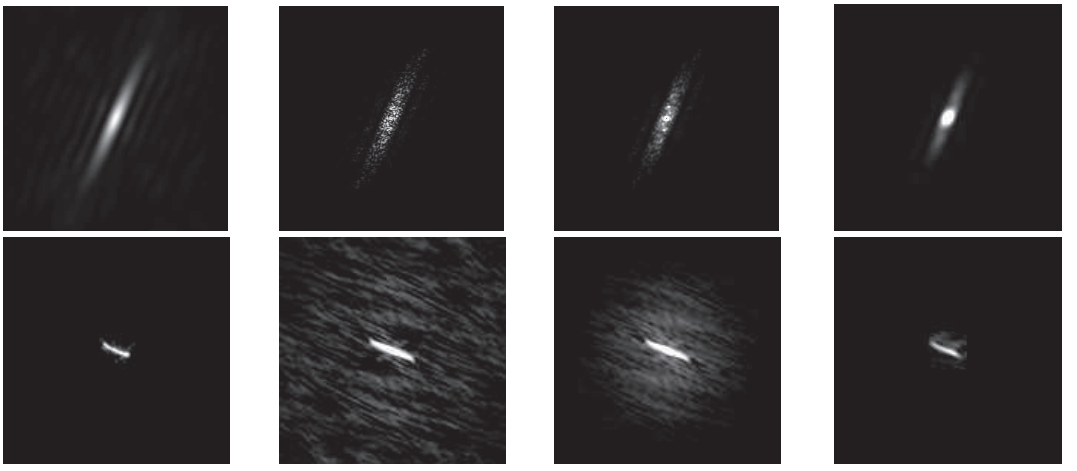
4. $\mathbf{p}(\mathbf{Y})$: One can also integrate X and express $p(Y|K)$ analytically. This is also a diagonal zero mean Gaussian with

$$Y \sim N\big(0, diag(\phi_\omega^2)\big), \quad \phi_\omega^2 = \sigma_\omega^2 |K_\omega|^2 + \eta^2. \tag{33}$$

Given (28)-(33), we can return to blind deconvolution. If we were to estimate every frequency K_ω independently, we could differentiate (33) and conclude it is maximized when $\phi_\omega^2 = |Y_\omega|^2$, which results in

$$|K_\omega|^2 = \max\left(0, \frac{|Y_\omega|^2 - \eta^2}{\sigma_\omega^2}\right). \tag{34}$$

Equation (34) essentially states that the optimal K leads to an X whose power spectrum equals the expected power spectrum σ^2. However, for frequencies ω in which the observed signal value is below the noise variance (i.e., $|Y_\omega|^2 < \eta^2$), the estimator acknowledges that K_ω cannot be



(a) Ground truth (b) Independent estimation (c)Smoothing PWS (d) Compact sup. constraint

Fig. 11. Power spectrum estimation and the compact support constraint. Top: Power spectrum, Bottom: Kernel in primal domain.

recovered and outputs 0. Below, we make usage of this point to derive a coarse-to-fine algorithm. In Fig. 11b we show the filter estimated using (34). The estimation nicely resembles the overall shape and power spectrum of the true filter (Fig. 11a), but is far too noisy to be acceptable. This noise is not surprising as every component of K was estimated from a single measurement.

The signal processing literature [15] addresses the problem of power spectrum estimation (also known as the Periodogram), suggesting that the power spectrum of the observed signal Y should be smoothed before applying (34). While such a smoothing operation increases the bias of the estimation, it significantly reduces its variance. Fig. 11c demonstrates the estimation from a smoothed power spectrum. One can note that as smoothing reduces the fluctuation in the frequency domain, the support of the filter in the primal domain becomes more compact. This leads to another important property of the problem that was ignored so far: While (34) estimates every Fourier coefficient independently, the number of free parameters to estimate in K is much smaller than the image size since a typical filter is assumed to have only a small compact support. Fig. 11d presents the estimated kernel once a compact support was enforced (according to the algorithm described below). This constraint significantly increases the stability of the estimation.

5.1 Phase Estimation

While (34) defines the power spectrum of K, it leaves us with a complete ambiguity regarding its phase. In fact, for every solution K, X such that $Y_\omega = K_\omega X_\omega$ and for any phase vector θ_ω, the pair $\widetilde{K}_\omega = K_\omega e^{i\theta_\omega}, \widetilde{X}_\omega = X_\omega e^{-i\theta_\omega}$ is an equally valid solution, satisfying $Y_\omega = \widetilde{K}_\omega \widetilde{X}_\omega$. The prior on X does not help in resolving this ambiguity—as the Gaussian prior in (29) depends only on the power spectrum, $p(\widetilde{X}) = p(X)$. However, while every phase can maintain the convolution model, most random phase choices destroy the finite support of K. The question of estimating the signal phase given the power spectrum has a long history in signal processing. Hayes [10] states that for most kernels, a finite support constraint uniquely defines the signal phase, up to 1) shift and 2) flipping (mirroring). While a shift ambiguity in deconvolution is reasonable and does not effect the visual quality of the deconvolved image, deconvolving the image with the mirrored filter leads to noticeable artifacts, as illustrated in Fig. 13. For the implementation in this paper, we escape this ambiguity by noticing that while the original image x (in the spatial domain) is nonnegative, deconvolving y with the mirrored filter often leads to negative x values. Yet, this ambiguity highlights one of the weaknesses of second order statistics. While the second order statistics of the images in Figs. 13a and 13b are equal, it is clear that every simple sparse measure will favor Fig. 13a. Nevertheless, we show that the second order statistics plus finite support constraint can get us surprisingly close to the true solution.

While a bounded support constraint removes most phase ambiguity, recovering the phase algorithmically is not a trivial question. A popular gradient-based optimization scheme is the Gerchberg-Saxton [7], [6] algorithm. This algorithm initializes the kernel phase randomly, and then alternates between primal-frequency transformations, enforcing the finite support constraint in the primal domain and the required power spectrum in the frequency domain.

5.2 EM Optimization

Applying the Gerchberg-Saxton algorithm [7], [6] to the independent power spectrum estimated from (33) provides a reasonable initialization for our algorithm. We then proceed with an EM algorithm. The E-step computes the expected mean and variance for the deblurred image X. The M-step uses the second order statistics of X to solve for k, enforcing two constraints: the finite support constraint discussed above, plus the simple requirement that the blur kernel k (in the spatial domain) is nonnegative.

E-step: Applying (32):

$$<X_\omega> = \left(|K_\omega|^2 + \frac{\eta^2}{\sigma_\omega^2}\right)^{-1} K_\omega^T Y_\omega, \tag{35}$$

$$<X_\omega^T X_\omega> = \left(|K_\omega|^2 + \frac{\eta^2}{\sigma_\omega^2}\right) + <X_\omega>^T <X_\omega>. \tag{36}$$

M-step: Transform $<X>$ and $<XX>$ to the spatial domain and solve for k minimizing $<k \otimes x - y>$ subject to finite support and nonnegativity.

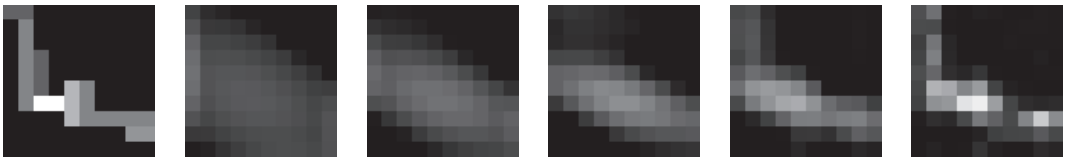


Fig. 12. Coarse to fine kernel estimation. (a) Ground truth. (b), (c), (d), (e), and (f) Estimated kernels with decreasing noise variance.

To express this minimization, suppose that k is an $l \times l$ filter. We denote by x_{w_i} the $l \times l$ window around the ith pixel such that $y_i = \sum_{j \in w_i} k_j x_j$. Let A be an $m \times l^2$ matrix whose rows are the windows x_{w_i}, and m is the number of windows included in the image. If x, y are known, the best filter k is the one minimizing

$$\|Ak(:) - y(:)\|^2 = k(:)^T A^T Ak(:) - 2y(:)^T Ak(:) + y(:)^T y(:) \quad s.t. k \geq 0, \tag{37}$$

where the $(:)$ notation indicates a lexicographical ordering of all the array elements. Note that the number of unknowns in this system is equal to the kernel size l^2, which is much lower than the number of pixels in the image. In practice, we do not precisely know x, but from the E-step, we have access to $<A^T A>$ and $<A>$.

This is a quadratic minimization subject to linear constraints, and thus a convex problem that can be solved using quadratic programming.

5.3 Coarse-to-Fine

Fergus et al. [5] estimated the kernel in a coarse-to-fine scheme. In our case, (33) provides an easy way to implement this. We initialize the optimization with a high-noise variance η^2. As a result, all frequencies with observation below the noise variance (usually the high frequencies) are set to zero, and we mostly optimize the low frequencies of the kernel. Once the low-frequency optimization starts to converge, we gradually reduce the noise variance η^2, allowing more and more bands of frequencies to be nailed down. The kernels estimated with varying η^2 values are illustrated in Fig. 12.

6 Discussion

This paper analyzes the major building blocks of recent blind deconvolution algorithms. We illustrate the limitation of the simple $\mathrm{MAP}_{x,k}$ approach, favoring the no-blur (delta kernel) explanation. One class of solutions involves explicit edge detection. A more principled strategy exploits the dimensionality asymmetry, and estimates MAP_k while marginalizing over x. While the computational aspects involved with this marginalization are more challenging, existing approximations are powerful.

We have collected motion blur data with ground truth and quantitatively compared existing algorithms. Our comparison suggests that the variational Bayes approximation [5] significantly outperforms all existing alternatives.

The conclusions from our analysis are useful for directing future blind deconvolution research. In particular, we note that modern natural image priors [24], [30] do not overcome the $\mathrm{MAP}_{x,k}$ limitation (and in our tests did not change the observation in Section 2). While it is possible that blind deconvolution can benefit from future research on natural image statistics, this paper suggests that better estimators for existing priors may have more impact on future blind deconvolution algorithms. Additionally, we observed that the popular spatially uniform blur assumption is usually unrealistic. Thus, it seems that blur models which can relax this assumption [27] have a high potential to improve blind deconvolution results.

(a) Deconvolution with correct filter (b) Deconvolution with mirrored filter

Fig. 13. Mirroring ambiguity with second order statistics.

Acknowledgments

The authors thank the Israel Science Foundation, US-Israel Bi-National Science Foundation, the Royal Dutch/Shell Group, NGA NEGI-1582-04-0004, MURI Grant N00014-06-1-0734, US National Science Foundation (NSF) CAREER award 0447561. F. Durand acknowledges a Microsoft Research New Faculty Fellowship and a Sloan Fellowship.

References

[1] G.R. Ayers and J.C. Dainty, "Interative Blind Deconvolution Method and Its Applications," *Optics Letters,* vol. 13, pp. 547-549, 1988.
[2] D. Brainard and W. Freeman, "Bayesian Color Constancy," *J. Optical Soc. of Am.,* vol. 14, pp. 1393-1411, 1997.
[3] M.M. Bronstein, A.M. Bronstein, M. Zibulevsky, and Y.Y. Zeevi, "Blind Deconvolution of Images Using Optimal Sparse Representations," *IEEE Trans. Image Processing,* vol. 14, no. 6, pp. 726-736, June 2005.
[4] S. Cho and S. Lee, "Fast Motion Deblurring," *Proc. ACM SIGGRAPH,* 2009.
[5] R. Fergus, B. Singh, A. Hertzmann, S.T. Roweis, and W.T. Freeman, "Removing Camera Shake from a Single Photograph," *Proc. ACM SIGGRAPH,* 2006.
[6] J.R. Fienup, "Phase Retrieval Algorithms: A Comparison," *Applied Optics,* vol. 21, pp. 2758-2769, Aug. 1982.
[7] R.W. Gerchberg and W.O. Saxton, "A Practical Algorithm for the Determination of Phase from Image and Diffraction Plane Pictures," *Optik,* vol. 35, pp. 237-246, 1972.
[8] D.N. Godard, "Self-Recovering Equalization and Carrier Tracking in Two-Dimensional Data Communication Systems," *IEEE Trans. Comm.,* vol. 28, no. 11, pp. 1867-1875, Nov. 1980.
[9] R.C. Gonzalez and R.E. Woods, *Digital Image Processing.* Prentice Hall, Jan. 2002.

[10] M. Hayes, "The Reconstruction of a Multidimensional Sequence from the Phase or Magnitude of Its Fourier Transform," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 30, no. 2, pp. 140-154, Apr. 1982.
[11] J. Jia, "Single Image Motion Deblurring Using Transparency," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[12] C.R. Johnson, P. Schniter, T.J. Endres, J.D. Behm, D.R. Brown, and R.A. Casas, "Blind Equalization Using the Constant Modulus Criterion: A Review," *Proc. IEEE,* vol. 86, no. 10, pp. 1927-1950, Oct. 1998.
[13] N. Joshi, R. Szeliski, and D. Kriegman, "PSF Estimation Using Sharp Edge Prediction," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[14] A.K. Katsaggelos and K.T. Lay, "Maximum Likelihood Blur Identification and Image Restoration Using the EM Algorithm," *IEEE Trans. Signal Processing,* vol. 39, no. 3, pp. 729-733, Mar. 1991.
[15] S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice Hall, 1997.
[16] D. Kundur and D. Hatzinakos, "Blind Image Deconvolution," *IEEE Signal Processing Magazine,* vol. 13, no. 3, pp. 43-64, May 1996.
[17] R.G. Lane and R.H.T. Bates, "Automatic Multidimensional Deconvolution," *J. Optical Soc. of Am. A,* vol. 4, no. 1, pp. 180-188, 1987.
[18] A. Levin, "Blind Motion Deblurring Using Image Statistics," *Proc. Advances in Neural Information Processing Systems,* 2006.
[19] A. Levin, R. Fergus, F. Durand, and W. Freeman, "Image and Depth from a Conventional Camera with a Coded Aperture," *Proc. ACM SIGGRAPH,* 2007.
[20] A. Levin, Y. Weiss, F. Durand, and W.T. Freeman, "Understanding and Evaluating Blind Deconvolution Algorithms," Technical Report MIT-CSAIL-TR-2009-014, 2009.
[21] A.C. Likas and N.P. Galatsanos, "A Variational Approach for Bayesian Blind Image Deconvolution," *IEEE Trans. Signal Processing,* vol. 52, no. 8, pp. 2222-2233, Aug. 2004.
[22] J.W. Miskin and D.J.C. MacKay, "Ensemble Learning for Blind Image Separation and Deconvolution," *Proc. Advances in Independent Component Analysis,* 2000.
[23] R. Molina, A.K. Katsaggelos, J. Abad, and J. Mateos, "A Bayesian Approach to Blind Deconvolution Based on Dirichlet Distributions," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* 1997.
[24] S. Roth and M.J. Black, "Fields of Experts: A Framework for Learning Image Priors," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2005.
[25] O. Shalvi and E. Weinstein, "New Criteria for Blind Deconvolution of Nonminimum Phase Systems (Channels)," *IEEE Trans. Information Theory,* vol. 36, no. 2, pp. 312-321, Mar. 1990.
[26] Q. Shan, J. Jia, and A. Agarwala, "High-Quality Motion Deblurring from a Single Image," *Proc. ACM SIGGRAPH,* 2008.
[27] Q. Shan, W. Xiong, and J. Jia, "Rotational Motion Deblurring of a Rigid Object from a Single Image," *Proc. IEEE 11th Int'l Conf. Computer Vision,* 2007.
[28] E.P. Simoncelli, "Bayesian Denoising of Visual Images in the Wavelet Domain," *Bayesian Inference in Wavelet Based Models,* Springer-Verlag, 1999.
[29] E. Thiébaut and J.-M. Conan, "Strict A Priori Constraints for Maximum-Likelihood Blind Deconvolution," *J. Optical Soc. of Am. A,* vol. 12, no. 3, pp. 485-492, 1995.
[30] Y. Weiss and W.T. Freeman, "What Makes a Good Model of Natural Images?" *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[31] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-Uniform Deblurring for Shaken Images," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.
[32] L. Xu and J. Jia, "Two-Phase Kernel Estimation for Robust Motion Deblurring," *Proc. 11th European Conf. Computer Vision,* 2010.

Anat Levin received the PhD degree from the Hebrew University in 2006 and did postdoctoral work at the Massachusetts Institute of Technology. She is a senior scientist at the Weizmann Institute of Science Department of Computer Science and Applied Mathematics, working on image processing, computer vision, and computer graphics.

Yair Weiss received the PhD degree from the Massachusetts Institute of Technology in 1998, where he worked with Ted Adelson, and he did postdoctoral work at the University of California Berkeley. He is a professor at the Hebrew University School of Computer Science and Engineering, working on machine vision, machine learning, and neural computation. Since 2005, he has been a fellow of the Canadian Institute for Advanced Research.

Fredo Durand received the PhD degree from Grenoble University, France, in 1999, supervised by Claude Puech and George Drettakis. He is an associate professor in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT) and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL). From 1999 till 2002, he was a postdoctoral research in the MIT Computer Graphics Group with Julie Dorsey. He works both on synthetic image generation and computational photography, where new algorithms afford powerful image enhancement and the design of imaging system that can record richer information about a scene. His research interests span most aspects of picture generation and creation, with emphasis on mathematical analysis, signal processing, and inspiration from perceptual sciences. He coorganized the first Symposium on Computational Photography and Video in 2005, the first International Conference on Computational Photography in 2009, and was on the advisory board of the Image and Meaning 2 conference. He received an inaugural Eurographics Young Researcher Award in 2004, an US National Science Foundation (NSF) CAREER award in 2005, an inaugural Microsoft Research New Faculty Fellowship in 2005, a Sloan fellowship in 2006, and a Spira award for distinguished teaching in 2007.

William T. Freeman is a professor of computer science at the Massachusetts Institute of Technology, joining the faculty in 2001. His current research interests include machine learning applied to computer vision and graphics, and computational photography. He is active in the program and organizing committees of the major computer vision, graphics, and machine learning conferences. He was the program cochair for the IEEE International Conference on Computer Vision (ICCV) in 2005, and will be the program cochair for the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2013. From 1981-1987, he worked at Polaroid, developing image processing algorithms for electronic cameras and printers. In 1987-1988, he was a foreign expert at the Taiyuan University of Technology, People's Republic of China. From 1992-2001, he worked at Mitsubishi Electric Research Labs (MERL), in Cambridge, Massachusetts, most recently as a senior research scientist and an associate director. His hobby is flying cameras in kites. He holds 30 patents and is fellow of the IEEE.