Subjective Evaluations of Example-based, Total Variation, and Combined Regularization for Image Processing

Hyrum S. Anderson, Maya R. Gupta
Department of Electrical Engineering
University of Washington
Seattle, WA 98195, USA
{gupta}@ee.washington.edu

and Jon Y. Hardeberg
Gjøvik University College,
Gjøvik, Norway
{jon.hardeberg}@hig.no
Abstract

We present subjective evaluations of example-based regularization, total variation regularization, and a proposed joint example-based and total variation regularization for image estimation problems. We focus on the noisy deblurring problem, which generalizes image superresolution and denoising. Controlled subjective experiments show that the proposed joint regularization can yield significant improvement over only using total variation or example-based regularization, particularly when the example images contain similar structural elements as the test image. We also investigate whether the regularization parameters can be trained by cross-validation, and the difference in cross-validation judgments made by humans or by fully automatic image quality metrics. Experiments show that of five image quality metrics tested, the structural similarity index (SSIM) correlates best with human judgement of image quality, and can be profitably used to cross-validate regularization parameters. However, there is a significant quality gap depending on whether the parameters are cross-validated by humans or with the best image quality metric.

Contents

1 Introduction 2

2 Image Processing By Examples 3

3 Joint Example-based Tikhonov and Total Variation Regularization 3

3.1 Proposed EB-TV Objective Function 4

3.2 Training Data Preparation 5

3.3 Minimizing the EB-TV Objective 5

3.4 Pixel-wise Example-based Regularization 6

3.5 Parameter Setting Via Crossvalidation 7

4 Experiments 7

4.1 Can EB-TV Improve Over TV? 9

4.2 Cross-validating Regularization Parameters Using Human Judgement 10

4.3 Cross-validating Deblurring Parameters Using IQMs 11

4.4 Subjective Experiments Comparing EB-TV to TV with Cross-validated Parameters 11

5 Conclusions 13
1 Introduction

Given the wide availability of high-quality images, many researchers have studied how one can use example images to solve image estimation problems. One of the first solutions to learn from example images was for inverse halftoning [1]: halftoned blocks of an image were mapped to continuous-tone values based on training example pairs of continuous blocks and their halftones. Another early direct use of example images was in restoring lossy-compressed images [2]. Recently, a number of researchers have investigated using example images on a block-by-block basis for single-frame superresolution (e.g. [3–5]). However, this example-based approach can lead to objectionable artifacts as illustrated in Fig. 1.

A more traditional paradigm for image estimation problems is to create prior probabilities over images based on aggregated natural image statistics. For example, researchers have shown that wavelet coefficients of natural images are well-modeled as being drawn from a Laplacian distribution [6], and that the wavelet coefficients tend to be correlated across scales to form what we recognize as edges [7]. Researchers have also hypothesized that a large class of natural images have relatively small total variation, corresponding to predominantly smooth regions separated by sharp transitions [8]. Armed with these statistical models, one can, for example, denoise more effectively by shrinking small wavelet coefficients towards zero [9], and deblur more effectively by regularizing deblurring solutions to have small total variation [8,10,11].

We questioned whether example-based (EB) methods can result in more appealing reconstructions than total variation (TV) regularization, but were unable to find comparative experiments in the literature; in this paper we present subjective experiments to address this question. We hypothesized that example-based regularization would be more effective if combined with TV regularization, and in this paper we present an efficient method for joint example-based total variation (EB-TV) regularization. We investigate how well these methods perform when using regularization parameters trained by maximizing quality on a training set, where quality is measured either by human judgement or by one of five full-reference image quality metrics (IQMs).

In this paper, we report on subjective experiments that compare the proposed EB-TV and TV regularizations for deblurring with optimized regularization parameters, in which regularization parameters are cross-validated by human judges and using full-reference IQMs. The results show that the combined regularization can perform better than either regularizer alone, but is most effective on images with strong structure and given similar example images. In subjective human experiments, we found that parameters can be effectively cross-validated by human judges, but are significantly less optimal when chosen by an IQM; we also found that SSIM correlated best with human judgements of deblurred image quality out of the five IQMs tested for the proposed algorithm.

We hypothesize that the proposed EB-TV regularization could be useful for many different image processing tasks where total variation regularization has proven effective, but in this paper we focus on the problem of reconstructing a high-quality image from an observed blurry image. Our preliminary results in this research area have been presented at a workshop [12].

![noisy blurred image](image1)

![example-based deblurring showing objectionable artifacts](image2)

![example of TV deblurring](image3)

![example of EB-TV deblurring](image4)

Figure 1: A noisy blurred image (far left), an example-based reconstruction that exhibits crisp edges at the expense of objectionable artifacts (second from left), total-variation (TV) reconstruction (second from right), and example-based TV (EB-TV) reconstruction (far left). Regularization parameters for TV and EB-TV were optimized over a finite set of parameter choices by maximizing the average score of three independent observers, as described in Section 4.1.
2 Image Processing By Examples

We build on research in example-based learning for superresolution. For the single-frame superresolution problem, the low-resolution image can be modeled as $y = D\tilde{H}x$, where $x$ is the unknown high-resolution image, $\tilde{H}$ is the ideal anti-aliasing filter, and $D$ is the decimation operator. Consider the non-decimated image $z = \tilde{H}x$ and let $x = x_{LR} + x_{HR}$ be the decomposition of $x$ into the low-resolution part $x_{LR}$ that lies in the row space of $\tilde{H}$ and the high-resolution part $x_{HR}$ that lies in the null-space of $\tilde{H}$ such that $\tilde{H}x = \tilde{H}x_{LR} + \tilde{H}x_{HR} = \tilde{H}x_{LR}$. The goal of single-frame superresolution is to estimate $x_{HR}$ given $y = Dz$, $D$, $\tilde{H}$, and a set of high-resolution example-images. Similarly, example-based deblurring aims to estimate $x_{HR}$ given $z = Hx + w$, $H$, and a set of high-resolution images. Thus, after upsampling $y$ to yield $z$, the superresolution problem can be considered a special case of the deblurring problem, where the noise $w = 0$ and $H = \tilde{H}$.

Freeman et al. proposed that for superresolution, $x_{HR}$ can be estimated by comparing blocks of $y$ (analogously, $z$) to blocks of example images after anti-aliasing and downsampling, finding the nearest neighbor match, and estimating $y$'s missing high frequency information to be the same as the nearest neighbor example image block [4]. Note that estimating high-frequencies from image blocks with similar mid-frequencies can be effective because edges induce correlation of adjacent frequency bins [6, 7]. Since the procedure is performed on a per-block basis, a boundary condition between adjacent blocks was used to regularize the nearest neighbor selection and promote continuity across block boundaries in the reconstructed image. A simple and effective implementation of such a boundary condition is to find the nearest-neighbor image patches in raster-scan order, and overlap each new patch with previously reconstructed image pixels [4].

Freeman et al. noted that this approach of learning from nearest-neighbor patches can overfit to noise in the test image and create objectionable artifacts in the reconstructed image [4]. When we applied the strict analogue of this approach to the deblurring problem, we saw similar objectionable reconstruction artifacts [13]. To reduce artifacts, we have previously proposed an optimal combination of the nearest-neighbor reconstruction with Wiener deblurring by treating them as multiple observations in a linear minimum mean squared error (LMMSE) framework [13]. However, that approach requires impractical knowledge about power spectra, assumes inaccurately that the image statistics are stationary, and focuses on maximizing PSNR rather than visual image quality.

A number of improvements to the basic idea of learning from nearest-neighbor patches have been proposed in the superresolution community. Chang et al. presented a method which requires fewer training samples by posing the nearest-neighbor problem similar to manifold learning [14]. Jiji and Chaudhury proposed a method in which high-resolution detail was learned from a database using contourlets [15]. Jiji et al. evaluated various local and global single-frame superresolution reconstruction techniques [16], and proposed a global PCA-based technique to remove blur, noise, and aliasing artifacts. Ma et al. have shown that example-based superresolution can be effective for error concealment in scalable video coding where previously decoded frames are used to create example images that can be highly similar to the image to be reconstructed [17].

In this paper we explicitly use examples to regularize the deblurring estimate (as opposed to model-based regularizations that use examples to learn the model parameters, such as the work of Zhu and Mumford [18]). Explicit example-based regularization was perhaps first proposed by Baker and Kanade in 2002, who regularized the gradients of their superresolution solution to match the gradient of the corresponding nearest-neighbor from a set of example images [19]; they termed this image hallucination. Datsenko and Elad proposed a regularizer based on an (image-dependent) example-based prior for a global maximum a posteriori objective, and showed good results superresolving scanned documents containing text, equations and graphs [20], and face images [5].

3 Joint Example-based Tikhonov and Total Variation Regularization

We propose a cost function for deblurring that is faithful to the observed image data and uses a joint total variation and example-based regularizer $^1$.

$^1$Matlab code corresponding to the following sections can be downloaded from http://idl.ee.washington.edu/publications.php.

UWEE1011-0005 3
We describe the proposed EB-TV objective in Section 3.1, followed by details of the pre-processing for the example images in Section 3.2, how to efficiently minimize the objective in Section 3.3, a variation of the proposed objective that uses pixelwise weighting of the example-based regularization in Section 3.4, and then a discussion about choosing the method’s parameters in Section 3.5.

3.1 Proposed EB-TV Objective Function

Given a column-scanned $M \times N$ noisy blurred image $z \in \mathbb{R}^{MN}$ and a known blur matrix $H \in \mathbb{R}^{MN \times MN}$, a simplified statement of the proposed objective is that we desire a restored image $x \in \mathbb{R}^{MN}$ that solves:

$$
\min_x \|Hx - z\|_2^2 + \lambda_{EB} \|x - \hat{x}_{EB}\|_2^2 + \lambda_{TV} \|x\|_{TV},
$$

where $\hat{x}_{EB}$ is an example-based reconstruction and $\lambda_{EB} \in \mathbb{R}$ and $\lambda_{TV} \in \mathbb{R}$ are regularization parameters.

However, in preliminary experiments we found that it is better to learn the example-based reconstruction jointly with the estimated $x$, and for efficiency we introduce an auxiliary image (following Chambolle [21]). Thus our proposed cost function solves for a restored image $x \in \mathbb{R}^{MN}$, an auxiliary image $v \in \mathbb{R}^{MN}$, and weights $\{\alpha_i \in \mathbb{R}^k\}_{i=1}^B$ that solve

$$
\min_{x,v,\{\alpha_i\}_{i=1}^B} \|HV - z\|_2^2 + \lambda_1 \|v - x\|_2^2 + 
+ \lambda_2 \left\|x - \sum_i T_i \alpha_i\right\|_2^2 + \lambda_3 \|x\|_{TV},
$$

where $\lambda_1, \lambda_2$ and $\lambda_3$ are regularization parameters, the columns of each $T_i \in \mathbb{R}^{MN \times k}$ contain the $k$ “most similar” $m \times m$ training blocks to the $i$th block of the input image $z$ (with entries outside the block set to zero), the $i$th training blocks are weighted by the corresponding vector $\alpha_i$, and the $\alpha_i$ are jointly solved for to minimize the error between $x$ and the nearest neighbor reconstruction $\sum_i T_i \alpha_i$. The norm $\| \cdot \|_{TV}$ denotes total variation,

$$
\|x\|_{TV} \triangleq \sum_{j,k} \sqrt{|(\nabla_v x)_{j,k}|^2 + |(\nabla_h x)_{j,k}|^2},
$$

where the subscripts $v$ and $h$ denote the vertical and horizontal gradients respectively, and the subscript $(j,k)$ indexes the ($k-1)N+j$th entry of the vector.

The terms in the proposed objective function (1) can be analyzed as follows:

i) The term $\|HV - z\|_2^2$ *faithful to the observations*: it ensures that the auxiliary image $v$ matches the measured signal if blurred by $H$; this is the standard least-squares approach to deblurring. If the noise is assumed to be Gaussian, this term alone would produce the maximum likelihood estimate of the original image.

ii) The term $\|v - x\|_2^2$ is *detail enhancing*: high-resolution information is inferred from middle-resolution information using training examples.

iii) The term $\|x\|_{TV}$ promotes *smoothness* of the reconstruction; TV also promotes sharp edges.

iv) The term $\|x - v\|_2^2$ forces the auxiliary image $v$ and the estimated image $x$ to be close. We use an auxiliary image for computational efficiency, as done by Chambolle [21]. We discuss solving (1) in Section 3.3.

Our example-based regularizer $\|x - \sum_i T_i \alpha_i\|_2^2$ differs from the example-based regularizer proposed by Datsenko and Elad [5,20] in a few respects: (a) we regularize towards a linear combination of neighborhood blocks rather than summing regularizations to each block, (b) we propose a more sophisticated method to adaptively regularize depending on the quality of the examples, see Section 3.4 for details, and (c) we argue that the example-based regularization should be coupled with total variation regularization and our objective promotes the interaction of the example-based regularizer and total variation regularizer, as detailed in the next paragraph.
A key contribution of the proposed objective (1) is the interaction of the \( \{\alpha_i\} \) in the example-based regularization with the other terms caused by jointly minimizing over \( x, v \), and \( \{\alpha_i\} \). Usually in example-based reconstruction, one or more neighbors that are close to \( z \) are solved for and combined in some way. Here, the neighbor blocks \( \{T_i\} \) are fixed based on closeness to \( z \), but the combination weights \( \{\alpha_i\} \) are solved for jointly with \( x \) and \( v \), and the example-based reconstruction \( \sum_i T_i \alpha_i \) is solved to match the reconstruction \( x \). This interaction tempers the propensity of the example-based reconstruction to add artifacts, because the total variation regularization will cause the reconstruction \( x \) to be smoother and have cleaner edges, and thus \( T_i \) that are smoother and have cleaner edges will be rewarded with stronger \( \alpha_i \)’s.

We considered a number of variants of (1) in preliminary experiments including (i) keeping the \( \alpha_i \)’s constant, with values determined by minimizing the distortion between the blurred reconstruction and the measured image: \( ||z - \sum_i HT_i \alpha_i|| \); and (ii) restricting the \( \alpha_i \)’s in (1) to be positive and normalized so that \( \sum_i T_i \alpha_i \) is a convex combination of the neighboring blocks rather than a linear combination. Such variants did not produce better image quality in the preliminary experiments, and were not pursued.

We also found in preliminary experiments that setting the example-based regularization parameter \( \lambda_2 \) once for the entire image was suboptimal because the example-based training blocks can be a good match in some parts of the image but a poor match in other parts, and a \( \lambda_2 \) value that is useful for parts of the image will cause objectionable artifacts in other parts. Thus it is better to set regularization parameters per block, but because the blocks themselves overlap, in Section 3.4 we detail a pixel-wise example-based regularization that reduces objectionable artifacts. Datsenko and Elad also noticed this issue and proposed weighting each block in their block-based regularizer [5, 20].

3.2 Training Data Preparation

To generate \( \{T_i\}_{i=1}^B \), we pre-process training images as prescribed in Freeman et al. [3]: edges are extracted using a high-pass filter, and the images are (locally) contrast normalized. Training image blocks are generated by extracting every \( m \times m \) block from the corpus of training images (we use \( m = 9 \)). For each \( m \times m \) block of the test image in raster-order, we select the \( k = 5 \) (rather than \( k = 1 \) as done in [3, 4]) blocks that, when blurred by \( H \), are the \( k \) nearest neighbors to the test block in terms of Euclidean distance. To promote continuity across block boundaries, we overlap subsequent blocks by one row (or column) of pixels, and use precisely the same distance calculation between blocks as Freeman et al. [3]: the distance between blocks is the Euclidean distance between the blurred blocks plus a scalar (\( \alpha = 0.1 \)) times the Euclidean distance between the current reconstruction estimate for the overlapped pixels and the unblurred training blocks. Each of the \( k \) nearest neighbor training blocks inherit the local contrast and low-frequency information of the test block; this is done so that the low-frequency content of \( x \) is not regularized towards zero for the \( \lambda_2 \) term in (1).

Searching for the nearest-neighbors can be expedited by creating a kd-tree, however this requires extracting, processing and storing every overlapping block from each training image, which leads to redundancy in a straightforward implementation. Instead, we found it quite efficient to exploit the fact that the cross-term of \( ||a - b_i||^2 = ||a||^2 - 2(a, b_i) + ||b_i||^2 \) can be computed by cross-correlation of image block \( a \) with \( \{b_i\}_{i=1}^B \) quickly using FFTs, while the other two terms can be computed once prior to the search.

3.3 Minimizing the EB-TV Objective

Since (1) is jointly convex in its variables, it can be implemented by alternating minimizations over each variable. Specifically, we first optimize over \( \{\alpha_i\}_{i=1}^B \) for fixed \( x \) and \( v \), then optimize over \( x \) and \( v \) for fixed \( \{\alpha_i\}_{i=1}^B \). The latter optimization over \( x \) and \( v \) is accomplished with an inner alternating minimizations loop. That is, (1) is implemented by solving

\[
\begin{align*}
\text{Step 1: } & \min_{\{\alpha_i\}_{i=1}^B} \|x - \sum_i T_i \alpha_i\|^2 \\
\text{Step 2a: } & \min_x \lambda_1 \|x - v\|^2 + \lambda_2 \|x - \sum_i T_i \alpha_i\|^2 \\
& + \lambda_3 \|x\|_{TV} \\
\text{Step 2b: } & \min_v \|Hv - z\|^2 + \lambda_1 \|x - v\|^2,
\end{align*}
\]
where each bracket denotes a loop of alternating minimizations. We next detail how we solve the subproblems that arise in each of the steps.

Since the blocks \( \{ T_i \} \) overlap, optimizing Step 1 requires that all \( \{ \alpha_i \}_{i=1}^T \) be found jointly using, e.g., a Markov network. However, for the related example-based superresolution problem, Freeman et al. found that a raster-order method performed well compared to a Markov network \cite{4}. With that motivation, we similarly solve Step 1 in (2) by solving for each block in raster scan order. Then solving for each \( \alpha_i \) is equivalent to solving a least-squares linear regression problem, and thus each \( \alpha_i \) has a closed-form solution \( \alpha_i = (T_i^T T_i)^{-1} T_i^T x \) (solved efficiently by exploiting the sparsity of \( T_i \); the only nonzero entries of \( T_i \) correspond to the support of the \( i \)th block).

We derived a solution for Steps 2a and 2b in (2) by extending recent work on efficient total variation deblurring. Chambolle showed that the solution to

\[
\min_x \frac{1}{2} \| x - g \|^2 + \gamma \| x \|_{TV}
\]

where the subscript \( T \) blocks match image blocks in the image reconstruction especially when \( \gamma \) is large. The severity of the artifacts depends on how well the training \( z \), based on the residual image \( r = z - Hf \), where \( f = \sum_i T_i \alpha_i \). For pixels where \( z - Hf \) is large, we wish to decrease the effect of the example-based regularization in favor of the total-variation regularization. To this end, we modify (5) so that the update on \( x \) is on a per-pixel basis:

\[
x_i = \frac{\lambda_1 v_i + \lambda_2 f_i}{\lambda_1 + \lambda_2} - \frac{\lambda_3}{2(\lambda_1 + \lambda_2)} (\nabla \cdot p)_i,
\]

where the subscript \( i \) denotes the \( i \)th element of the vector. In our experiments, we form \( \lambda_2 \) by starting with the residual image \( r = z - Hf \), transforming each pixel’s residual into a weight \( q_i = e^{-50r^2} \), then spatially
smoothing the image \( q \) by convolving it with a spatial filter \( s = \begin{pmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{pmatrix} \), resulting in the pixel-wise regularization parameter \( \lambda_2 = \lambda_2(q ** s) \), where ** denotes spatial (two-dimensional) convolution. We found that the performance of this pixel-wise regularizer was robust to the exact specification of the \( q \) and \( s \).

### 3.5 Parameter Setting Via Crossvalidation

Adjusting the values for parameters \( \lambda_1 \), \( \lambda_2 \) and \( \lambda_3 \) in (1) can lead to a variety of reconstructions. For \( \lambda_2 = \lambda_3 = 0 \), equation (1) results in a least-squares reconstruction \( x = v \); if only \( \lambda_2 = 0 \), then \( x \) is the TV reconstruction as implemented by Bresson and Chan [10]; and \( \lambda_1 = \lambda_3 = 0 \) produces a blockwise \( k \)-nearest neighbor reconstruction \( x \), analogous to [4]. For many cases we hypothesize that the best reconstruction will have \( \lambda_1 > 0, \lambda_2 > 0, \lambda_3 > 0 \); we test this hypothesis in the next section. However, manually finding the optimal choice of parameters for every blurred image is impractical for many applications.

We hypothesize that it is possible to learn a good set of parameter choices for an image by finding good parameters for similar images and that similar images are required to train the parameters because of two issues. First, since the optimal value of \( \lambda_2 \) depends on how well local edge content in the test image is captured by training blocks extracted from the training set—in order to tune the relative strengths of \( \lambda_1, \lambda_2, \lambda_3 \) it would be best if the test and training images have similar edge content, e.g., straight edges for structural images. Second, human perceptual quality is generally scene dependent. Thus, an optimal choice for \( \{ \lambda_1, \lambda_2, \lambda_3 \} \) on a structural image may not be appropriate for an image of a human face as there will be different perceptual tolerance for noise, edge sharpness, etc. Thus, images used to train the deblurring parameters should be similar to the test image in a psycho-visual sense.

### 4 Experiments

We conducted controlled subjective experiments to determine whether EB-TV deblurring is preferable to TV deblurring, to investigate how well deblurring parameters for each method can be optimized by cross-validation on a training set, and finally, to determine how well IQM-optimized parameters compare to human-optimized parameters.

Blind subjective experiments were done with two sets of subjects: three image processing experts (the authors: one female and two males, all thirtysomething), and a set of sixteen observers consisting of six females and ten males, aged between 21 and 40 years, all with normal or corrected-to-normal color vision. All experiments took place in a room with dimmed D50 lighting, using a calibrated NEC MultiSync LCD 2490WUXi monitor in sRGB mode, and with the observer sitting at roughly 18 inches from the monitor.

We selected training images from the Kodak benchmark dataset that contain similar content where similar is defined by coarse image content categories such as faces and houses. Training images and test images are shown in Fig. 2. The training images shown in Fig. 2 are also used as examples to generate \( \{ T_i \}_{i=1}^B \) for example-based regularization. In preliminary experiments we studied the effect of the number of training images and found that using a fewer number of training images results in greater diversity in the nearest neighbors. Especially for severe blur, when a large number of images is used, the nearest neighbor search leads to overfitting, and the reconstruction contains objectionable artifacts.

We employ leave-one-out cross-validation to optimize \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) on the training set. Leave-on-out cross-validation ensures that the nearest neighbor blocks \( T_i \) for a given training image are extracted from the other images.

As detailed in the next section, we tested our hypothesis about parameter-training using humans to score training image reconstructions over a set of parameter choices. In addition, we investigated how well it works to learn parameters automatically by using a full-reference IQM to rate the performance of different parameter choices on the training images.

For all experiments, each sharp image was blurred by convolving it with a Gaussian blur kernel with a 2.25 pixel bandwidth and adding Gaussian white noise such that the PSNR with respect to the original sharp image was 34 dB PSNR. For the reconstructions where \( \lambda_2 > 0 \), example image blocks were extracted from all of the images in the training set if reconstructing a test image, or if a training image was being
reconstructed (to cross-validate parameters), then the example image blocks were extracted from only those images in the training set that were not being reconstructed (that is, at no point did we deblur an image using examples from the same image).

For all experiments, each color plane of a high-quality 24 bit RGB color image was independently subjected to blur and noise. Wen et al. have investigated color-plane-dependent TV deblurring [11], but to simplify our experiments, we converted the corrupted RGB image to the YCbCr colorspace, and only deblur the Y plane. Because the Y plane roughly corresponds to human perception of luminance and luminance edges are most visible, this is a reasonable approximation to full deblurring. We chose the YCbCr colorspace over a more perceptually accurate colorspace such as CIELAB because the Y channel is a linear function of the original RGB channels, and thus the blur affecting Y is still $H$. After deblurring the Y channel, we converted the YCbCr image back to an RGB image, and then all visual comparisons were performed with RGB images. However, the IQM’s only used the Y color plane as specified below.

Figure 2: Test images **house**, **boat** and **face** (top), **structure** training set (middle), and **human** training set (bottom); all images are $768 \times 512$ pixels.
4.1 Can EB-TV Improve Over TV?

Our first experiment was to determine whether the EB-TV algorithm can produce better image quality than TV alone if a human chooses the deblurring parameters to maximize quality on the test image. As noted previously, we hypothesize that EB-TV may perform better when the test image is similar to the training images both in structural content, and in a psycho-visual sense. Therefore, we reconstructed the test images using the structure and human training sets in Fig. 2. For EB-TV and TV, we restricted our attention to a finite set of possible parameter values $\lambda_1, \lambda_2$, and $\lambda_3$: $\lambda_1 \in \{0.1, 0.5, 1\}$, $\lambda_2 \in \{0, 0.001, 0.1\}$ and $\lambda_3 \in \{0, 0.001, 0.1\}$; these values were chosen from a large set of possible parameter values to be those parameters that produced perceptually different reconstructions in a preliminary experiment.

For each of the three test images in Fig. 2, we generated a human rating for the 36 possible reconstructions (all combinations of $\lambda_1, \lambda_2$ and $\lambda_3$), once using the structure training set, and once using the human training set. Furthermore, we generated scores for each of the training images; for EB-TV, we employ the leave-one-out method, in which one training image is reconstructed using $k$-NN training blocks extracted from the remaining images in the training set. In all, there were a total of 11 reconstruction scenarios, each with 36 image reconstructions.

To generate human scores, three expert observers independently performed a blind experiment on the 36 × 11 images. For each of the 11 experiments, the 36 reconstructions were ordered randomly, and each of the three experts independently scored each of the reconstructions according to the following categories: very good (score=5), good (score=4), fair (score=3), unsatisfactory (score=2) and objectionable (score=1). Scoring was averaged across the three human observers to yield the “human score.” Although the three expert scores were highly correlated, they did not agree exactly as the experts differed in their sensitivity to blur, to overall noise, and to objectionable artifacts (of the type shown in Fig. 1, second from right).

An example noisy blurred image and three of the reconstructions are shown in Fig. 1 that used the structure example images. The middle-left image shows one of the example-based reconstructions with $\lambda_2 = 0$ such that there is no total variation. This creates objectionable artifacts in the reconstruction of the numbers on the sail. The middle-right image shows the TV reconstruction with the highest quality as judged by the human experts. The far-right image shows the EB-TV reconstruction with the highest quality as judged by the human experts. The EB-TV reconstruction is sharper but noisier.

Fig. 3 shows a scatterplot of the human score for each test image (house, boat and face) in the three columns versus the human scores averaged over each training set (structure and human) in the two rows. Each of the markers in the plot correspond to one of the 36 sets of parameter choices; note that close inspection reveals that some markers (parameter choices) overlap. The red squares correspond to parameter choices where only $\lambda_3 = 0$ (TV regularization), the green diamonds correspond to parameter choices where $\lambda_3 = 0$ (EB regularization with no TV regularization), and the blue triangles mark parameter choices that use both EB regularization and TV regularization.

To understand the figure, consider first the top-left plot of Fig. 3. The x-axis shows how each parameter choice was scored, averaged over the three human experts and averaged over the three structure training images (where for each structure training image the example-based term was trained on the other two structure images). One sees that the highest score was achieved with a parameter choice marked by a blue triangle, which signifies that it was a parameter choice that used the EB-TV regularization (that is, $\lambda_2, \lambda_3 > 0$). The y-axis shows how each parameter choice was scored by the three human experts on the test image house when the structure training images were used for the example-based term in (1). Again, the parameter choice that yielded the highest score on the house image was marked by a blue triangle signifying $\lambda_2, \lambda_3 > 0$.

Examining the rest of the plots in Fig. 3 shows that the parameter set that produced the highest-quality boat reconstruction given the structure images did not use example-based regularization (top, middle), but the highest-quality face reconstruction did (top, right). The bottom row of plots shows the results from using the human images for the example-based term. Two parameter choices tied for the highest score averaged over the human training images, one parameter choice had only used TV (red square) and the other parameter choice used both EB and TV regularization (blue triangle). On the house test image the best score only used TV. However, on the boat and face test images, the best scores were achieved using the EB regularization. Notably, on the face test image, the best score achieved with only total variation is a 3$\frac{1}{2}$, whereas nine parameter choices that used the example-based regularization scored 3$\frac{2}{3}$ to 4$\frac{1}{2}$. 
Overall, we see that sometimes the example-based regularization can yield better performance than total variation alone, and not surprisingly the example-based regularization seems to be more useful when the test image is better matched to the training images. Next, we investigated which method performed better when the parameter choices were cross-validated on similar or dissimilar images.

### 4.2 Cross-validating Regularization Parameters Using Human Judgement

We used the 11 × 36 averaged expert-scores described in the previous subsection to investigate the hypothesis that effective regularization parameters for deblurring a test image can be reliably selected via cross-validation on the training set, as long as the test image is similar to the training set.

Consider again Fig. 3, but now focusing on the correlation between the two dimensions in each plot. Reconstruction parameters for test images house and boat exhibit high correlation with reconstruction parameters for the structure training set, and worse correlation for the human training set. Conversely, face exhibits better correlation with human than with structure. This results supports our hypothesis that if a test image is similar to the training set, parameters preferred for image reconstruction of training images may be applied to the test image to yield a reconstruction close to that produced by the optimal parameter set for that test image.

In Section 4.4, we test whether EB-TV regularization is preferred to TV-regularization when the parameters are cross-validated.
4.3 Cross-validating Deblurring Parameters Using IQMs

For some applications it may be desirable for the cross-validation of the parameters to be fully automatic, to that end we consider whether a full-reference IQM can be used to assess the best parameter choices. We evaluated five candidate IQMs: PSNR, the 95% highest error as measured by $\Delta E$ CIELAB, the 95% highest error as measured by $\Delta E$ of s-CIELAB [23], structural similarity index (SSIM) [24], and visual information (VIF) [25]. The $\Delta E$ CIELAB and sCIELab measures act on the $L$ channel of an sRGB-to-CIELab transformation. The SSIM and VIF measures act on the $Y$ channel after an RGB-to-YCbCr transformation. The PSNR measure acts on the full RGB image.

The scatterplot in Fig. 4 shows how IQM scores correlate with human scores when averaged over the training set. It would be desirable for PSNR, SSIM and VIF to have a strong positive correlation and 95% $\Delta E$ and 95% s-$\Delta E$ to have a strong negative correlation (since low $\Delta E$ scores are desirable). However, only SSIM exhibits strongly favorable correlation. In fact, for the human training set (bottom half of Fig. 4), the other IQMs give good scores to images that received poor human scores.

To determine which IQM performs closest to the human observers, we consider two metrics. First, the correlation between each IQM and the average human score gives a global indication of how well the IQM mimics human performance. However, since we will select $(\lambda_1, \lambda_2, \lambda_3)$ that achieve the best IQM score, we are especially interested in whether the highest-ranking IQM parameters correspond to high-ranking human scores. Thus for a second metric, we compute the average human score of the top five IQM-selected parameters. For both metrics and for both the structure and human training sets, SSIM performs better than the other IQMs.

Based on these results, in the subjective experiments detailed in the next section comparing EB-TV and TV with cross-validated parameters, we used the SSIM cross-validated parameters. That is, we select the parameter set $(\lambda_1, \lambda_2, \lambda_3)$ for which SSIM is maximized on the training set of images. This parameter set is then applied to the test image for reconstruction. For this, we require knowledge of the blur kernel $H$ and the noise power so that for cross-validation, training images can be corrupted to mimic the corruption of the test image.

4.4 Subjective Experiments Comparing EB-TV to TV with Cross-validated Parameters

Given cross-validated parameters chosen by humans or SSIM, we performed blind subjective tests to evaluate how the EB-TV deblurring given by (1) compares to TV deblurring given by (1) with $\lambda_2 = 0$. The parameters selected by optimizing the human score and SSIM for each algorithm are shown in Table 1.

The forced-choice pair comparison protocol was used, where each image, observers compared four reconstructions of each of the three test images using two different training sets separately, so that there were $2 \times 3$ separate experiments each comparing four reconstructions, for a total of 24 image pairs evaluated by each subject. Sixteen observers, six females and ten males, aged between 21 and 40 years, participated in this experiment, all with normal or corrected to normal color vision.

In the beginning of the experiment, the observers were shown the original (non-blurred) images as a reference, while during the comparison only each pair of reconstructed images to be compared were shown at any one time. The observers were informed that the images they would be comparing were high-quality images that had been subjected to blur and noise and then processed with different deblurring methods, and they were instructed to select the reconstruction in each pair that they preferred. All pairs were presented twice with the two reconstructions on different sides of the monitor to avoid possible systematic errors in which people prefer one side of the monitor to the other when reconstructions seem indistinguishable. The observers’ responses were analyzed according to Case V of Thurstone’s Law of Comparative Judgement in order to obtain Z-scores [26,27].

Results for this experiment are given in Fig. 5. The left column shows results using the structure training images. EB-TV is statistically significantly better than TV when both use human cross-validated parameters on the house and boat images, which are a better psychovisual match to the structure training images than the face image. On the face image, the performance of EB-TV and TV was not statistically significantly different when trained with the structure images. However, when trained on human, the EB-TV solution is preferred to the TV solution, but there are no statistically significant differences between
Figure 4: Average human score versus IQM score over training set structure (top) and human (bottom). The marks denote images for which $\lambda_2 = \lambda_3 = 0$ (●), $\lambda_2 = 0$, $\lambda_3 > 0$ (■), $\lambda_2 > 0$, $\lambda_3 = 0$ (♦), and $\lambda_2, \lambda_3 > 0$ (▲). The table lists the correlation coefficient (negative correlation for the $\Delta E$ metrics) and the average human score of the top 5 IQM parameters.
EB-TV and TV with human cross-validated parameters.

Fig. 5 shows that SSIM cross-validated parameters almost always led to worse results than using human cross-validated parameters, and sometimes dramatically worse. EB-TV and TV trained with SSIM cross-validated parameters are only statistically significantly different for the case of face and house trained with structure, and in both cases EB-TV is statistically significantly better.

Most observers used between 5 and 15 minutes to evaluate the 2x4x3=24 image pairs, while two observers used as much as 30 minutes. When asked afterwards to comment on how they made their judgements, three main attributes were mentioned: sharpness, presence of artifacts (described as jagged edges/ghosting/ringing), and noise (described as graininess/dottedness/pockiness). Typically the observers reported making tradeoffs between these attributes. Notably, according to these post-experiment interviews, there was a large variation in how the observers ranked the different attributes.

5 Conclusions

In this paper we proposed a combined example-based and total variation regularization (EB-TV) for reconstructing images, and conducted subjective experiments to determine if EB-TV could offer better performance than TV alone, and how well the regularization parameters for both methods can be learned by cross-validation, either with human or IQM scores. Our experiments suggest that good deblurring parameters are learnable by cross-validation with human scores, but more so if the training and test data are well-matched, and even more so for EB-TV if the images have structure. Of five IQM’s, we found that SSIM correlated best with human judgements of quality of the reconstructions for all images. However, our concept of well-matched was based on a coarse categorization, and quantifying exactly how the training test should be matched is an open question.

Comparing the EB-TV deblurring to TV deblurring, we found that the EB-TV deblurring was significantly better than TV deblurring when the test and training images were matched in structural content and given parameters cross-validated by humans. However with SSIM-chosen parameters, EB-TV deblurring was only statistically significantly better given structured training examples and the house or face test images, and was otherwise not statistically different than TV. Overall, the parameters chosen by SSIM yielded much lower quality reconstructions than the parameters chosen by humans.

We found the results depended heavily on the training set. We believe there are two main reasons the training set matters. First, different training sets may have different amounts of edge content. Second, as shown in Fig. 3, the appropriate set of parameters for different images may differ; we hypothesize that this is primarily because humans judge image quality differently depending on the image content, for example noise appears to be judged more harshly in face images, and that affects the choice of regularization parameters. We did not try to quantify how the training set should match the test set, which is an important open question for future research.

Although we focused on known blur here, we hypothesize that the proposed EB-TV regularizer may be useful for many image processing problems where total variation has already been shown to be effective.
Acknowledgements

This work was supported by the United States Office of Naval Research and the United States PECASE award.

References


Figure 5: Results of the psychophysical experiment. Overall, the proposed algorithm is preferred over TV with parameters chosen by a human during crossvalidation. When SSIM is used as a surrogate, our method is preferred twice, and otherwise statistically insignificant to TV.