Perceptually Based Downscaling of Images

A. Cengiz Öztireli
ETH Zurich

Markus Gross
ETH Zurich

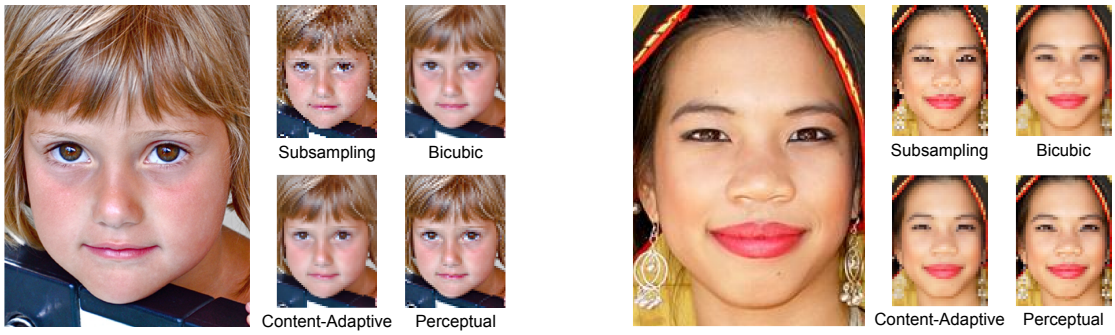

Figure 1: *It is challenging to reproduce a perceptually similar downscaled version of an image, as shown here for simple subsampling, the commonly used bicubic filter, and the state-of-the-art method of Kopf et al. [2013] ("Content-Adaptive"). Relying on a perceptual image quality measure instead of standard metrics, our method ("Perceptual") is able to preserve perceptually important features and the overall look of the original images. Left input image courtesy of Flickr user Matteo Catanese.*

Abstract

We propose a perceptually based method for downscaling images that provides a better apparent depiction of the input image. We formulate image downscaling as an optimization problem where the difference between the input and output images is measured using a widely adopted perceptual image quality metric. The downscaled images retain perceptually important features and details, resulting in an accurate and spatio-temporally consistent representation of the high resolution input. We derive the solution of the optimization problem in closed-form, which leads to a simple, efficient and parallelizable implementation with sums and convolutions. The algorithm has running times similar to linear filtering and is orders of magnitude faster than the state-of-the-art for image downscaling. We validate the effectiveness of the technique with extensive tests on many images, video, and by performing a user study, which indicates a clear preference for the results of the new algorithm.

CR Categories: I.4.1 [Image Processing and Computer Vision]: Digitization and Image Capture—Sampling;

Keywords: images, video, perceptual, downscaling, structural similarity, unsharp masking

1 Introduction

Image downscaling is a fundamental operation performed constantly in digital imaging. The abundance of high resolution capture devices and the variety of displays with different resolutions make it an essential component of virtually any application involving images or video. However, this problem has so far received substantially less attention than other sampling alterations.

Classical downscaling algorithms aim at minimizing aliasing artifacts by linearly filtering the image via convolution with a kernel before subsampling and subsequent reconstruction, following the sampling theorem [Shannon 1998]. However, along with aliasing, these strategies also smooth out some of the perceptually important details and features, as shown in Figure 1, since the kernels used are agnostic to the image content. A solution to this problem is adapting the kernel shapes to local image patches [Kopf et al. 2013] in the spirit of bilateral filtering [Tomasi and Manduchi 1998], so that they are better aligned with the local image features to be preserved. This strategy can significantly increase the crispness of the features while avoiding ringing artifacts typical for post-sharpening filters. However, it still cannot capture all perceptually relevant details, and as a result, might distort some of the perceptually important features and the overall look of the input image (Figure 1, content-adaptive), or lead to artifacts such as jagged edges [Kopf et al. 2013].

Loss of some of the perceptually important features and details stems from the common shortcoming of these methods that they operate with simple error metrics that are known to correlate poorly with human perception [Wang and Bovik 2009]. Significant improvements have been obtained for many problems in image processing by replacing these classical metrics with perceptually based image quality metrics [Zhang et al. 2012; He et al. 2014].

In this paper, we propose a perceptually based method for downscaling images. We formulate the downscaling problem as an optimization where we solve for the downscaled output image given the input image. The error between the two images is measured using the widely adopted structural similarity (SSIM) index [Wang et al. 2004]. The use of SSIM in optimization problems has been hindered by the resulting non-linear non-convex error functions [Brunet et al. 2012]. However, we show that for the downsampling problem, it is possible to derive a closed-form solution to this optimization. The solution leads to a non-linear filter, which involves computing local luminance and contrast measures on the original and a smoothed version of the input image. Although the filter is seemingly different than SSIM without any covariance term, we show that it maximizes the mean SSIM between the original and

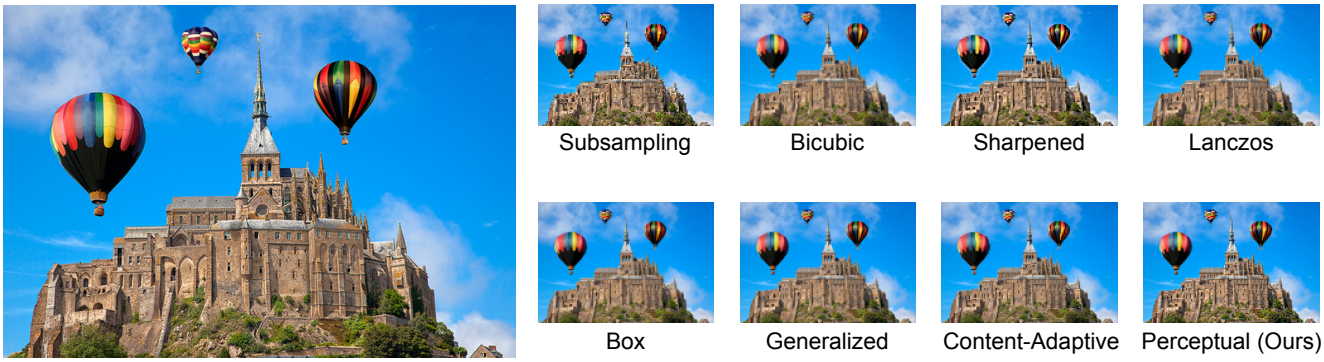

Figure 2: *Commonly used filters for downscaling such as the box or bicubic filter result in oversmoothing. Trying to avoid oversmoothing by post-sharpening the downscaled images or using the Lanczos filter can lead to ringing artifacts and the small-scale features can still not be recovered. The recent methods, generalized sampling [Nehab and Hoppe 2011] and content-adaptive downscaling [Kopf et al. 2013] can produce crisper images, but cannot preserve all perceptually important details. In contrast to the previous methods, we utilize a perceptual metric and generate a perceptually optimum image as measured by this metric. This makes our results look closer to the original high resolution images by retaining apparent details. Input image courtesy of Flickr user Nicolas Raymond.*

downscaled images.

Our downscaling method preserves perceptually important fine details and features that cannot be captured with other metrics, resulting in crisper images that provide a better depiction of the original image. The downscaled images do not exhibit disturbing aliasing artifacts for natural images and are spatio-temporally more coherent than methods based on kernel optimizations [Kopf et al. 2013]. This allows us to apply the technique to video downscaling as well. The resulting algorithm has a very simple, efficient, and parallelizable implementation with sums and convolutions. It thus has a computational complexity similar to the classical filtering methods, and runs orders of magnitude faster than the state-of-the-art [Kopf et al. 2013]. We illustrate that our method significantly improves the quality of downscaled images by comparisons to a variety of classical and state-of-the-art downscaling methods on many images. We validate our results with a user study, which indicates a clear preference for our algorithm over previous methods.

2 Related Work

2.1 Image Downscaling

The standard approach to image downscaling involves limiting the spectral bandwidth of the input high resolution image by applying a low-pass filter, subsampling, and reconstructing the result. As is well-known in signal processing, this avoids aliasing in the frequency domain and can be considered optimal if only smooth image features are desired. Approximations of the theoretically optimum sinc filter such as the Lanczos filter, or filters that avoid ringing artifacts such as the bicubic filter are typically used in practice [Mitchell and Netravali 1988]. However, these filters often result in oversmoothed images as the filtering kernels do not adapt to the image content. The same is true for more recent image interpolation techniques [Thévenaz et al. 2000; Nehab and Hoppe 2011].

Recently, Kopf et al. [2013] showed that significantly better downscaling results with crisper details can be obtained by adapting the shapes of these kernels to the local input image content. Since the kernels better align with the features in the input image, they capture small scale details when present. However, the method does not take perceptual importance of the features into account, resulting in loss of apparent details and hence leading to a rather abstract view of the input image. Indeed, the method is shown to provide excellent results for generating pixel-art images [Kopf et al. 2013].

For natural images, we show that significantly better and crisper depictions of a high resolution input image can be obtained by incorporating a perceptual metric. Our method also has better spatio-temporal consistency with less apparent aliasing artifacts, and runs orders of magnitude faster with a simple and robust implementation.

Downscaling operators are also designed for other related problems. Several algorithms carefully tune the downscaling operators and filters to the interpolation method used for subsequent upscaling [Wu et al. 2009; Zhang et al. 2011; Dong and Ye 2012]. Unlike these methods, we are interested in the perceptual quality of the downscaled image itself. Thumbnail generation tries to preserve the imperfections, in particular blurriness in the original images for accurate quality assessment from the downscaled images [Trentacoste et al. 2011; Didyk et al. 2012]. In contrast, the downscaling problem can be regarded as selectively adjusting the blur to preserve the important details and overall look of an input image [Kopf et al. 2013]. Another related set of algorithms deals with retargeting images by changing the aspect ratios of input images [Banterle et al. 2011], while preserving important parts such as foreground objects in the image by carefully modifying the image content. Instead, we would like to keep the image content as close as possible to that of the original image, and target resolution reductions far more than the retargeting algorithms are designed for. Finally, image abstraction methods can be used to generate artistic depictions of an input image such as via pixel art [Gerstner et al. 2012] by reducing the resolution as well as the color palette. We instead target realistic depictions of the input image.

2.2 Image Quality Metrics and SSIM

Standard error metrics such as the mean squared error is well-known to correlate poorly with human perception when measuring image differences [Wang and Bovik 2009]. Instead, for the assessment of the quality of images and video, a variety of perceptually based image quality metrics has been proposed. Full reference quality metrics refer to the assumption that an input image can be compared to an available reference image for quality assessment. For the downscaling problem, we interpret the input image as the reference, and the downscaled output as the image to be assessed. Please refer to the recent survey papers for an overview of full reference image quality assessment metrics [Zhang et al. 2012; He et al. 2014].

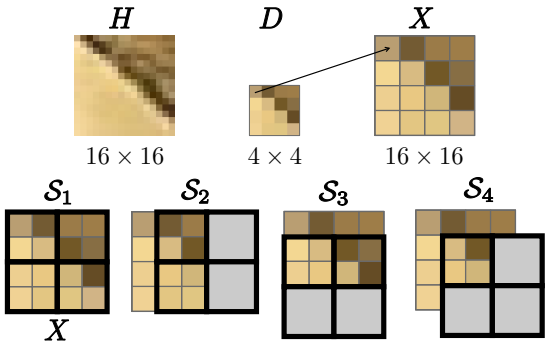

Figure 3: *(Top) The input high resolution image H (16×16 pixels), the downscaled image D (4×4 pixels), and its scaled version X (16×16 pixels). Each pixel of D is replicated in 16 pixels of X. (Bottom) Different patch sets $\mathcal{S}_k$ for this example with $n_p = 4$ (2×2 patches on D). Each patch set $\mathcal{S}_k$ contains patches that do not overlap each other. The patch sets are shifted by 4 pixels in X and H, which corresponds to a shift of 1 pixel in D.*

We utilize the structural similarity (SSIM) index [Wang et al. 2004], which is one of the most widely used and successful full reference image quality metrics [Brunet et al. 2012]. SSIM computes a matching score between two images by local luminance, contrast, and structure comparisons. Despite its simplicity, it performs consistently well on image quality assessment tests and is thus widely used [Zhang et al. 2012] with many modifications and extensions [Wang and Bovik 2009].

An intuitive idea for image processing problems involving error minimization is replacing the usual mathematical metrics with perceptually based metrics such as SSIM. This idea has been explored in some works for problems such as halftoning [Pang et al. 2008], image denoising [Channappayya et al. 2006; Channappayya et al. 2008c; Rehman et al. 2011; Shao et al. 2014], inpainting [Ogawa and Haseyama 2013], superresolution [Zhou and Liao 2015], quantization [Channappayya et al. 2008b], and coding [Wang et al. 2011]. However, in spite of the popularity of SSIM as a quality measure, its wider acceptance as a standard metric for image processing tasks has been hindered by the resulting non-linear non-convex optimization problems [Channappayya et al. 2008c; Brunet et al. 2012].

This has led to proposing application dependent assumptions on local or global image properties [Channappayya et al. 2006; Channappayya et al. 2008c; Brunet et al. 2010; Zhou and Liao 2015], and approximations of the SSIM measure [Brunet et al. 2012]. The resulting quasi-convex or convex problems are typically solved via iterative optimization methods [Channappayya et al. 2008c; Rehman et al. 2011; Ogawa and Haseyama 2013; Brunet et al. 2012], although a few works show that careful selection of simplifying assumptions can lead to closed-form solutions [Channappayya et al. 2006; Brunet et al. 2010]. We formulate image downscaling as an optimization problem with SSIM as the error metric. Similar to previous applications, this gives our method a significant advantage for preserving perceptually important features. In contrast to many other applications, we show that a closed-form solution can be derived for the downscaling problem.

3 Perceptual Downscaling Method

Given a high resolution input image H, our goal is to find the downscaled output image D that is as close as possible to H as measured by the SSIM index. We denote the dissimilarity measure between images H and D with $d(H, D)$. We would like to get the image D^* that minimizes this measure. We assume images with a single channel such that each pixel of H and D contains a single number in the dynamic range $[0, 1]$, and further assume for simplicity that the width and height of H is downscaled by an integer factor s to produce D. If the actual downscaling factor is not an integer, we upscale the input image by bicubic filtering such that the factor becomes an integer.

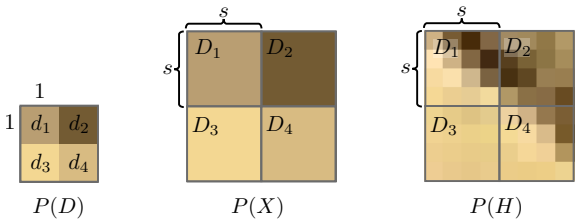

Figure 4: *Each pixel d_i in the output downscaled image patch $P(D)$ is mapped to a set D_i of s^2 pixels in the patches $P(X)$ and $P(H)$. All s^2 pixels in D_i of $P(X)$ have value d_i.*

3.1 The Downscaling Problem

Most image quality assessment measures are not designed to compare images of different spatial resolutions [Yeganeh 2014]. For images of different resolutions, there are two common simple approaches: downscaling the higher resolution image, or upscaling the lower resolution one [Demirtas et al. 2014]. Since we do not want to lose the information present in H, we upscale D. We call this upscaled image X, which has the same dimensions as H. Assuming a piecewise constant interpolation, each pixel of D is replicated in s^2 pixels of X (Figure 3, top).

The SSIM index is a local measure of similarity computed between local patches of images. These similarity scores are then summed for all patches to compute the mean SSIM. Denoting the i^{th} patch of an image X by $P_i(X)$, the downscaling problem can thus be written as finding the optimum X^* that satisfies

$$X^* = \underset{X}{\operatorname{argmin}} \sum_{P_i \in \mathcal{S}} d(P_i(H), P_i(X)), \tag{1}$$

for some set $\mathcal{S}$ of patches, with the constraint that each group of pixels of X that corresponds to a single pixel of D has the same pixel value. Note that we do not restrict the pixel values of X to be in $[0, 1]$. We will see in Section 3.3 that the optimized D contains a small number of pixels negligibly outside of the dynamic range.

The shapes and the set $\mathcal{S}$ of the patches can be defined in various ways, depending on the application considered [Silvestre-Blanes 2011]. For a given patch size n_p, we propose to use the set $\mathcal{S}$ of all possible square patches of width (and height) $s\sqrt{n_p}$ (excluding the patches not completely within image limits), but in patch sets $\mathcal{S}_k$, such that each $\mathcal{S}_k$ contains only non-overlapping patches, and $\mathcal{S} = \bigcup \mathcal{S}_k$. The final X^* is computed by averaging the solutions X_k^* of the problem in Equation 1 for different $\mathcal{S}_k$. Since each group of s^2 pixels in X actually corresponds to a single pixel in D, integer patch shifts in D leads to shifts by s in H and X. The patch sets S_k for a small example image with $n_p = 4$ are shown in Figure 3, bottom. We will elaborate in Section 3.3 that the solution does not deviate much for other choices of patch sets, while n_p should be chosen carefully.

Since the patches in $\mathcal{S}_k$ do not overlap, the pixels of each patch can be optimized independently of the other patches in $\mathcal{S}_k$. Hence, for a

patch $P \in \mathcal{S}_k$, the optimum patch $P^*(X)$ of the image X is given by

$$P^*(X) = \arg\min_{P(X)} d(P(H), P(X)). \quad (2)$$

We stack the pixels of the patches into the vectors $\mathbf{h}$ and $\mathbf{x}$. Similarly, we represent the pixels of D that correspond to $\mathbf{x}$ with $\mathbf{d}$, and denote the set of pixels in $P(X)$ that corresponds to the i^{th} pixel in the patch in D with D_i (Figure 4). Hence, $\mathbf{x} = \mathbf{V}\mathbf{d}$, where the j^{th} component of each column vector $\mathbf{v}_i$ of $\mathbf{V}$ is 1 if $x_j \in D_i$, and 0 otherwise. Then, the above problem can be rewritten in the following form

$$\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} d(\mathbf{h}, \mathbf{x}), \quad \mathbf{x} = \mathbf{V}\mathbf{d}. \quad (3)$$

3.2 Optimization with SSIM

The SSIM index is computed by multiplying three components corresponding to luminance, contrast, and covariance based comparisons. The widely used form of SSIM is given by [Brunet et al. 2012]

$$SSIM(\mathbf{h}, \mathbf{x}) = \frac{(2\mu_h\mu_x + c_1)(2\sigma_{xh} + c_2)}{(\mu_h^2 + \mu_x^2 + c_1)(\sigma_h^2 + \sigma_x^2 + c_2)}, \quad (4)$$

where $\mu_x = \sum w_i x_i$ denotes the mean, $\sigma_x^2 = \sum w_i (x_i - \mu_x)^2$ the variance, and $\sigma_{xh} = \sum w_i (x_i - \mu_x)(h_i - \mu_h)$ the covariance with weights w_i, and x_i denoting the i^{th} component of $\mathbf{x}$. The c_1 and c_2 are small constants added to avoid instability. For the simplicity of the expressions, and since the small values used in practice do not affect our results for the downscaling problem, we will assume that the constants are selected as $c_1 = c_2 = 0$. Since x_i and h_i are in $[0, 1]$, $SSIM(\mathbf{x}, \mathbf{h}) \in [0, 1]$. It is 1 when $\mathbf{x} = \mathbf{h}$, and decreases as the patches become less similar. We thus define the dissimilarity measure as $d(\mathbf{h}, \mathbf{x}) = 1 - SSIM(\mathbf{h}, \mathbf{x})$.

The $d(\cdot, \cdot)$ is not a distance function, and not even convex. Instead of directly trying to solve the problem in Equation 3, we thus define another problem that is easy to be solved, by parametrizing the solution to the original problem. Specifically, we fix the mean μ_x and variance σ_x of $\mathbf{x}$ to arbitrary values, leaving only σ_{xh} as the free term in SSIM (Equation 4). We thus optimize for σ_{xh} under these constraints to get the optimum for this subproblem. Finally, we find the μ_x and σ_x that gives the global optimum. In Appendix A, we show that the global optimum is obtained by setting $\mu_x = \mu_h$, and $\sigma_x = \sigma_h$, and solving the resulting problem:

$$\max_{\mathbf{x}} \ \sigma_{xh}$$
$$\mu_x = \mu_h, \quad \sigma_x = \sigma_h, \quad \mathbf{x} = \mathbf{V}\mathbf{d}. \quad (5)$$

Note that since $\mathbf{x} = \mathbf{V}\mathbf{d}$, the terms μ_x, σ_x, and σ_{xh} can also be expressed in terms of $\mathbf{d}$. For example, we can write $\mu_x = \mathbf{w}^T\mathbf{x} = (\mathbf{V}^T\mathbf{w})^T\mathbf{d} = \mathbf{m}^T\mathbf{d}$ with $\mathbf{m} = \left[\sum_{x_i \in D_1} w_i \cdots \sum_{x_i \in D_{n_p}} w_i\right]^T$. Similarly, $\sigma_x^2 = \mathbf{d}^T\mathbf{M}\mathbf{d} - \mu_x^2$ and $\sigma_{xh} = \mathbf{a}^T\mathbf{d} - \mu_x\mu_h$, where $\mathbf{M}$ is a diagonal matrix with $M_{ii} = m_i$, and $a_i = \sum_{h_j \in D_i} w_j h_j$. With these substitutions, the problem in Equation 5 becomes

$$\max_{\mathbf{d}} \ \mathbf{a}^T\mathbf{d}$$
$$\mathbf{m}^T\mathbf{d} = \mu_h, \quad \mathbf{d}^T\mathbf{M}\mathbf{d} = \mu_h^2 + \sigma_h^2. \quad (6)$$

The solution is given by the following expression (Appendix A):

$$d_i^* = \mu_h + \frac{\sigma_h}{\sigma_l}(l_i - \mu_h), \quad (7)$$

with $l_i = a_i/m_i$, and $\sigma_l^2 = \sum_{i=1}^{n_p} m_i (l_i - \mu_h)^2$.

Figure 5: *Post-sharpening after filtering (top) results in severe ringing and fails to capture the small-scale details in the background. The Lanczos filter (middle) can reduce ringing but still cannot capture the details well. Our method (bottom) utilizes the local content in the input image to avoid artifacts while preserving details. Input image courtesy of Flickr user nevynxxx.*

Solutions of optimization problems involving the SSIM index by fixing the mean have been utilized for other applications, where the optimum is then searched for using iterative methods [Channappayya et al. 2008a; Ogawa and Haseyama 2013; Shao et al. 2014]. However, closed-form solutions could only be derived for simple image models [Channappayya et al. 2006; Chai et al. 2014], or expansions on Fourier type bases [Brunet et al. 2010]. Although the images H and D, or basis vectors $\mathbf{v}_i$ do not satisfy the properties required for these solutions, we could still derive a closed-form solution, due to the structure of the downscaling problem.

For each pixel in the output image D, we thus get an optimum value from each patch overlapping that pixel. Each of these patches belongs to a different patch set S_k. The final value of the pixel is found by averaging these values. The weights w_i are usually taken from a Gaussian or constant window [Silvestre-Blanes 2011; Brunet 2012]. Following the latter, we assume that the weights are uniform summing to 1, since our patches are rather small as explained in the next section. We thus get the following value for the i^{th} pixel in image D (the i is now defined as a global index in D)

$$d_i^* = \frac{1}{n_p}\sum_{P_k} \mu_h^k + \frac{\sigma_h^k}{\sigma_l^k}(l_i - \mu_h^k), \quad (8)$$

where P_k denote the n_p patches overlapping this pixel. The form of the optimum image in Equation 8 is a non-linear filter on the input image H. The filter adapts to the image content in a perceptually optimal way as measured by the SSIM index. Our construction of the solution makes it clear that, it preserves the local luminance and contrast of the input image H, while maximizing local structural similarity. Although the filter is non-linear, it can be implemented with a series of linear operations as apparent from Equation 8 and we show in Appendix B.

3.3 Discussion and Analysis

We can view Equation 8 as an adaptive unsharp masking filter [Polesel et al. 1997] applied to the averaged l_i values, where the sharpening factor depends non-linearly on the local image content with the ratio σ_h^k/σ_l^k of the standard deviations of the input image, and a filtered version of it. This ratio thus adaptively adjusts the filter using H as the reference image so as to preserve the local features. Unsharp masking combined with pixel-wise contrast measures extracted from a reference image has previously generated excellent results for enhancing images generated by tone mapping [Krawczyk et al. 2007] or color to greyscale conversion [Smith

Figure 6: *Increasing the patch size n_p leads to loss of small scale features, from left to right $\sqrt{n_p} = 2, 8, 32$. Original image courtesy of Flickr user Salva Barbera.*

et al. 2008], as well as for rendered scenes [Ritschel et al. 2008]. It is interesting to see that our SSIM-optimal filter leads to a similar term for the downscaling problem.

It is well-known that trying to get sharper results by using a post-sharpening step after filtering, or a filter that generates sharper results by better approximating the sinc filter leads to artifacts when used for image downscaling [Kopf et al. 2013]. In Figure 5, we illustrate that our method avoids such problems and leads to better preservation of image features. Post-sharpening (with the sharpen filter of Adobe Photoshop) after filtering leads to severe ringing on the foreground object while failing to preserve the contrast in the background. This approach is fundamentally disadvantaged since the sharpening filter cannot use information from the original high resolution image to enhance the downscaled image. The Lanczos filter reduces the artifacts, but also fails to preserve the background. The adaptivity of the derived filter in Equation 8 ensures that all features are preserved while avoiding the ringing artifacts.

Patch size The only free parameter of our method is the patch or window size given by n_p. In general, determining the patch size for SSIM to best correlate the results with the response of the human visual system is a difficult problem. However, recent works confirm that as the image complexity increases, the window size should be reduced [Silvestre-Blanes 2011]. For the downscaling problem, it is crucial to capture the local structures in the input image H as well as possible. However, as the downscaling factor s increases, the patch size $s\sqrt{n_p}$ in H also gets bigger. Thus, for our problem, it is best to keep the patch size n_p as small as possible. For all our results, we set $n_p = 4$, corresponding to the smallest patch possible with size 2×2. A similar conclusion stems from the interpretation of the filter as an adaptive unsharp mask. The smoothed image in unsharp masking, corresponding to the averaged means μ_h^k of the patches in our case, can be made smoother to capture lower frequency bands. However, many lower bands are already captured in D. Furthermore, as the patch size gets larger, the ratio of the standard deviations decrease, leading to less enhancement. We show the effect of the patch size on the downscaled images in Figure 6. As the patch size increases, we start to lose small scale features. In the limit that the whole image is covered by one patch, the downscaled image approaches the filtered image given by l_i, since the contrasts σ_h and σ_l can be matched almost exactly.

Pixel values outside the dynamic range Since we do not restrict that the values of the pixels in D lie in $[0, 1]$ in the optimization, some pixels might end up having values outside this dynamic range. However, since the mean and standard deviations match for the optimum solution, in practice, the percentage of these pixels and their distance to the dynamic range is negligible for natural images. We show the percentage of pixel values outside the dynamic range averaged over 3000 randomly chosen natural images from the MSRA Salient Object Database [Liu et al. 2011] for 7 different downscaling factors ranging from $s = 1.75$ to 12.5 in Figure 7, a. There is a very small percentage of slightly off-range pixels.

Choice of patch sets Since we work with a small patch size of 2×2, the choice of the patch sets does not lead to a noticeable difference.

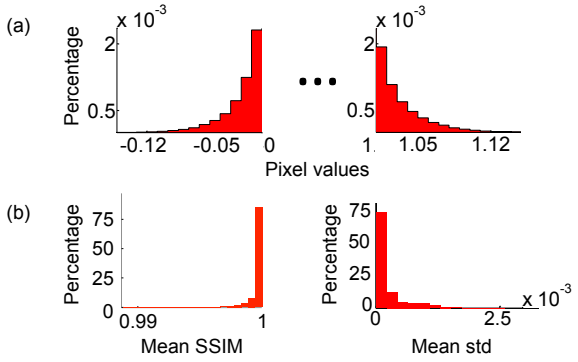

Figure 7: *(a) The percentage of pixel values outside the dynamic range for 3000 random natural images for 7 different sizes. (b) For each of the input images and sizes, the mean SSIM index and mean standard deviation between the downscaled image generated using all $\mathcal{S}_k$ by averaging (our solution), and those generated using individual $\mathcal{S}_k$'s, are computed. The figure shows the histograms of these values over the same set of images and sizes as in (a). Both measures show that optimizing over different sets does not alter the solution significantly.*

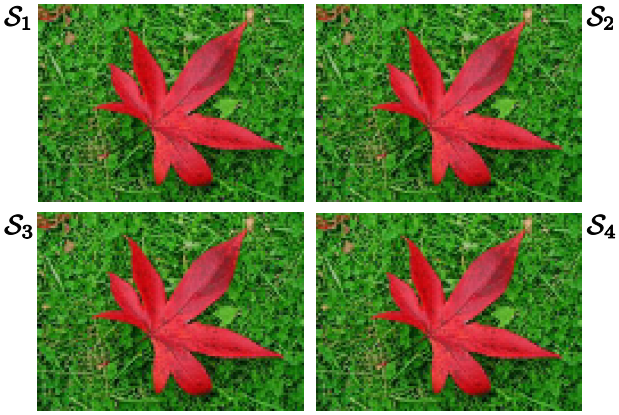

Figure 8: *Downscaled images obtained by optimizing over different patch sets $\mathcal{S}_k$. The images are almost identical, due to the small patch size we use in the downscaling algorithm.*

The resulting optimized images for different patch sets $\mathcal{S}_k$ and their mean (our SSIM-optimal image) are almost identical. In Figure 7, b, we show the distributions of the mean SSIM indices and mean standard deviations computed between the mean image (our solution) and the images optimized over different S_k's, for the same set of 3000 images and 7 sizes as above. Both measures indicate that the resulting images are almost identical. We show an example image optimized over different $\mathcal{S}_k$ in Figure 8. The images are almost identical and differ slightly in some of the patches where the texture has large and high frequency variations. We also ran a full optimization over all patches in $\mathcal{S}$ for some of the images, starting from the mean image (our solution), and have not observed a mean SSIM difference of more than 0.02 between our solution and the final solution obtained by the optimization over all $\mathcal{S}$ simultaneously.

Patches with constant values For some of the patches, the intensities l_i can be constant such that we get $\sigma_l = 0$. For these cases, there is no way to match the contrast, as required by the solution, and only the mean can be matched. Hence, for a patch with $\sigma_l < 10^{-6}$, we set the values of the pixels of the downscaled image in this patch to the mean μ_h of the patch.

Figure 9: *From top to bottom: the original image, bicubic filtering, content-adaptive downscaling [Kopf et al. 2013], our result. Our algorithm preserves the details better while leading to less jaggy edge effects. Original image courtesy of Flickr user Yasuhiko Ito.*

Figure 10: *From top to bottom: the original image, content-adaptive downscaling [Kopf et al. 2013], our result. The features are kept intact with our method. Left original image courtesy of Flickr user Ian Wilson, right original image courtesy of Flickr user foeoc kannilc.*

Color images SSIM is defined for images with a single channel, although some works explore utilizing extracted features [Lissner et al. 2013], or working in various color spaces [Bonnier et al. 2006]. We experimented with different color spaces including CIELAB, but did not see a significant difference in the results. Hence, we simply use the RGB space for all our results, and apply our algorithm to each channel independently.

4 Results

We performed a large number of experiments to validate the practical value of our method with thousands of images and many different downscaling factors, a detailed analysis, comparisons to existing methods, and a formal user study.

4.1 Downscaling Results and Analysis

We show several example results in Figures 1, 2, 5, 9–11, 15–17. Please refer to the supplementary images and video for many additional downscaling results. Our technique generates local pixel patterns that form structures resembling those in the input image, when viewed by a human observer. This effect is most apparent when there are perceptually important features (Figures 1, 10), textures (Figures 15, 16), or other small-scale details (Figures 1, 2, 15, 16, 17) in the input images. While trying to capture as much structure as possible, it also preserves the local contrast and luminance of the input image, which makes the overall look of the downscaled image close to the input (e.g. Figures 1, 16).

The algorithm does not significantly alter the features that are already captured by low-pass filters. This results in less jagged edge artifacts than previous high quality downscaling methods. We show an example downscaled edge in Figure 9. Our method performs a slight enhancement on the edge, resulting in less artifacts than with

Figure 11: *From left to right: Bicubic filtering, subsampling, our result. We get crisp details without Moirré patterns. Original image courtesy of Flickr user Chi King.*

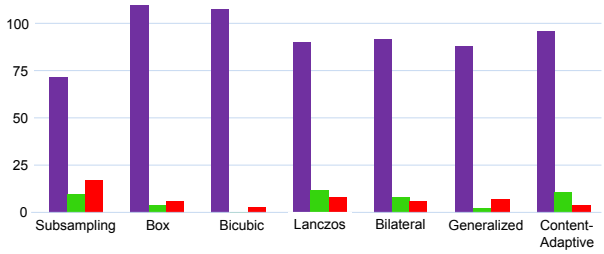

Figure 12: *User study results. The blue bars represent how many times our algorithm is selected, green is for no preference, and red is for the other algorithm.*

the previous content-adaptive method of Kopf et al. [2013]. If some details cannot be captured with the pixel budget in the downscaled image, they are mapped to noise-like structures that resemble those in the input image if viewed at the native resolution, as opposed to Moirré patterns, as with subsampling (Figure 11).

The method is also spatio-temporally consistent, leading to accurate representation of features, as can be clearly seen in Figures 1, right, and Figure 10. Classical filtering methods such as bicubic filtering are also consistent, but fail to generate crisp images. Aligning the kernels to local image features [Kopf et al. 2013] can generate crisper results, but the resulting kernels can miss or distort some features as in Figure 10, and small changes in input images are sometimes amplified, leading to flickering, as we illustrate in the accompanying video.

4.2 User Study

There are numerous studies on the correlation of the SSIM index with human perception when used as an image quality measure [Wang and Bovik 2009]. However, our particular problem of downscaling calls for a tailored formal user study. The design of our user study follows that of the previous study performed by Kopf et al. [2013], including the images used and all design choices.

The study is based on presenting the participants a large image, and two downscaled versions of that image. The participant is then asked to select the small image that she/he thinks represents a better downscaled version of the large image, or indicate no preference. One of the small images presented for each test is computed using our algorithm, and the other by a previous algorithm. We included subsampling, the classical box, bicubic, and Lanczos filtering, as well as bilateral filtering, and the state-of-the-art algorithms generalized sampling [Nehab and Hoppe 2011] and content-adaptive downscaling [Kopf et al. 2013]. There were 125 participants in the study.

The 13 natural images used in the study, originally from the MSRA Salient Object Database [Liu et al. 2011], are the same as the ones used in the previous study [Kopf et al. 2013]. We show some example results in Figure 16 (please see the supplementary material

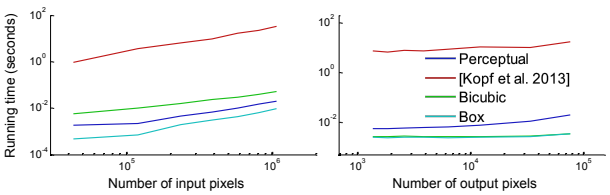

Figure 13: *Our algorithm runs only a few times slower than basic filters, and orders of magnitude faster than the state-of-the-art [Kopf et al. 2013].*

for the complete set of images and downscaling algorithms). They cover a variety of scenes with different types and scales of structures. The images were shown at the native resolution of the display, and zooming was not provided. The long side of the large images is 400 pixels, and that of the small images is 128 pixels. The study was performed online with participants from different parts of the world, educational backgrounds, occupations, and computer experience. Similar to the previous study [Kopf et al. 2013], we allowed the participants to move closer to the screen if they would like to, as would happen in practice for real-world situations. Each test for a particular participant involved a different image, and was repeated twice to check for consistency. All the results coming from subjects with consistency lower than 80% were discarded [Kopf et al. 2013], leaving results from 64 participants (the results do not change significantly for other rejection rates). There was no time limit to finish the study.

We present the results of the study in Figure 12. There is a clear preference for the results of our algorithm against all methods. The best competing downscaling method is simple subsampling, which was also the case for the previous study [Kopf et al. 2013]. Since subsampling does not involve any filtering, it preserves the crisp look of the images well, of course at the cost of well-known strong aliasing artifacts. For the user study images where these artifacts are not visible, the participants could not decide which algorithm to choose. For other images where the artifacts are noticeable, there is a clear preference for our algorithm. Hence, our algorithm preserves the crisp look of the images as in subsampling, but without the visible aliasing artifacts.

4.3 Implementation and Performance

Although the final method is based on a non-linear filter on the input image, it can be implemented very efficiently and robustly with simple convolutions and sums. We present the pseudo code of the algorithm in Appendix B. We implemented our algorithm in Matlab with native Matlab operators, some of which use multiple CPU cores.

We performed a performance test with 100 randomly chosen images on a computer with the configuration Intel Core i7 3770K CPU @350GHz. The method of Kopf et al. [2013] was run as a native executable. The results of the test are reported in Figure 13 for different input image sizes (with output image size fixed to 80×60), and output sizes (with input image size 640×480).

Our algorithm is only a few times slower than the box filter we used in the implementation of our algorithm, and $\times 500 - \times 5000$ faster than the method of Kopf et al. [2013] that relies on an iterative expectation-maximization based optimization. Our algorithm involves two box filterings followed by supsampling on the input image, and further operations on images of size proportional to the output image, as can be seen in the pseudo code in Appendix B. Hence, for smaller output sizes relative to the input size, it performs

Figure 14: *Insets from left to right: original image, bicubic filtering, our result. Since our method lacks scene semantics, it tries to preserve the noise in the input image. Original image courtesy of Flickr user City of Boston Archives.*

closer to the initial box filter we used, while increasing the output size slows it down a few times, as can be seen in Figure 13, right.

4.4 Limitations

A fundamental limitation of our method is its indifference to scene semantics. Like all previous methods, ours see the local structures in an image without any reference to what they actually represent. This, for example, leads to preservation of undesired details such as noise present in the input image, as we show in Figure 14, which is smoothed out by non-adaptive filters.

Our results exhibit fewer jagged edges (Figure 9) and aliasing artifacts (Figure 11) than methods that generate crisp images. However, if the image contains very regular repeating structures with a high frequency, aliasing can happen. The SSIM index tends to not prefer patches with a constant value, since this makes the index 0. Instead, our algorithm tries to reproduce the local contrast and structure. However, for perfectly regular structures, this might not be possible and a constant patch value is preferred instead. For those cases, such as on standard aliasing tests, we get artifacts similar to those produced by previous enhancement methods [Kopf et al. 2013]. Fortunately, such regular structures are rarely present in natural images. We observed that the small perturbations to regular structures that exist in most natural images can break the artifacts, as in Figure 11.

The SSIM index is known to not preserve the blur in the images [Chen et al. 2006]. We also observed that as opposed to thumbnail generation methods [Trentacoste et al. 2011; Didyk et al. 2012], our downscaling results do not contain the same amount of blur in the input image, especially for high downscaling ratios. We experimented with an extension of SSIM in the gradient domain [Chen et al. 2006], by solving for the gradients of the downscaled image, and subsequently a Poisson equation to get the actual image, but could not get satisfactory results so far.

5 Conclusions

We presented a novel method for image downscaling that aims to optimize for the perceptual quality of the downscaled results. Our extensive tests involving hundreds of images, and the user study clearly indicate that it generates perceptually accurate and appealing downscaling results, outperforming previous techniques. Despite its effectiveness and non-linear nature, it has a very simple, robust, efficient, and parallelizable implementation, making the algorithm a practical addition to the arsenal of image filters.

Future work

We used the basic form of the SSIM index. There are numerous extensions that modify the local similarity measure, the patch averaging stage, or extend it to feature and color spaces. It will be interesting to see how such extensions can affect the resulting downscaled images. However, some might also come with additional computational complexity. Although the downscaled videos exhibit less

Figure 15: *Our technique is able to capture small-scale details and textures while preserving local contrast and luminance to produce a perceptually accurate downscaled image. Input image courtesy of Flickr user ruru.*

flickering due to the consistency of the filter, better downscaling results can be obtained by incorporating extensions of the SSIM index to videos, e.g. models of speed perception [Wang and Li 2007]. The SSIM index does not model all aspects of human perception. We believe it is a very interesting direction to investigate how other perceptual measures can be utilized to improve image scaling results.

The SSIM index sees the image at the level of patches, and cannot by itself adapt to scene semantics. This leads to problems such as the noise amplification in Figure 14. Scene semantics such as background/foreground separation, properties of the objects in the scene, or saliency maps can be integrated into our algorithm by adaptively weighting the patches, or adjusting the parameters (α, γ) and patch size locally.

Finally, we believe that the SSIM index, with its simple definition and excellent correlation with perception, can be utilized more widely in similar image processing problems, and in particular for algorithms that rely on matching patches [Darabi et al. 2012].

Acknowledgements

We thank the anonymous reviewers for their helpful comments, Johannes Kopf for providing the implementation of his downscaling algorithm, Oliver Wang and Jean-Charles Bazin for helpful discussions, Liu et al. [2011] for providing the MSRA Salient Object Database, and all Flickr users that provide their images under the Creative Commons licence.

A Solution of the SSIM based Optimization

We parametrize the solution of the optimization problem by setting $\mu_x = \alpha\mu_h$, and $\sigma_x = \gamma\sigma_h$, for arbitrary (α, γ). Then, to maximize $SSIM(\mathbf{h}, \mathbf{x})$ for this particular (α, γ), we only need to maximize σ_{xh}. This leads to the following constrained optimization problem:

$$\begin{aligned} &\max_{\mathbf{d}} \ \mathbf{a}^T\mathbf{d} \\ &\mathbf{m}^T\mathbf{d} = \alpha\mu_h, \ \ \mathbf{d}^T\mathbf{M}\mathbf{d} = \alpha^2\mu_h^2 + \gamma^2\sigma_h^2. \end{aligned} \tag{9}$$

This problem can be solved by standard methods such as the method of Lagrange multipliers as we show in the supplementary material. The solution is given by the following expression

$$d_i^*(\alpha, \gamma) = \alpha\mu_h + \gamma\frac{\sigma_h}{\sigma_l}(l_i - \mu_h). \tag{10}$$

For each (α, γ), the $\mathbf{d}^*$ with the components d_i^* thus maximizes the covariance σ_{hx} and hence SSIM. If we plug in this expression for $\mathbf{d}^*$ into the expression for SSIM in Equation 4, we get the following maximum SSIM

$$SSIM(\mathbf{h}, \mathbf{d}^*(\alpha, \gamma)) = 4\frac{\sigma_l}{\sigma_h}\frac{\alpha\gamma}{(1+\alpha^2)(1+\gamma^2)}. \tag{11}$$

This expression is maximized if we select $\alpha = \gamma = 1$, giving us the global optimum $\mathbf{d}^*$. Hence, the solution of the problem in Equation 9 with the choice $(\alpha, \gamma) = (1, 1)$ coincides with the solution of the original problem in Equation 3.

B Pseudo Code of the Algorithm

In the algorithm below, all operations are element-wise on the single channel images, denoted with big letters. The function convValid$(X, P(y))$ convolves the image X with an averaging filter of size $y \times y$ for the valid range of the image such that the kernel stays within the image limits. The function convFull is similar but the image is assumed to be padded with zeros to allow the kernel go out of the image limits. The function subSample(X, y) subsamples the image X at intervals of y, I_X produces an image of the size of X with all ones, $X(C)$ gets all entries of the image X for which the corresponding entry in the image C returns true, and $\epsilon = 10^{-6}$.

Figure 16: *Example results from the user study. For each image, top-left: subsampling, top-right: bicubic filtering, bottom left: content-adaptive downscaling [Kopf et al. 2013], bottom-right: perceptual (ours).*

Algorithm 1 Downscale Image

Input: Input image H, downscaling factor s, patch size n_p.
Output: Downscaled image D.

procedure DOWNSCALEIMAGE
  $L \leftarrow \text{subSample}(\text{convValid}(H, P(s)), s)$
  $L_2 \leftarrow \text{subSample}(\text{convValid}(H^2, P(s)), s)$
  $M \leftarrow \text{convValid}(L, P(\sqrt{n_p}))$
  $S_l \leftarrow \text{convValid}(L^2, P(\sqrt{n_p})) - M^2$
  $S_h \leftarrow \text{convValid}(L_2, P(\sqrt{n_p})) - M^2$
  $R \leftarrow \sqrt{S_h / S_l}$
  $R(S_l < \epsilon) \leftarrow 0$
  $N \leftarrow \text{convFull}(I_M, P(\sqrt{n_p}))$
  $T \leftarrow \text{convFull}(R \times M, P(\sqrt{n_p}))$
  $M \leftarrow \text{convFull}(M, P(\sqrt{n_p}))$
  $R \leftarrow \text{convFull}(R, P(\sqrt{n_p}))$
  $D \leftarrow (M + R \times L - T)/N$

References

BANTERLE, F., ARTUSI, A., AYDIN, T., DIDYK, P., EISEMANN, E., GUTIERREZ, D., MANTIUK, R., AND MYSZKOWSKI, K. 2011. Multidimensional image retargeting. In *ACM Siggraph Asia 2011 Courses*, ACM, ACM Siggraph Asia.

BONNIER, N., SCHMITT, F., BRETTEL, H., AND BERCHE, S. 2006. Evaluation of spatial gamut mapping algorithms. In *Proc. 14th Color Imag. Conf.*, 56–61.

BRUNET, D., VRSCAY, E., AND WANG, Z. 2010. Structural similarity-based approximation of signals and images using orthogonal bases. In *Image Analysis and Recognition*, A. Campilho and M. Kamel, Eds., vol. 6111 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 11–22.

BRUNET, D., VRSCAY, E., AND WANG, Z. 2012. On the mathematical properties of the structural similarity index. *Image Processing, IEEE Trans. on 21*, 4 (April), 1488–1499.

BRUNET, D. 2012. *A Study of the Structural Similarity Image Quality Measure with Applications to Image Processing*. PhD thesis, University of Waterloo.

CHAI, L., SHENG, Y., AND ZHANG, J. 2014. Ssim performance limitation of linear equalizers. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 1220–1224.

CHANNAPPAYYA, S., BOVIK, A., AND HEATH, R. 2006. A linear estimator optimized for the structural similarity index and its application to image denoising. In *Image Processing, 2006 IEEE International Conference on*, 2637–2640.

CHANNAPPAYYA, S., BOVIK, A., CARAMANIS, C., AND HEATH, R. 2008. Ssim-optimal linear image restoration. In *Acoustics, Speech and Signal Processing (ICASSP), 2008. IEEE International Conference on*, 765–768.

CHANNAPPAYYA, S., BOVIK, A., AND HEATH, R. 2008. Rate bounds on ssim index of quantized images. *Image Processing, IEEE Trans. on 17*, 9 (Sept), 1624–1639.

CHANNAPPAYYA, S. S., BOVIK, A. C., CARAMANIS, C., AND JR., R. W. H. 2008. Design of linear equalizers optimized for the structural similarity index. *Image Processing, IEEE Trans. on 17*, 6, 857–872.

Figure 17: *Our downscaling method adaptively adjusts the local details such that downscaled images perceptually close to the original image are generated.*

CHEN, G.-H., YANG, C.-L., AND XIE, S.-L. 2006. Gradient-based structural similarity for image quality assessment. In *Image Processing, IEEE International Conference on*, 2929–2932.

DARABI, S., SHECHTMAN, E., BARNES, C., GOLDMAN, D. B., AND SEN, P. 2012. Image Melding: Combining inconsistent images using patch-based synthesis. *ACM Trans. Graph. (Proc. of SIGGRAPH 2012) 31*, 4, 82:1–82:10.

DEMIRTAS, A., REIBMAN, A., AND JAFARKHANI, H. 2014. Full-reference quality estimation for images with different spatial resolutions. *Image Processing, IEEE Trans. on 23*, 5 (May), 2069–2080.

DIDYK, P., RITSCHEL, T., EISEMANN, E., AND MYSZKOWSKI, K. 2012. *Perceptual Digital Imaging: Methods and Applications*. CRC Press, ch. Exceeding Physical Limitations: Apparent Display Qualities.

DONG, J., AND YE, Y. 2012. Adaptive downsampling for high-definition video coding. In *ICIP 2012*, 2925–2928.

GERSTNER, T., DECARLO, D., ALEXA, M., FINKELSTEIN, A., GINGOLD, Y., AND NEALEN, A. 2012. Pixelated image abstraction. In *NPAR 2012, Proc. of the 10th International Symposium on Non-photorealistic Animation and Rendering*.

HE, L., GAO, F., HOU, W., AND HAO, L. 2014. Objective image quality assessment: A survey. *Int. J. Comput. Math. 91*, 11 (Nov.), 2374–2388.

KOPF, J., SHAMIR, A., AND PEERS, P. 2013. Content-adaptive image downscaling. *ACM Trans. Graph. 32*, 6 (Nov.), 173:1–173:8.

KRAWCZYK, G., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2007. Contrast restoration by adaptive countershading. In *Proc. of Eurographics 2007*, Blackwell, vol. 26 of *Computer Graphics Forum*.

LISSNER, I., PREISS, J., URBAN, P., LICHTENAUER, M. S., AND ZOLLIKER, P. 2013. Image-difference prediction: From grayscale to color. *Image Processing, IEEE Trans. on 22*, 2, 435–446.

LIU, T., YUAN, Z., SUN, J., WANG, J., ZHENG, N., TANG, X., AND SHUM, H.-Y. 2011. Learning to detect a salient object. *Pattern Analysis and Machine Intelligence, IEEE Trans. on 33*, 2 (Feb), 353–367.

MITCHELL, D. P., AND NETRAVALI, A. N. 1988. Reconstruction filters in computer-graphics. In *Proc. of SIGGRAPH '88*, ACM, New York, NY, USA, 221–228.

NEHAB, D., AND HOPPE, H. 2011. Generalized sampling in computer graphics. Tech. Rep. MSR-TR-2011-16, February.

OGAWA, T., AND HASEYAMA, M. 2013. Image inpainting based on sparse representations with a perceptual metric. *EURASIP Journal on Advances in Signal Processing 2013*, 1.

PANG, W.-M., QU, Y., WONG, T.-T., COHEN-OR, D., AND HENG, P.-A. 2008. Structure-aware halftoning. *ACM Trans. Graph. 27*, 3 (Aug.), 89:1–89:8.

POLESEL, A., RAMPONI, G., AND MATHEWS, V. J. 1997. Adaptive unsharp masking for contrast enhancement. In *ICIP '97 3-Volume Set-Volume 1 - Volume 1*, IEEE Computer Society, Washington, DC, USA, 267–.

REHMAN, A., WANG, Z., BRUNET, D., AND VRSCAY, E. 2011. Ssim-inspired image denoising using sparse representations. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 1121–1124.

RITSCHEL, T., SMITH, K., IHRKE, M., GROSCH, T., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2008. 3D Unsharp Masking for Scene Coherent Enhancement. *ACM Trans. Graph. (Proc. of SIGGRAPH 2008) 27*, 3.

SHANNON, C. 1998. Communication in the presence of noise. *Proc. of the IEEE 86*, 2 (Feb), 447–457.

SHAO, Y., SUN, F., LI, H., AND LIU, Y. 2014. Structural similarity-optimal total variation algorithm for image denoising. In *Foundations and Practical Applications of Cognitive Systems and Information Processing*, vol. 215. Springer Berlin Heidelberg, 833–843.

SILVESTRE-BLANES, J. 2011. Structural similarity image quality reliability: Determining parameters and window size. *Signal Processing 91*, 4, 1012 – 1020.

SMITH, K., LANDES, P.-E., THOLLOT, J., AND MYSZKOWSKI, K. 2008. Apparent greyscale: A simple and fast conversion to perceptually accurate images and video. *Computer Graphics Forum (Proc. of Eurographics 2008) 27*, 2 (apr).

THÉVENAZ, P., BLU, T., AND UNSER, M. 2000. Interpolation revisited. *Medical Imaging, IEEE Trans. on 19*, 7, 739–758.

TOMASI, C., AND MANDUCHI, R. 1998. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, 839–846.

TRENTACOSTE, M., MANTIUK, R., AND HEIDRICH, W. 2011. Blur-Aware Image Downsizing. In *Proc. of Eurographics*.

WANG, Z., AND BOVIK, A. 2009. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *Signal Processing Magazine, IEEE 26*, 1 (Jan), 98–117.

WANG, Z., AND LI, Q. 2007. Video quality assessment using a statistical model of human visual speed perception. *J. Opt. Soc. Am. A 24*, 12, B61B69.

WANG, Z., BOVIK, A., SHEIKH, H., AND SIMONCELLI, E. 2004. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Trans. on 13*, 4 (April), 600–612.

WANG, S., REHMAN, A., WANG, Z., MA, S., AND GAO, W. 2011. Rate-ssim optimization for video coding. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 833–836.

WU, X., ZHANG, X., AND WANG, X. 2009. Low bit-rate image compression via adaptive down-sampling and constrained least squares upconversion. *Trans. Img. Proc. 18*, 3 (Mar.), 552–561.

YEGANEH, H. 2014. *Cross Dynamic Range And Cross Resolution Objective Image Quality Assessment With Applications*. PhD thesis, University of Waterloo.

ZHANG, Y., ZHAO, D., ZHANG, J., XIONG, R., AND GAO, W. 2011. Interpolation-dependent image downsampling. *Image Processing, IEEE Trans. on 20*, 11 (Nov), 3291–3296.

ZHANG, L., ZHANG, L., MOU, X., AND ZHANG, D. 2012. A comprehensive evaluation of full reference image quality assessment algorithms. In *ICIP 2012*, 1477–1480.

ZHOU, F., AND LIAO, Q. 2015. Single-frame image super-resolution inspired by perceptual criteria. *Image Processing, IET 9*, 1, 1–11.