# Quality-optimized Encoding of JPEG Images using Transform Domain Sparsification

Junichi Ishida $^{\$1}$, Gene Cheung $^{*2}$, Akira Kubota $^{\$3}$, Antonio Ortega $^{\#4}$

$^{\$}$ *Department of Electrical, Electronic and Communication Engineering, Chuo University*
*1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan*
[1] j.ishida@kubotalab.jp, [3] kubota@elect.chuo-u.ac.jp

$^{*}$ *Digital Content and Media Sciences Research Division, National Institute of Informatics*
*2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 180-0022, Japan*
[2] cheung@nii.ac.jp

$^{\#}$ *Department of Electrical Engineering, University of Southern California*
*3740 McClintock Ave., Los Angeles, CA 90089*
[4] antonio.ortega@sipi.usc.edu

***Abstract*—To account for the unique characteristics and limitations of the human visual system (HVS) when perceiving images, a variety of perceptual quality metrics have been proposed in the literature. Tailoring rate-distortion (RD) optimization for each metric is cumbersome and time-consuming. In this paper, we propose a general RD-optimization strategy called "transform domain bounding box" (BB) that can easily adapt to different quality metrics for JPEG-like block-based encoding of images. First, we define an objective function that is a weighted sum of the $l_0$-norm of the transform coefficients (a proxy for rate) and distortion from the transform domain representation. Next, for a given distortion target $\tau$, we define a don't care region (DCR) that specifies a search region of representations with distortion $\leq \tau$. We then show that the sparsest transform domain representation (lowest encoding rate) inside a BB that tightly contains the DCR can be constructed efficiently. Varying $\tau$ to induce different DCRs and corresponding BBs results in a set of constructed sparse representations of different sparsity counts, and the one that optimally trades off rate and distortion can be easily identified as solution to our objective. We show that our proposed BB strategy can be easily re-targeted for three common quality metrics: MSE, MSE-HVS-M and SSIM. Experimental results show that our BB strategy outperformed unoptimized JPEG compression by up to $1$dB in PSNR when distortion metric is MSE, up to $2$dB when metric is MSE-HVS-M, and up to $0.005$ when metric is SSIM.**

## I. Introduction

It is now well accepted in the signal processing community that classical signal distortion metrics such as mean square error (MSE) do not correspond well to how human visual system (HVS) perceives quality in images or videos. For example, spatial regions with larger intensities have stronger error-masking effects [1], and structural errors in an image are more objectionable than mosquito-like random noise [2]. In response, *quality assessment* has become a popular research topic, where the goal is to derive computational metrics that are more aligned to human's perceptual quality. However, there is currently no consensus on which derived quality metric is best. Given perceived quality is also influenced by the viewer's visual attention, driven by a complicated mixture of low-level visual stimulus and contextual information [3], the "best" quality metric is often application- and context-dependent, and there likely will never be one single metric that is optimal for all cases.

Given this state of affairs, individually optimizing image encoding for a variety of quality metrics becomes a necessary but cumbersome and painstaking process. To aleviate the burden of tailoring coding optimization for each quality metric, in this paper we propose a general rate-distortion (RD) optimization strategy called *transform domain bounding box* (BB) that can easily adapt to different quality metrics for JPEG-like block-based encoding of images. First, leveraging on our previous work on transform domain sparsification (TDS) [4], we define an objective function that is a weighted sum of the $l_0$-norm of the transform coefficients of representation $\mathbf{Y}$ (a proxy for coding rate) and distortion due to selected representation $\mathbf{Y}$. Next, we define a *don't care region* (DCR) that specifies a search region $\mathcal{S}(\tau)$ of representations with distortion less than or equal to a distortion target $\tau$. Given $\mathcal{S}(\tau)$, we then construct a BB $\mathcal{B}$ that tightly contains $\mathcal{S}(\tau)$ and whose sides are either parallel or perpendicular to the transform axes. Finding the sparsest representation $\mathbf{Y}^*$ inside BB $\mathcal{B}$ turns out to be easy, so if we perform this operation iteratively for different $\tau$, we can identify a set of sparse representations $\mathbf{Y}^*$'s with different sparsity counts. The one that optimally trades off rate with distortion is the the solution to our objective.

We show how BB strategy can be easily adapted to three popular quality metrics in the literature: MSE, MSE-HVS-M [1] and *Structural Similarity* (SSIM) [2]. In our experiments, we show that our proposed BB strategy outperformed unoptimized JPEG compression by up to 1dB in PSNR when distortion metric is MSE, up to 2dB when metric is MSE-HVS-M, and up to $0.005$ when metric is SSIM.

The outline of the paper is as follows. We first briefly

*MMSP'12, September 17-19, 2012, Banff, Canada.*
*???-?-????-????-?/10/\$??.??* ©*2012 IEEE.*

discuss related work in Section II. We next overview three popular quality metrics, MSE, MSE-HVS-M and SSIM, in Section III. We then describe our general transform domain BB strategy in Section IV, where we also discuss how the strategy can be implemented for each of the three metrics. Experiments for all three metrics are discussed in Section V. Finally, we present concluding remarks in Section VI.

## II. Related Work

As new quality metrics are still actively being investigated and proposed [2], [1], RD-optimized coding tailored specifically for an individual metric remains a popular research topic [5], [6]. Our current work is unique in that a general RD-optimization strategy is first sought, so that subsequent re-targeting for a specific metric only requires minimum investment in time and effort. We note that for distortion metric MSE, the re-targeted implementation of our BB strategy becomes very similar to thresholding algorithms like [5] (though instead of $l_0$-norm as a proxy for rate, [5] captures the cost of run-length coding as well, so that equally sparse transform domain representations will have different encoding costs). We do not claim strictly superior performance over all metric-specific algorithms; rather, we stress that the value of our proposal lies in the generality of the optimization framework, and the ease in re-targeting for any distortion metric that satisfies a transform-axis-aligned property (to be discussed in Section IV-A).

Transform domain sparsification (TDS) was studied in our previous work [4] for encoding of depth maps in texture-plus-depth format of multiview video, where the depth maps are used at decoder for view synthesis via depth-image-based rendering (DIBR). Though the concept of don't care region (DCR) and the usage of $l_0$-norm of transform coefficients as a proxy for coding rate are the same, the general optimization strategy using bounding box (BB) is new in our current work. Note also that DCR for depth maps in general is not transform-axis-aligned, while our BB strategy applies only for transform-axis-aligned DCRs.

## III. Image Quality Metrics

In this section, we overview three popular metrics for image/video quality assessment in the literature: Mean Squared Error (MSE), MSE-HVS-M, and Structure Similarity (SSIM). Our purpose is not to argue the merits of one metric over another, but that our BB optimization strategy can be applied to a variety of proposed quality metrics in the literature. See [7] for an extensive discussion on image quality metrics.

### A. Mean Squared Error

One of the most commonly used quality metrics in the image and video coding community is *mean squared error* (MSE): given original signal $\mathbf{x}$ and reconstructed $\mathbf{y}$ of equal dimension $\mathcal{R}^N$, we calculate the average of the component-wise squared differences between them:

$$\text{MSE}(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=0}^{N-1}(x_i - y_i)^2 \tag{1}$$

After computing MSE, Peak Signal-to-Noise Ratio (PSNR) is often computed as a function of MSE to reflect the quality of reconstructed signal $\mathbf{y}$:

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{2}$$

where $MAX_I$ is the maximum pixel value.

### B. MSE with Contrast Sensitivity and Masking

It is known that MSE does not capture HVS's varying sensitivity to different DCT frequencies. *PSNR-HVS-M* [1] is a relatively new metric that takes into account Contrast Sensitivity Function (CSF) and between-coefficient contrast masking of DCT basis functions. It is computed as follows. First, the weighted energy of DCT coefficients of a $8\times8$ image block $\mathbf{X}$ (in transform domain) is computed as:

$$E_w(\mathbf{X}) = \sum_{i=0}^{N-1} X_i^2 C_i \tag{3}$$

where $X_i$ is the $i$th DCT coefficient, and $C_i$ is the corresponding scaling factor determined by CSF. An error $\mathbf{X}-\mathbf{Y}$ between original block $\mathbf{X}$ and reconstructed block $\mathbf{Y}$ cannot be visually distinguished if it is smaller than $\max(E_w(\mathbf{X})/16, E_w(\mathbf{Y})/16)$.

This masking effect can be too large if there exists an edge in block $\mathbf{x}$ (in pixel domain). To take this into account, we compute and use $E_m(\mathbf{x})$ below instead:

$$E_m(\mathbf{x}) = E_w(\mathbf{x})\delta(\mathbf{x})/16 \tag{4}$$

where $\delta(\mathbf{x}) = (V(\mathbf{x}^{(1)}) + V(\mathbf{x}^{(2)}) + V(\mathbf{x}^{(3)}) + V(\mathbf{x}^{(4)}))/4V(\mathbf{x})$, $\mathbf{x}^{(k)}$ is the pixel sub-block in the $k$-th quadrant, and $V(\mathbf{x})$ is the variance of the pixel values in block $\mathbf{x}$. We can hence conclude that the maximum masking effect is $E_{\max} = \max(E_m(\mathbf{x}), E_m(\mathbf{y}))$.

Masking reduces error sensitivity for all coefficients except DC, and so we can write the resulting noticeable difference $\Delta_i$ for coefficient $i$ as:

$$\Delta_i = \begin{cases} |X_i - Y_i| & \text{if } i=0 \\ 0 & \text{elseif } |X_i - Y_i| \le E_{norm}/C_i \\ |X_i - Y_i| - E_{norm}/C_i & \text{o.w.} \end{cases} \tag{5}$$

where $E_{norm} = \sqrt{E_{\max}/64}$.

Finally, we can compute the metric MSE-HVS-M $MSE_H$ using obtained $\Delta_i$'s as follows:

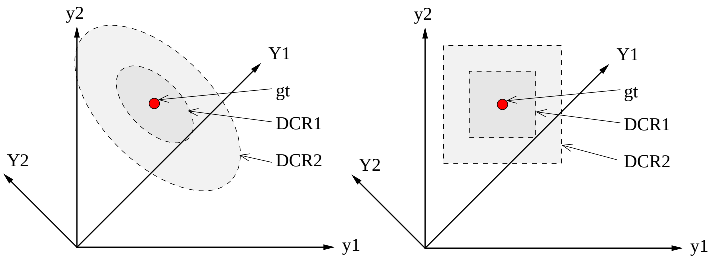
$$MSE_H = \sum_{i=0}^{N-1} \Delta_i^2 S_i \tag{6}$$

where $S_i$ is another scaling factor based on CSF [8]. PSNR-HVS-M is computed straightforwardly using $MSE_H$.

### C. Structural Similarity

Yet another popular alternative quality metric to MSE is the recently proposed *Structural Similarity* (SSIM) [2], defined as follows:

$$\text{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{7}$$



(a) DCRs for weighted MSE

(b) DCRs for $l_1$-norm

Fig. 1. Examples of transform-axis-aligned Don't Care Regions (DCR) using weighted MSE and $l_1$-norm of transform coefficients as distortion metrics for two-dimensional signals.

where $\mu_x$ and $\sigma_x^2$ are the pixel mean and variance of signal $\mathbf{x}$, and $\sigma_{xy}$ is the cross-correlation between signal $\mathbf{x}$ and $\mathbf{y}$. $c_1$ and $c_2$ are constants pre-set for stability reasons; SSIM is not sensitive to particular values of $c_1$ and $c_2$. SSIM has a maximum value of 1.0, which indicates the reconstructed signal $\mathbf{y}$ is exactly the same as the target signal $\mathbf{x}$.

SSIM is typically calculated locally for a small local patch ($11 \times 11$ is calculated in [2]), and quality for the entire image, mean SSIM (MSSIM), is computed simply as the average of calculated SSIMs of all patches in the image.

SSIM of a block of $N$ pixels can also be expressed in the block DCT domain[1] as follows [9]:

$$\text{SSIM}(\mathbf{X},\mathbf{Y}) = \left( \frac{2\frac{X_0 Y_0}{N} + C_1}{\frac{X_0^2+Y_0^2}{N} + C_2} \right) \times \left( \frac{2\frac{\sum_{k=1}^{N-1} X_k Y_k}{N-1} + C_1}{\frac{\sum_{k=1}^{N-1} X_k^2+Y_k^2}{N-1} + C_2} \right) \tag{8}$$

where $X_0$ is the DC coefficient in the block $\mathbf{Y}$ in DCT domain. For optimization convenience, we will use (8) in our SSIM computation. Also for convenience, we will define and use *distortion of SSIM* (dSSIM) instead of quality SSIM during optimization, as done in [6]:

$$\text{dSSIM}(\mathbf{X},\mathbf{Y}) = \frac{1}{\text{SSIM}(\mathbf{X},\mathbf{Y})} \tag{9}$$

## IV. Transform Domain Sparsification

Given orthogonal transform $\Phi$, our goal is to find an RD-optimal sparse representation $\mathbf{Y}$ of dimension $N$ in the transform domain. Specifically, we first assume that the number of non-zero transform coefficients for a code block is a good proxy for encoding rate; it has been shown theoretically for low-rate [10] and experimentally [4] that this is a reasonable approximation. We then seek to minimize the weighted sum of $l_0$-norm of $\mathbf{Y}$ (sparsity count) and distortion $d$ of the reconstructed pixel-domain signal $\mathbf{y} = \Phi^{-1}\mathbf{Y}$ compared to original signal (or *ground truth* (gt)) $\mathbf{x}^o$:

$$\mathbf{Y}^* = \arg\min_{\mathbf{Y}} \|\mathbf{Y}\|_0 + \lambda \; d(\mathbf{x}^o, \Phi^{-1}\mathbf{Y}) \tag{10}$$

where $\lambda > 0$ is a constant that specifies the relative importance of rate to distortion.

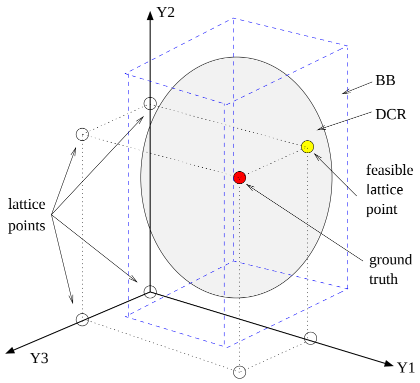


Fig. 2. Example of DCR and BB for a 3-dimensional signal. There are 7 lattice points in this case, one of which is feasible (inside DCR).

To solve (10) efficiently, we first provide an overview of a general *transform domain bounding box* strategy, which we will re-target for metrics MSE, MSE-HVS-M and dSSIM later.

### A. Don't Care Region

We first define the notion of *don't care region* (DCR), which is a restricted search region for sparse representations, given a distortion tolerance level. Specifically, we define $\mathcal{S}(\tau)$ for distortion level $\tau$ as a region of representations $\mathbf{Y}$'s with distortion less than or equal to $\tau$; i.e., $\mathcal{S}(\tau) = \{\mathbf{Y} \mid d(\mathbf{x}^o, \Phi^{-1}\mathbf{Y}) \leq \tau\}$.

The shape of the DCR obviously depends on the metric used to define distortion $d(\mathbf{x}^o, \Phi^{-1}\mathbf{Y})$. In this paper, we will restrict our consideration to DCRs that satisfy a *transform-basis-aligned* property. To properly define this property, we first note that frequency components of a transform domain representation $\mathbf{Y}$ can be divided into the following three types:

1) *zero-components*: frequency components that equal to zero, i.e., $\mathcal{A}^0 = \{Y_i | \; Y_i = 0\}$.
2) *gt-components*: non-zero frequency components that equal to gt's components, i.e., $\mathcal{A}^= = \{Y_i | \; Y_i = X_i^o, X_i^o \neq 0\}$.
3) *ngt-components*: non-zero frequency components that are different from gt's components, i.e., $\mathcal{A}^{\neq} = \{Y_i | \; Y_i \neq X_i^o, Y_i \neq 0\}$.

We can now define the transform-basis-aligned property for a DCR as follows:

> A DCR is *transform-basis-aligned* if by reassigning a subset of ngt-components $\mathcal{A}^{\neq}$ in a representation $\mathbf{Y}$ to gt-components $\mathcal{A}^=$ to construct $\mathbf{Y}'$, the resulting distortion is no worse; i.e., $d(\mathbf{x}^o, \Phi^{-1}\mathbf{Y}) \geq d(\mathbf{x}^o, \Phi^{-1}\mathbf{Y}')$.

Geometrically, transform-basis-aligned means that the DCR $\mathcal{S}(\tau)$ is widest along a dimension $i$ when representation $\mathbf{Y}$ has all other components $Y_j$'s, $j \neq i$, equal to gt components $X_j^o$. See Fig. 1 for examples of basis-aligned DCRs. An example of DCR that is not transform-basis-aligned would be the ellipse in Fig. 1(a) rotated clockwise by $45^o$, i.e., when the distortion metric is a weighted MSE of the pixel values in the *pixel* domain.

An important corollary of transform-axis-aligned is that when searching for representation $\mathbf{Y}$ inside a DCR $\mathcal{S}(\tau)$ that has the smallest Lagrangian cost (10), it is sufficient to consider *only* representations $\mathbf{Y}$'s with no ngt-components

[1] We will use the convention that the representation of a signal $\mathbf{x}$ in the transform domain, given orthogonal transform $\Phi$, is capital letter $\mathbf{X} = \Phi\mathbf{x}$.

$\mathcal{A}^{\neq}$. The reason is as follows. Any representation $\mathbf{Y}$ inside $\mathcal{S}(\tau)$ that has non-empty $\mathcal{A}^{\neq}$ can be converted to $\mathbf{Y}'$ that is also inside $\mathcal{S}(\tau)$, by reassigning components in $\mathcal{A}^{\neq}$ to $\mathcal{A}^{=}$ without increasing distortion. Further, $\mathbf{Y}'$ has the same sparsity count as $\mathbf{Y}$; i.e., $\|\mathbf{Y}\|_0 = \|\mathbf{Y}'\|_0$. Hence, $\mathbf{Y}'$ has no larger Lagrangian cost (10) than $\mathbf{Y}$, and it is sufficient to consider only representations with $\mathcal{A}^{\neq}$ as empty set. This is a discrete set of representations, and we call these *lattice points*. See Fig. 2 for examples of lattice points for a three-dimensional signal. Note that not all lattice points are feasible (inside $\mathcal{S}(\tau)$).

### B. Transform Domain Bounding Box Strategy

For a constructed DCR $\mathcal{S}(\tau)$, we next construct a *bounding box* (BB) $\mathcal{B}$, with boundaries either parallel or perpendicular to all axes in the transform domain, that properly contains DCR $\mathcal{S}(\tau)$, i.e., $\mathcal{S}(\tau) \subseteq \mathcal{B}$. In other words, bounding box $\mathcal{B}$ is defined with boundary $[L_j, U_j]$ in each dimension $j$ in the transform domain as follows:

$$\mathcal{S}(\tau) \subseteq \mathcal{B} = \{\mathbf{Y} \mid L_j \le Y_j \le U_j, \ \forall j = 0, \ldots, N-1\} \quad (11)$$

As an example, we see in Fig. 2 a DCR in grey for a three-dimensional signal is contained inside a BB in blue.

Constructing a *tight* (smallest possible) BB $\mathcal{B}$ that contains DCR $\mathcal{S}(\tau)$ in general is non-trivial. However, if DCR is transform-axis-aligned, then finding lower and upper bound $L_i$ and $U_i$ for dimension $i$ of BB $\mathcal{B}$ is much easier; by setting all other frequencies $Y_j$'s, $j \neq i$, to gt's $X_j$, one only needs to identify range of $Y_i$ where distortion $d(\mathbf{x}^o, \Phi^{-1}\mathbf{Y})$ does not exceed $\tau$. We discuss how this is done specifically for MSE, MSE-HVS-M and dSSIM in the following sections.

Having constructed BB $\mathcal{B}$, since transform $\Phi$ is orthogonal, we can construct a *sparsest* lattice point $\mathbf{Y}^*$ (in transform domain) inside $\mathcal{B}$ easily; i.e., $\mathbf{Y}^* = \arg\min_{\mathbf{Y}\in\mathcal{B}} \|\mathbf{Y}\|_0$. Specifically, for each defined boundary $[L_j, U_j]$ of $\mathcal{B}$, we set coefficient $Y_j = 0$ if $L_j \le 0 \le U_j$, and set $Y_j = X_j$ otherwise. Continuing with our example in Fig. 2, $L_3 \le 0 \le U_3$, so we can set coefficient $Y_3 = 0$ while keeping $Y_1 = X_1$ and $Y_2 = X_2$, resulting a 2-sparse representation shown in yellow. See Appendix for a proof for the minimum sparsity count of constructed $\mathbf{Y}^*$.

Because $\mathcal{B}$ is a superset that contains $\mathcal{S}(\tau)$, we can establish the following useful lemma:

*Lemma 1:* The sparsity count $\|\mathbf{Y}^*\|_0$ of the constructed sparsest lattice point $\mathbf{Y}^*$ inside BB $\mathcal{B}$ that contains DCR $\mathcal{S}$, i.e., $\mathcal{S}(\tau) \subseteq \mathcal{B}$, is a sparsity lower bound for any representation inside DCR $\mathcal{S}(\tau)$.

The corollary of lemma 1 is that if $\mathbf{Y}^*$ is also inside $\mathcal{S}(\tau)$, then it is also the sparsest representation in $\mathcal{S}(\tau)$. In practice, very often $\mathbf{Y}^* \in \mathcal{S}(\tau)$ is then the sparse solution we sought for given $\tau$. If $\mathbf{Y}^* \notin \mathcal{S}(\tau)$, then a simple greedy procedure can be taken where we iteratively restore a zero-component $Y_k^* = 0$ to $X_k^o$ (choosing one that results in the largest decrease in distortion $d(\mathbf{x}, \Phi^{-1}\mathbf{Y})$) until $\mathbf{Y}^* \in \mathcal{S}(\tau)$.

If we now iteratively vary $\tau$ to induce different DCRs $\mathcal{S}(\tau)$'s and resulting in different sparse lattice points $\mathbf{Y}^*$'s, we can find a series of representations with different sparsity / distortion tradeoffs. From among the discovered sparse representations $\mathbf{Y}^*$'s, we can find a near-optimal solution to (10) by identifying the one that has the smallest Lagrangian cost. See Fig. 3 for a summary of this BB strategy.

```
1) Compute suitable target distortions $\tau$'s.
2) For each computed $\tau$,
    a) Construct BB $\mathcal{B}$ that contains DCR $\mathcal{S}(\tau)$. Construct
       sparsest lattice point $\mathbf{Y}^*$ inside $\mathcal{B}$.
    b) If $\mathbf{Y}^* \notin \mathcal{S}(\tau)$, iteratively restore zero-component $Y_j^* =$
       $0$ to gt's $X_j^o$ (one with largest decrease in distortion),
       until $\mathbf{Y}^* \in \mathcal{S}(\tau)$.
    c) Compute Lagrangian cost of $\mathbf{Y}^*$.
3) Identify $\mathbf{Y}^*$ for all $\tau$'s with smallest Lagrangian cost as
   solution to (10).
```

Fig. 3. Generic transform domain bounding box strategy.

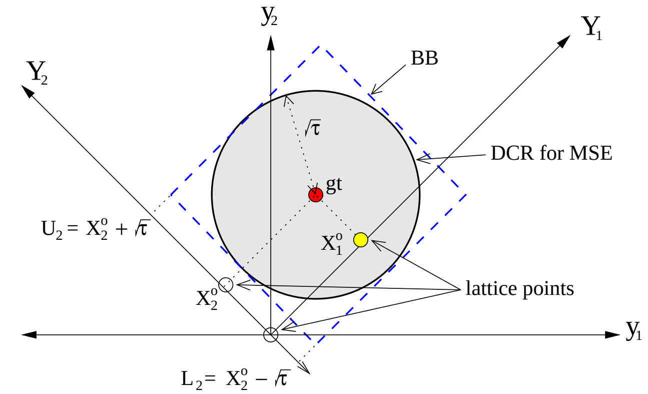


Fig. 4. DCR for tolerable MSE $\tau$ is shown in grey. Bounding box in transform domain is shown in blue.

There are two remaining problems that need to be solved to implement the BB strategy: i) how to identify suitable $\tau$'s for construction of DCRs $\mathcal{S}(\tau)$, and ii) for given $\tau$, how to construct tight BB $\mathcal{B}$ that contains $\mathcal{S}(\tau)$. We next discuss these problems specifically for distortion metrics MSE, MSE-HVS-M and dSSIM in order.

### C. Bounding Box Strategy for MSE

Suppose MSE is chosen as the distortion metric. The uniqueness of MSE is that DCR $\mathcal{S}(\tau) = \{\mathbf{y} \mid d(\mathbf{x}^o, \mathbf{y}) \le \tau\}$, translates simply to a sphere with radius $\sqrt{\tau}$. It is thus clear that the DCR is transform-axis-aligned. Then, given DCR sphere $\mathcal{S}(\tau)$ with radius $\sqrt{\tau}$, a bounding box $\mathcal{B}$ that tightly contains $\mathcal{S}$ can be very easily found:

$$\mathcal{B} = \left\{\mathbf{Y} \mid X_j^o - \sqrt{\tau} \le Y_j \le X_j^o + \sqrt{\tau}, \quad \forall j = 0, \ldots, N-1\right\} \quad (12)$$

See Fig. 4 for an illustration. Given this simple geometric interpretation, the previously discussed BB strategy can be implemented simply as follows.

For each non-zero coefficients $X_j^o$ of gt $\mathbf{X}$, we can compute a distortion $\tau = (X_j^o)^2$, which is the minimum distortion $\tau$ at which the lower and upper bound of frequency $j$ includes zero; i.e., $0 \in [L_j, U_j]$. Computing $\tau$'s for all frequencies provides us the set of suitable target distortions that we need for the BB strategy.

We note that while we cannot guarantee that constructed lattice point $\mathbf{Y}^*$ will also be inside DCR $\mathcal{S}(\tau)$, for the same sparsity count $\|\mathbf{Y}^*\|_0$, $\mathbf{Y}^*$ in fact has the smallest distortion of all representations, since the $\|\mathbf{Y}^*\|_0$ zero-components of $\mathbf{Y}^*$ correspond to gt's $\|\mathbf{Y}^*\|_0$ components $X_j$'s with the smallest magnitudes. Hence without performing step 2(b) in Fig. 3, we can nonetheless find the *optimal* solution to (10) using the BB strategy. The resulting algorithm is actually similar to thresholding algorithms designed explicitly for MSE in the literature [5].

### D. Bounding Box Strategy for MSE-HVS-M

Suppose MSE-HVS-M is chosen as the distortion metric. We first show DCR using MSE-HVS-M as distortion metric is transform-basis-aligned. If a representation $\mathbf{Y}$ has non-empty $\mathcal{A}^{\neq}$, it is easy to see that by reassigning those components in $\mathcal{A}^{\neq}$ to $\mathcal{A}^{=}$ to construct $\mathbf{Y}'$, the resulting distortion is no worse:

$$d(\mathbf{x}, \Phi^{-1}\mathbf{Y}) - d(\mathbf{x}, \Phi^{-1}\mathbf{Y}') = \sum_{i \in \mathcal{A}^{\neq} \in \mathbf{Y}} \Delta_i^2 S_i \geq 0 \qquad (13)$$

For a given distortion $\tau$ and DCR $\mathcal{S}(\tau)$, we can compute a tight BB enclosing DCR $\mathcal{S}(\tau)$ as follows. For DC coefficient $Y_0$, because there is no masking effect, we can compute the lower and upper bound for $Y_0$ as follows:

$$\begin{aligned} \tau &= (X_0 - Y_0)^2 S_0 \\ Y_0 &= X_0 \pm \sqrt{\frac{\tau}{S_0}} \end{aligned} \qquad (14)$$

For AC coefficient $Y_i$, it is slightly more involved because of masking:

$$\tau = (|X_i - Y_i| - E_{norm}/C_i)^2 S_i \qquad (15)$$

We first assume $\delta(\mathbf{X}) \approx \delta(\mathbf{Y})$. To find the lower bound $L_i$ of $Y_i$, we know $L_i < X_i$, and hence $E_{\max} = E_m(\mathbf{x})$. We can then derive $L_i$ as:

$$L_i = X_i - \sqrt{\frac{\tau}{S_i}} - \sqrt{E_w(\mathbf{X})\delta(\mathbf{X})/16/64}\ /C_i \qquad (16)$$

For upper bound $U_i$ of $Y_i$ where $U_i > X_i$, we know $E_{\max} = E_m(\mathbf{y})$. Using again (15), we can derive the following quadratic equation and solve for $U_i$:

$$\begin{aligned} 0 \ &= \ \left(C_i^2 - \frac{\delta(\mathbf{X})C_i}{1024}\right) U_i^2 \ - \ 2C_i^2 \left(\sqrt{\frac{\tau}{S_i}} + X_i\right) U_i \\ &\quad + C_i^2 \left(\sqrt{\frac{\tau}{S_i}} + X_i\right)^2 - \frac{\delta(\mathbf{X}) \sum_{j \neq i} X_i^2 C_i}{1024} \end{aligned} \qquad (17)$$

$U_i$ will be the *larger* of the two roots, since by assumption $U_i > X_i$.

We now need to find a suitable sequence of $\tau$'s for the algorithm to seek sparse solutions. For DC coefficient $Y_0$, it is simply $\tau = X_0^2 S_0$. For AC coefficient $Y_i$, it is computed as:

$$\tau = \begin{cases} \left[\max(0, X_i - \sqrt{E_w(\mathbf{X})\delta(\mathbf{X})/1024}/C_i)\right]^2 S_i & \text{if } X_i \geq 0 \\ \left[\min(0, X_i + \sqrt{E_w(\mathbf{X})\delta(\mathbf{X})/1024}/C_i)\right]^2 S_i & \text{o.w.} \end{cases} \qquad (18)$$

Having computed a suitable set of $\tau$'s, the BB strategy in Fig. 3 can be implemented for MSE-HVS-M to find a solution to (10).

### E. Bounding Box Strategy for dSSIM

When the distortion metric is dSSIM, we apply the transform domain BB strategy as follows. First, we argue that DCR using dSSIM as metric is transform-axis-aligned. The argument is that each frequency component $Y_j$ of representation $\mathbf{Y}$ contributes $Y_j^2$ to the numerator and $X_j Y_j$ to denominator of dSSIM, where the ratio is smallest when $Y_j = X_j$. Thus, DCR is widest at dimension $i$ when other frequency components $Y_j$'s, $j \neq i$, are the same as gt's $X_j$'s.

For a given DCR $\mathcal{S}$ of maximum dSSIM $\tau$, we compute the largest and smallest DC components, $L_0$ and $U_0$, of a tight BB $\mathcal{B}$ that contains $\mathcal{S}$ by letting $Y_j = X_j$ for $j = 1, \ldots, N-1$, and solving for $Y_0$ using (8):

$$\begin{aligned} \tau \ &= \ \left(\frac{\frac{X_0^2 + Y_0^2}{N} + C_2}{2\frac{X_0 Y_0}{N} + C_1}\right) \times \underbrace{\left(\frac{\frac{\sum_{k=1}^{N-1} 2X_k^2}{N-1} + C_2}{2\frac{\sum_{k=1}^{N-1} X_k^2}{N-1} + C_1}\right)}_{K_1} \\ 0 \ &= \ \left(\frac{1}{N}\right) Y_0^2 - \left(\frac{2\tau X_0}{K_1 N}\right) Y_0 + \left(\frac{X_0^2}{N} + C_2 - \frac{C_1 \tau}{K_1}\right) \end{aligned} \qquad (19)$$

$L_0$ and $U_0$ are the smaller and the larger values when $Y_0$ is sought in quadratic equation (19).

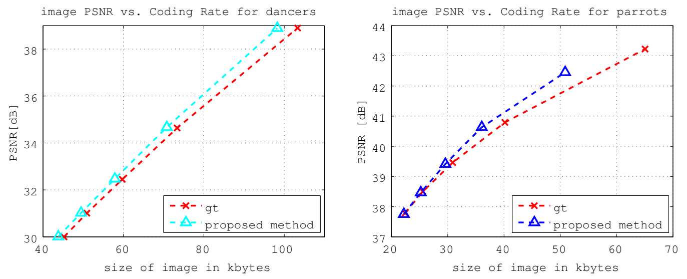
For each AC component $Y_j$, we follow similar procedure to solve for lower and upper bound, $L_j$ and $U_j$. Having derived all limits $[L_j, U_j]$'s, BB $\mathcal{B}$ that contains DCR $\mathcal{S}(\tau)$ is well defined.

The remaining task is how to identify a suitable set of dSSIM $\tau$'s so that corresponding DCR $\mathcal{S}(\tau)$'s will induce different sparsity count. We again follow similar procedure as we have done for MSE and MSE-HVS-M. For each coefficient $Y_j$, we set $Y_j = 0$ and all other coefficients $Y_k$'s, $k \neq j$, to original signal $X_k^o$'s. This results in SSIM using (8) and corresponding dSSIM, which we label $\tau_j$. This is in fact the minimum dSSIM value at which BB $\mathcal{B}$ will induce sparse lattice point $\mathbf{Y}$ with $Y_j = 0$, i.e., $0 \in [L_j, U_j]$. Computing $\tau$ in this fashion for all frequency components yields a suitable set of target distortions $\tau$'s for BB strategy in Fig. 3.

## V. Experimentation

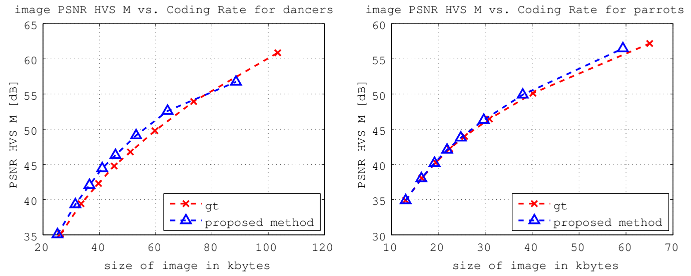
### A. Experimental Results for MSE

To test the effectiveness of our proposed transform domain BB strategy, we first investigate the effective of our strategy for MSE. Fig. 5 shows the coding performance (PSNR versus image encoding size) for our proposed scheme and the unoptimized JPEG compression implementation (gt), for images `dancers` and `parrots`. We see that our strategy outperformed gt by noticeable amount; the largest coding gain is 1dB in PSNR.



(a) dancers (b) parrots

Fig. 5. MSE comparison for dancers and parrots.



(a) dancers (b) parrots

Fig. 6. MSE-HVS-M comparison for dancers and parrots.

### B. Experimental Results for MSE-HVS-M

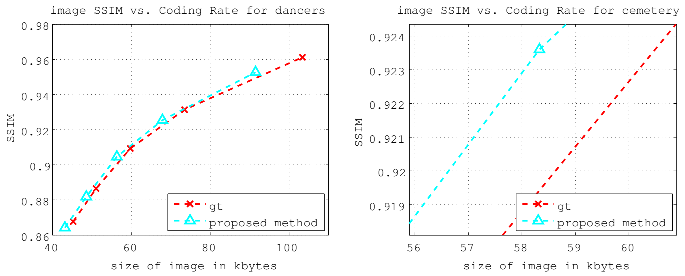
Next, we make the same comparison for distortion metric MSE-MVS-M. The coding results for the same dancers and parrots are shown in Fig. 6. We see again that our proposal outperformed gt in general. Specifically, our BB-based scheme outperformed gt by up to 2dB at mid-encoding rate.

### C. Experimental Results for SSIM

Finally, we made the same comparison when SSIM is the quality metric. Fig. 7 shows the coding performance of gt and our proposed strategy for images dancers and cemetery. The coding gain here is not as significant, though we do observe a 0.005 gain in SSIM.

## VI. Conclusion

To account for the unique characteristics and limitations of the human visual system (HVS) when perceiving images, a variety of quality metrics have been proposed in the literature. In this paper, we present a general RD-optimization strategy based on transform domain sparsification that can easily adapt to diffferent quality metrics for JPEG-like block-based encoding of images. In particular, we first define a don't care region (DCR) that specifies a restricted search region of representations with distortion no larger than a distortion target $\tau$. Then, using $l_0$-norm as a proxy for encoding rate, we show that the sparsest transform domain representation in a bounding box (BB) that tightly contains the DCR can be constructed efficiently. Varying $\tau$ to induce different DCRs results in different discovered sparse solutions, and the one that optimally trades off rate and distortion can be identified. Experimental results show that our BB strategy outperformed unoptimized JPEG compression by up to 1dB in PSNR when distortion metric is MSE, up to 2dB when metric is MSE-HVS-M, and up to 0.005 when metric is SSIM.



(a) dancers (b) cemetery

Fig. 7. SSIM comparison for dancers and cemetery.

## Appendix

We prove by contradiction that the constructed sparse representation $\mathbf{Y}^*$ inside a BB $\mathcal{B}$ in Section IV-B is indeed the sparsest one possible. Suppose there exists a feasible representation $\mathbf{Z}$ inside BB $\mathcal{B}$ with sparsity count strictly smaller than $\mathbf{Y}^*$; i.e., $\|\mathbf{Z}\|_0 < \|\mathbf{Y}^*\|_0$. Given $\mathbf{Z}$ has $N - \|\mathbf{Z}\|_0$ zero frequency components, it follows that there must be at least one zero frequency component $Z_k = 0$, where $0 \notin [L_k, U_k]$, since there are only $N - \|\mathbf{Y}^*\|_0$ frequency components $j$'s with $0 \in [L_j, U_j]$. However, having a zero component $Z_k = 0$ where $0 \notin [L_k, U_k]$ means $\mathbf{Z}$ must be outside BB $\mathcal{B}$ by definition of BB in (11). A contradiction.

## References

[1] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Eletronics*, Scottsdale, AZ, January 2007.

[2] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," in *IEEE Transactions on Image Processing*, vol. 13, no.4, August 2005, pp. 600–612.

[3] O. L. Meur and P. L. Callet, "What we see is most likely to be what matters: Visual attention and applications," in *IEEE International Conference on Image Processing*, Cairo, Egypt, November 2009.

[4] G. Cheung, A. Kubota, and A. Ortega, "Sparse representation of depth maps for efficient transform coding," in *IEEE Picture Coding Symposium*, Nagoya, Japan, December 2010.

[5] K. Ramchandran and M. Vetterli, "Rate-distortion optimal fast thresholding with complete JPEG/MPEG decoder compatibility," in *IEEE Transactions on Image Processing*, vol. 3, no.5, September 1994.

[6] C. Yeo, H. L. Tan, and Y. H. Tan, "On rate distortion optimization using SSIM," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, March 2012.

[7] F. D. Simone, D. Ticca, F. Dufaux, M. Ansorge, and T. Ebrahimi, "A comparative study of color image compression standards using perceptually driven quality metrics," in *SPIE Applications of Digital Image Processing*, San Diego, CA, August 2008.

[8] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on hvs," in *Second International Workshop on Video Processing and Quality Metrics*, Scottsdale, AZ, January 2006.

[9] S. S. Channappayya, A. C. Bovik, and R. W. Heath Jr., "Rate bounds on SSIM index of quantized images," in *IEEE Transactions on Image Processing*, vol. 17, no.9, September 2008, pp. 1624–1639.

[10] S. Mallat and F. Falzon, "Analysis of low bit rate image transform coding," in *IEEE Transactions on Signal Processing*, vol. 46, no.4, April 1998.