Saliency-driven image coding improves overall perceived JPEG quality

Vlad Hosu,[1] Franz Hahn,[1] Oliver Wiedemann,[1] Sung-Hwan Jung,[2] Dietmar Saupe[1]
[1]Department of Computer and Information Science, University of Konstanz, Germany
[2]Department of Computer Engineering, Changwon National University, S. Korea

***Abstract*—Saliency-driven image coding is well worth pursuing. Previous studies on JPEG and JPEG2000 have suggested that region-of-interest coding brings little overall benefit compared to the standard implementation. We show that our saliency-driven variable quantization JPEG coding method significantly improves perceived image quality. To validate our findings, we performed large crowdsourcing experiments involving several hundred contributors, on 44 representative images. To quantify the level of improvement, we devised an approach to equate Likert-type opinions to bitrate differences. Our saliency-driven coding showed 11% bpp average benefit over the standard JPEG.**

I. Introduction and motivation

Region-of-interest (ROI) based image compression techniques propose to compress the background more than the foreground in order to improve the perceived image quality. In his studies on two-level ROI based JPEG2000 image coding Bradley [1] has shown that his strategy did not improve on the standard JPEG2000 overall. It only did so at very low bitrates. Harding et al. [2] proposed a binary "visual interest"-guided JPEG2000 compression technique that increased image quality as measured objectively, but did not account for the fact that images of the same bitrate should be compared. Furthermore, in a recent study on perceptual quality in images Alers et al. [3] have shown that image foreground regions are much more important than the background. In view of these results we reconsider ROI based image compression. To better understand the importance of the ROI in perceived quality, we designed our own saliency-driven variable coding strategy.

Due to its simplicity and popularity we decided to base our variable quantization technique on JPEG Part 3 rather than working with JPEG2000. One of the intended purposes of variable quantization is "the ability to use the masking properties of the human visual system more effectively, and thereby achieving greater compression rates for the same subjective quality", see [4]. Variable quantization has already been shown to produce better results than the standard JPEG for special applications. For instance, Konstantinides et al. [5] have adjusted the quantization scaling factors in composite documents. Memon et al. [6] used a measure of block activity and type. None of these works evaluate results in terms of perceptual improvement (user studies). Harding et al. [2] have performed limited subjective studies, however due the low numbers of participants their results are inconclusive. Yu et al. [7] perform subjective evaluation as well using a sequential paired comparison quality assessment methodology. Their results are overall not in favor of their encoding technique.

II. Problem and choices

Our main research question is whether and by how much saliency driven image coding can improve on standard JPEG. We devise a saliency-based coding method that accounts for perceptual factors and an approach to evaluate the amount of perceived improvement in image quality.

A. Saliency-based image coding

In ROI-based image coding the regions of interest and their level of importance need to be defined. Motivated by the results of Alers et al. [3], we consider several factors which can affect image coding:

- choice of ROIs and their connection to saliency
- number of regions, their span, and level of importance
- overhead of encoding multiple quantization levels

Our proposed JPEG variable quantization technique uses multiple quality levels rather than just foreground and background. The number of ROIs and their span can be derived from simple transformations of the saliency map using Gaussian filtering with different standard deviations and a subsequent range adjustment. The importance levels of the quantization map relate to the range of saliency values. The final discrete block level JPEG quality factors result from the quantization of the transformed saliency map. Parameterising such a coding strategy allows to evaluate the impact of each factor on the perceived quality of the encoded images. A suitable technique for performance evaluation has to be sensitive to small differences in perceived quality.

B. Evaluation of results

The best judges for image quality are the end-users. Objective quality measures such as PSNR, MSE, SSIM, etc. are important, but do not capture the fine differences in personal opinion. However, paired comparison user-studies, have shown to be well suited for understanding fine opinion differences [8]. Thus, the subjective quality assessments in our work are also based on paired comparisons.

We propose a large-scale crowdsourced evaluation for our approach. Whereas some saliency based coding approaches were evaluated using objective quality measures [9], [10] others have done subjective evaluation. The latter have performed small scale lab studies [1], [7], [2] which are not conclusive on the matter of the performance of variable quantization techniques. In our crowdsourcing studies we involve a larger number of participants. Our images are chosen automatically

to have diverse content. This reduces the selection bias that an experimenter might create otherwise.

Another contribution of this paper is a way to go beyond opinion scores, aiming for a more tangible evaluation. In Section IV-D we introduce a way to estimate the bitrate difference between images of equal subjective quality, encoded by standard and by variable quantization JPEG. The method is based on differential mean opinion scores (DMOS) from paired comparisons. This relates opinion scores to objective coding factors, namely the bitrate difference between two compared images, in the spirit of the common Bjøntegaard-Delta bitrate (BD-BR) [11].

III. Methodology for Saliency-based Variable JPEG Coding

We derive our JPEG quality levels from saliency maps, generated from many eye-tracking fixation points.

A. Saliency Maps

Saliency maps quantify the level of relevance on a pixel-basis, which we utilize to maximize perceived subjective image quality. We model saliency simply by taking into account a set of eye fixation locations $[x_i, y_i]$ with weights w_i, $i = 0, \ldots, k-1$. Then the saliency map is obtained by applying a Gaussian filter,

$$s[x,y] = \frac{1}{S} \sum_{i=0}^{k-1} w_i \delta[x - x_i, y - y_i] * g[x,y]$$

where $\delta[x,y]$ is the 2D unit impulse signal, $g[x,y] = (2\pi\sigma^2)\exp(-(x^2x+y^2)/(2\sigma^2))$ is the 2D Gaussian filter with standard deviation σ, and S is a scaling value to normalize the maximal saliency on the image support to 1.

To reduce the (possibly large) set of eye fixation points to a small, but still representative subset we used K-Means clustering with $k = 8$ clusters $[x_i, y_i]$. The weights w_i were defined relative to the sizes of the K-Means clusters. Details about the source of the data are given in Subsection IV-A.

B. Saliency-based Variable Quantization JPEG

In variable quantization JPEG there is for each 8×8-image block with upper left corner at $[x,y]$ a quality factor $0 \leq q_B[x,y] \leq 100$ that controls the quantization for the block. The larger $q_B[x,y]$, the better is the reconstruction quality and, as a consequence, also the bitrate associated with the block. This factor is based on the average saliency per block,

$$s_B[x,y] = \frac{1}{64} \sum_{i=0}^{7} \sum_{j=0}^{7} s[x+i, y+j]$$

and set to

$$q_B[x,y] = \min(s_B[x,y] \cdot \Delta + q_{\min}, 100) \qquad (1)$$

where $0 \leq \Delta \leq 100$ is a parameter denoting the quality difference between foreground ($s_B = 1$) and background ($s_B = 0$) blocks, and $0 \leq q_{\min} \leq 100$ is an offset equal to the JPEG quality of a background block (Fig. 1). Note that the bitrate of the coded image will be a monotonically increasing function of the base quality given by $q_{\min}$.

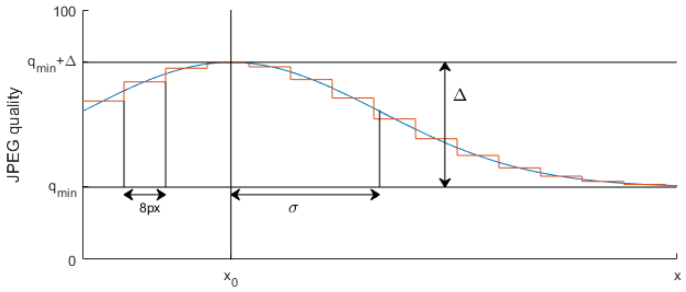

Fig. 1: Illustration of the saliency-based variable quantization JPEG coding strategy. For simplicity we show the JPEG variable quantization quality factor $q_B[x, y_0]$ (the staircase curve) for just one scan line ($y = y_0$) of an image that has just one eye fixation at $[x_0, y_0]$, in the same scan line. The smooth curve is the graph of the linear function of the saliency $x \mapsto s[x, y_0] \cdot \Delta + q_{\min}$ (compare Eq. (1)). The JPEG quality is bounded to the interval $[q_{\min}, q_{\min} + \Delta]$ and $q_{\min}$ is adjusted to achieve the target bitrate.

Thus, our approach to saliency-based variable JPEG coding has three parameters, σ, Δ, and $q_{\min}$. The parameter σ controls the size of the salient image region(s), Δ governs the quality difference between foreground and background blocks, and $q_{\min}$ is the base quality assigned to background blocks. Note that standard JPEG coding, i.e., not using variable quantization, is given by the special case of $\Delta = 0$, in which the JPEG quality is equal to $q_{\min}$.

When comparing a coding strategy, parametrized by $(\sigma, \Delta, q_{\min})$, with a standard JPEG approach it is necessary to compare only images of the same bitrate which is a function of all three parameters. Given a JPEG coded image at a certain bitrate, we thus choose the base quality $q_{\min}$ for the variable quantization JPEG coded image such that the target bitrate is achieved as closely as possible. Computationally, we apply the bisection method for efficiency. This reduces the set of free parameters to just two of them, (σ, Δ).

C. Side Information for Variable Quantization JPEG

For the reconstruction of an encoded image by variable quantization the decoder requires the block quantization factors $q_B[x,y]$. For this purpose JPEG prescribes a simple procedure similar to PCM. However, we found in experiments that for low bitrates the induced overhead is large and annihilates any gains that could be achieved using a saliency-based adaptive bitrate coding strategy. In our case, we can propose a more efficient coding scheme for this side information. We simply pass to the decoder the image saliency model itself together with the parameters $(\sigma, \Delta, q_{\min})$ such that the decoder can reconstruct all quality factors at the block level. For this purpose the x- and y-coordinates of the $k = 8$ fixation points are uniformly quantized to 8 bits, and their weights by only 3 bits. Thus, together with storage for σ (2 bits), Δ (2 bits) and $q_{\min}$ (7 bits), this side information amounts to only

$8(8+8+3)+2+2+7 = 163$ bits. Of course, the encoder must also use these same (quantized) saliency data and parameters.

IV. Experimental Design

We evaluate the proposed approach on a large set of images in a subjective crowd study. A number of parameter combinations of our model are compared against the standard JPEG compression algorithm. In the following sections we explain the choices and decisions made in this regard.

A. Dataset

We constructed the dataset based on the MIT-1003 dataset by Judd et al. [12]. The dataset includes eye-tracking information collected from laboratory experiments. We computed the SSIM [13] between the lowest and highest bitrate versions of each image and selected the 70% that had the lowest similarity between the lowest and highest desired bitrates. For this subset we binned images according to the 8-bit entropy of their respective saliency maps to ensure a variety of test images. 11 images were randomly selected from each bin, resulting in a total of 44 test images (41 color, 3 monochrome).

B. Parameters

For the purpose of this study we considered 5 bitrates, 4 values for σ and 3 values for Δ, listed in the table below. For each image source and each bitrate we compared a standard JPEG coded image with the 12 saliency-based adaptive bitrate coded variations. Thus, overall we had 2640 pairs for comparison.

Parameter	Number	Values	Unit
Bitrates	5	0.3, 0.36, 0.42, 0.5, 0.6	bpp
σ	4	5, 10, 15, 20	% of image width
Δ	3	15, 25, 35	JPEG quality factor

C. Crowd Design

Subjective evaluation of the paired comparison was performed in two separate crowd experiments using the CrowdFlower platform. In the first experiment subjects were presented with paired comparisons consisting of standard JPEG images and different realizations of our proposed variable JPEG approach of the same bitrates. In the second experiment paired comparisons were comprised of standard JPEGs of 10 different bitrates, including the original set of 5 bitrates of the first experiment: 0.30, 0.33, 0.36, 0.39, 0.42, 0.50, 0.60, 0.68, 0.82, 1.00. Viewers were asked to denote which of two pictures "shows more clear and sharp details". The evaluation was done on a 5-point Likert-type scale ranging from 1 (definitely the first) to 5 (definitely the second).

In each crowdsourcing experiment the contributors were tested based on questions with known answers. Only contributors maintaining a 70% accuracy on both the current task and overall on the Crowdflower platform were allowed to participate. Contributors that failed to maintain their accuracy at any point in the task were disqualified.

Test questions were comprised of paired comparisons between standard JPEG images encoded at different bitrates.

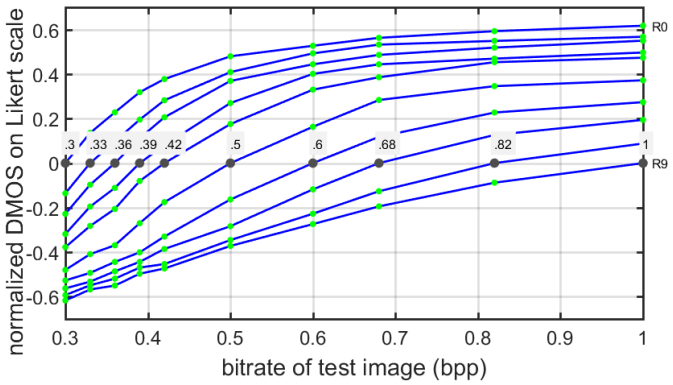

Fig. 2: DMOS results (averaged over all 44 image sources) for standard JPEG images encoded at bitrate R (horizontal axis) with respect to reference bitrates $R_j = 0.3, \ldots, 1.0$, indicated at the DMOS $= 0$ line. These DMOS curves are used to estimate the bitrate advantage of our saliency-based adaptive bitrate JPEG encodings. The 95% confidence intervals are ± 0.045 on average with a maximum of ± 0.05.

Each test question had multiple allowed answers to allow for some variability. Contributors were presented a short qualification quiz comprised of test questions. Upon its completion contributors were allowed to enter the experiment. Throughout the experiment random hidden test questions were presented.

Viewers were given very brief instructions. They were asked to denote their answers quickly, so as to indicate their first impression of each paired comparison. No time constraint was imposed on the task, and each contributor was allowed to rate at most 500 pairs of images (19% of the total).

D. Performance Evaluation

In the first experiment each of the 2640 test images was compared to the corresponding standard JPEG image of the same bitrate. The order of the images was randomized. We assigned a score to a test image as follows: 1.0 when the test image was strongly preferred, 0.5 slightly preferred, 0 no preference, -0.5 the standard JPEG was slightly preferred, or -1.0 standard JPEG strongly preferred. The average score of the test image is thus a normalized DMOS. A positive DMOS indicates that the saliency-based coding was preferred over the corresponding standard JPEG.

To better understand the meaning behind a given DMOS we propose a quantification in terms of bitrate difference similar to the Bjøntegaard-Delta bitrate. The algorithm is more complex than for the standard case of the Bjøntegaard-Delta metric because it can only rely on DMOS in place of MOS or PSNR quality functions of bitrate. For a given test image that was encoded to a bitrate R_j using one of the 12 coding strategies we ask for the (larger) bitrate $R^\star$ of a corresponding standard JPEG image (of the same source) that has the same perceptual image quality. Then the bitrate improvement achieved by the saliency-based adaptive encoding is the difference $R^\star - R_j$.

In order to avoid the large cost for directly comparing all 2640 test images with a set of standard JPEG images, we

Fig. 3: Standard JPEG vs. our saliency-based variable coding results at the same bitrate. Some of the best parameter settings for the images in our collection with respect to DMOS. The salient parts of the image are considerably better quality than the less important and less noticeable background details. Refer to Table I for the parameters and study results for each image.

estimate the bitrate difference as follows. The method is based on the data acquired from the second experiment in which we carried out paired comparisons for standard JPEG images with 10 different bitrates, $R_0, \ldots, R_9$. For all such comparisons of two images of different bitrates R_i, R_j, we computed the DMOS $D_{R_j}(R_i)$ of bitrate R_i with respect to the reference bitrate R_j, averaged over all 44 image sources. See Fig. 2 for piecewise linear interpolations $D_{R_j}(R)$ for the 10 reference bitrates $R_0 = 0.3, \ldots, R_9 = 1.0$ bpp and $0.3 \leq R \leq 1.0$. Note that by design these functions $D_{R_j}(R)$ are monotonically increasing with the bitrate R.

Now, assume we are given a test image, adaptively encoded at bitrate R_j and with a DMOS advantage of $\epsilon > 0$ in comparison with the corresponding standard JPEG image at the same bitrate, R_j. Then it can be estimated that the corresponding standard JPEG encoded image at bitrate $R^\star = D^{-1}_{R_j}(\epsilon) > R_j$ has the same DMOS advantage. Thus, $R^\star - R_j$ can be taken as the bitrate difference, and in Section V we report the relative bitrate difference $(R^\star - R_j)/R_j$, being more meaningful than the absolute values.

V. Results

In order to evaluate the performance of our approach we performed two crowdsourcing experiments:

- The first compared the same bitrate encoded image using two approaches: one was the result of our variable quantization approach and the other the corresponding standard JPEG (VAR–STD)
- The second experiment compared different bitrates of the same image coded using standard JPEG (STD–STD).

We used the CrowdFlower crowdsourcing platform to perform the experiments. A total of 453 viewers from 53 countries participated in the first experiment, 96% passed the qualification test and staying above 70% accuracy. For the second experiment a total of 950 viewers from 72 countries participated, 96% of which qualified and stayed within accuracy bounds.

	DMOS	Bitrate	Bitrate difference	Bitrate (%) difference	Δ	σ(%)
musician	0.42	0.30	0.16	53%	35	20%
animal	0.32	0.30	0.09	30%	15	10%
flights	0.23	0.36	0.04	11%	35	20%
beach	0.37	0.42	0.09	21%	35	20%

TABLE I: Parameter settings and results for the example images shown in Fig. 3. Bitrate differences show an improvement over standard JPEG.

A. Performance

We aggregated the results of each experiment by computing the normalized DMOS for each version of each image. A positive DMOS score in the first experiment shows a preference for our approach, whereas a negative one implies a preference for the standard coding. We did this for the best parameter combinations (σ, Δ) for all bitrate versions of the variable quantization approach. This amounts to 220 image versions: 44 originals at 5 bitrates each. The results are shown in the histogram in Fig. 4.

Using the DMOS data from the second experiment (STD–STD) we computed the curves shown in Fig. 2. Relying on these trend lines we computed the relative bitrate difference for our VAR–STD experiment. The results are shown in Fig. 5. We notice that in most cases our variable quantization approach shows an advantage over the standard JPEG encoding. On average we obtained an 11% improvement in bitrate. See Figure 3 for a visual comparison.

VI. Conclusions

The results of our evaluation are promising. They show that the approach works well in some situations and it could be further improved in others. However, to apply it in practice we need to devise solutions for the limitations of our approach.

A. Limitations and Future Work

Our adaptive coding approach relies on a small set of representative fixation points to derive a simple saliency map

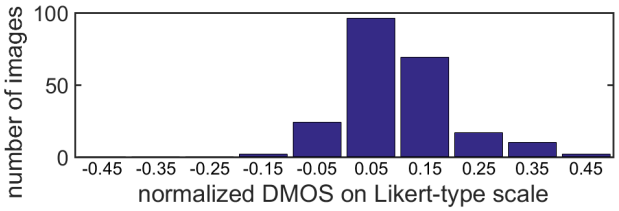

Fig. 4: DMOS for the best parameters of our variable quantization approach in the VAR–STD experiment for all bitrate versions of each image. Positive values show the preference for our approach compared to the standard JPEG at the same bitrate. The average DMOS is 0.1 in favor of our method.

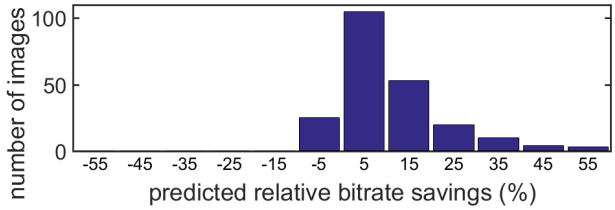

Fig. 5: Relative bitrate differences (%) for the best parameters with respect to DMOS in the VAR–STD experiment for all bitrate versions of each image. Positive percentages mean that our saliency-driven JPEG approach is better when compared to the standard JPEG at the same physical bitrate. This equates to an average of 11% bitrate difference.

used to prioritize the quality in the more salient areas. To make the approach practical a prediction of these fixation points is required. Such predictions could be obtained from saliency prediction methods such as in [14] and later extensions [15].

In this contribution we considered the ideal parameters for encoding an image based on the results of the crowdsourcing study. To apply the method in a practical codec, we would need to either fix the parameters or choose them based on image content. To solve this issue in future research, more examples of optimal parameters from crowdsourcing studies can facilitate a machine learning approach to automatically estimate these parameters.

Finally, the choice of applying our approach to the dated JPEG codec can be seen as a limitation. However, the purpose of this research merely was to prove in contrast to previous findings that saliency-driven image coding can achieve a measurable and significant improvement not only in special cases like composite documents. Based on these findings a complete codec based on JPEG and suitable adaptations for JPEG2000 and other current codecs should now be considered.

B. Contributions

Saliency driven coding is open for further improvement. Contrary to the findings of previous works [1], [16], our user studies conclusively show that variable coding works across multiple bitrates. In some cases, our implementation prototype gives excellent results reaching over 50% relative bitrate improvement, Fig. 5. However, in other cases the best improvement is only marginal. We still need to learn what makes images suitable for variable coding and how to best optimize our coding strategy.

We propose a solution for joining two important evaluation strategies for image coding. On one hand, objective measures such as PSNR, SSIM have been widely used for comparing coding strategies. On the other hand, subjective user studies provide the means to inspect human perception. Computational measures are universal, but weakly related to perception. However, subjective scores are contextual depending on the experimental setup. We ground our subjective results relating them to an objective factor, i.e., coding bitrate. The fused evaluation gives a more accurate measure of the quality of the results.

Acknowledgment. We thank the German Research Foundation (DFG) for financial support within project A05 of SFB/Transregio 161.

REFERENCES

[1] A. P. Bradley and F. W. M. Stentiford, "Visual attention for region of interest coding in JPEG 2000," *Journal of Visual Communication and Image Representation*, vol. 14, no. 3, pp. 232–250, 2003.

[2] P. Harding and N. M. Robertson, "Visual saliency from image features with application to compression," *Cognitive Computation*, vol. 5, no. 1, pp. 76–98, 2013.

[3] H. Alers, J. Redi, H. Liu, and I. Heynderickx, "Studying the effect of optimizing image quality in salient regions at the expense of background content," *Journal of Electronic Imaging*, vol. 22, no. 4, 2013.

[4] International Telecommunication Union (ITU), *Information technology — Digital compression and coding of continuous-tone still images: Extensions, Annex C (ITU-T Recommendation T.84)*, July 1996.

[5] K. Konstantinides and D. Tretter, "A JPEG variable quantization method for compound documents," *IEEE Transactions on Image Processing*, vol. 9, no. 7, pp. 1282–1287, 2000.

[6] N. D. Memon and D. R. Tretter, "Method for variable quantization in jpeg for improved perceptual quality," in *Electronic Imaging*. International Society for Optics and Photonics, 2000, pp. 24–34.

[7] S. X. Yu and D. A. Lisin, "Image compression based on visual saliency at individual scales," in *Advances in Visual Computing: 5th International Symposium, ISVC*, 2009, pp. 157–166.

[8] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[9] Y. Hu, F. Meng, and Y. Wang, "Improved jpeg compression algorithm based on saliency maps," in *Image and Signal Processing (CISP), 2012 5th International Congress on*, Oct 2012, pp. 262–266.

[10] G. Gowripushpa, G. Santoshi, B. Ravikiran, J. S. Rani, and K. S. Harsha, "Implementation of ROI based baseline sequential adaptive quantization," *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, no. 2, pp. 361–367, 2014.

[11] G. Bjontegaard, "Calcuation of average psnr differences between rd-curves," *Doc. VCEG-M33 ITU-T Q6/16*, 2001.

[12] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *12th IEEE International Conference on Computer Vision*, 2009, pp. 2106–2113.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Irocessing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[15] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *2013 IEEE Intern. Conf. on Computer Vision*. IEEE, 2013, pp. 921–928.

[16] A. P. Bradley and F. W. M. Stentiford, "JPEG 2000 and region of interest coding," *Digital Image Computing Techniques and Applications (DICTA)*, vol. 303, no. 1, pp. 303–308, 2002.