3
Saliency Enabled Screen Content Coding

The distinguishing properties of screen content images (SCIs) compared to the camera content images (CCIs) were discussed in detail in Chapter 1. The main difference between these two sets of images was the naturalness property. Due to the dominance of textual and graphical information in the SCIs, it was observed to have lower variance than CCIs. Due to these distinguishing properties of SCIs, the current image and video compression methods, which are being extensively used on multimedia communication channels, fail to provide satisfactory results. The main reason is that the methods like JPEG [Wallace, 1992] or HEVC [Sullivan *et al.*, 2012] are designed in such a way that the high-frequency regions are heavily quantized to achieve a low bitrate requirement. As a result of this the textual regions, which are having high frequencies, get illegible at high compression ratio (CR). One way to solve this problem is to identify the location of the textual regions and prevent those regions from getting heavily quantized. This motivated us to design a two-level saliency-based compression framework for SCIs instead of a multi-level framework which is discussed in the previous chapter. The reason for this was, if the textual regions of an SCI can be extracted and mapped as salient then it would be easy to preserve these regions from getting distorted even at high CR.

We proceeded towards developing a compression framework on the similar lines of the previous work as discussed in Chapter 2, where the JPEG quantizer was made intelligent in order to perform a judicious quantization. The major difference here was that instead of a multi-level saliency map, a binary saliency map was provided to the quantizer. Moreover, it was also observed from Chapter 1 that the frequency in the textual regions is higher than the non-textual ones. As JPEG framework converts the image blocks into DCT at the initial state, we decided to use this DCT information in order to mark every block as textual or non-textual. The motivation behind this was to enable the proposed encoder to create the saliency map with low computation cost.

Through this chapter a two-level saliency enabled compression method is proposed for SCIs with the aim to preserve the textual regions at high CR [Rahul and Tiwari, 2019b]. The proposed method identifies the textual regions as salient regions using first few DCT coefficients. The saliency map is then provided to the JPEG quantizer in order to judiciously retain the salient textual regions in the SCI.

The rest of this chapter is organized as follows. Section 3.1 explains the proposed two-level saliency detection and image compression algorithm. The experimental results are shown in Section 3.2, where the performance of the proposed method is compared with the state-of-the-art similar methods. The concluding remarks are given in Section 3.3

3.1 PROPOSED SCI COMPRESSION METHOD

The framework of the proposed encoding method is shown in Figure 3.1. The proposed framework is divided into two paths, Path A and Path B. Path A follows the JPEG framework with only change in the quantization step. Path B, on the other hand, shows the contributions of this chapter and provides the saliency map for the adaptive quantization. For ease of implementation,

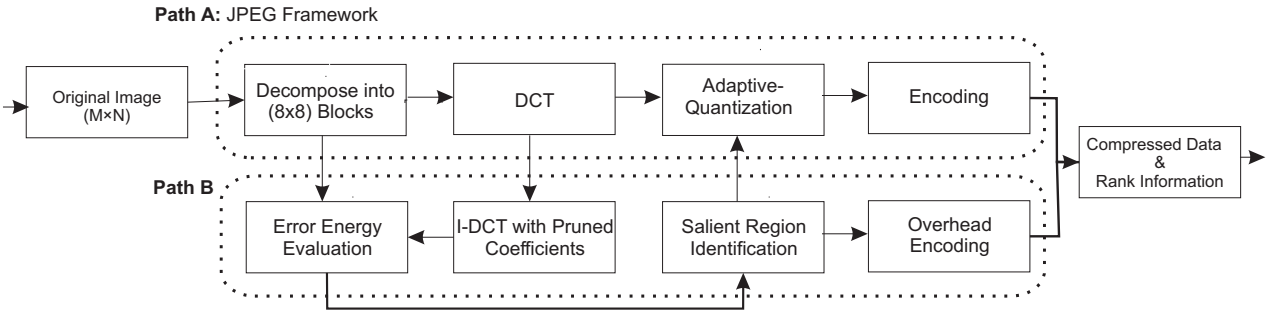

Figure 3.1 : Proposed encoding framework

various steps are briefly explained as follows:

In an example implementation of the present subject matter, an image for transmission is partitioned into a plurality of non-overlapping image blocks of size 8×8. The plurality of non-overlapping image blocks is transformed into the frequency domain by using discrete cosine transformation (DCT). For each non-overlapping image block first K number of DCT coefficients are selected by a zig-zag technique as shown in Figure 3.3 (a-b). The image block is reconstructed by performing reverse discrete cosine transformation to the selected first K number of discrete cosine transformed coefficients for the non-overlapping image block.

To identify the textual or salient region in the image, the error energy is evaluated between the image block and the reconstructed ones from DCT blocks using pruned coefficients. The error energy of every block is then compared to a threshold value in order to decide whether it belongs to a salient or non-salient region. The saliency map is then provided to the quantizater for judiciously quantizing the salient and non-salient regions in the image. The saliency map information is arithmetically encoded in order to reduce the overhead information. The detailed flowchart is shown in Figure 3.2.

The reconstruction of the image at the decoder end is the inverse of the encoding steps. Unlike other saliency enabled methods as in [Christopoulos *et al.*, 2000], where decoder requires to reproduce the ROI mask, making the decoder complex, the decoder of the proposed method is simple. The detailed description of the key steps in the encoding process is given as follows:

3.1.1 I-DCT with Pruned Coefficients

Block-based algorithms such as JPEG or HEVC are very effective in order to compress images which have continuous tone. However, such methods do not perform well with images where substantial textual information or graphics are present such as SCIs.

The developed method in this chapter lies on the fact that it may be beneficial to be able to classify each block as being either textual or non-textual. With such classification, different quantization parameters can be employed for the textual and non-textual data in order to obtain a good compression performance with minimal loss in the perceptual quality.

Classification methods which have been proposed over the years for separating the

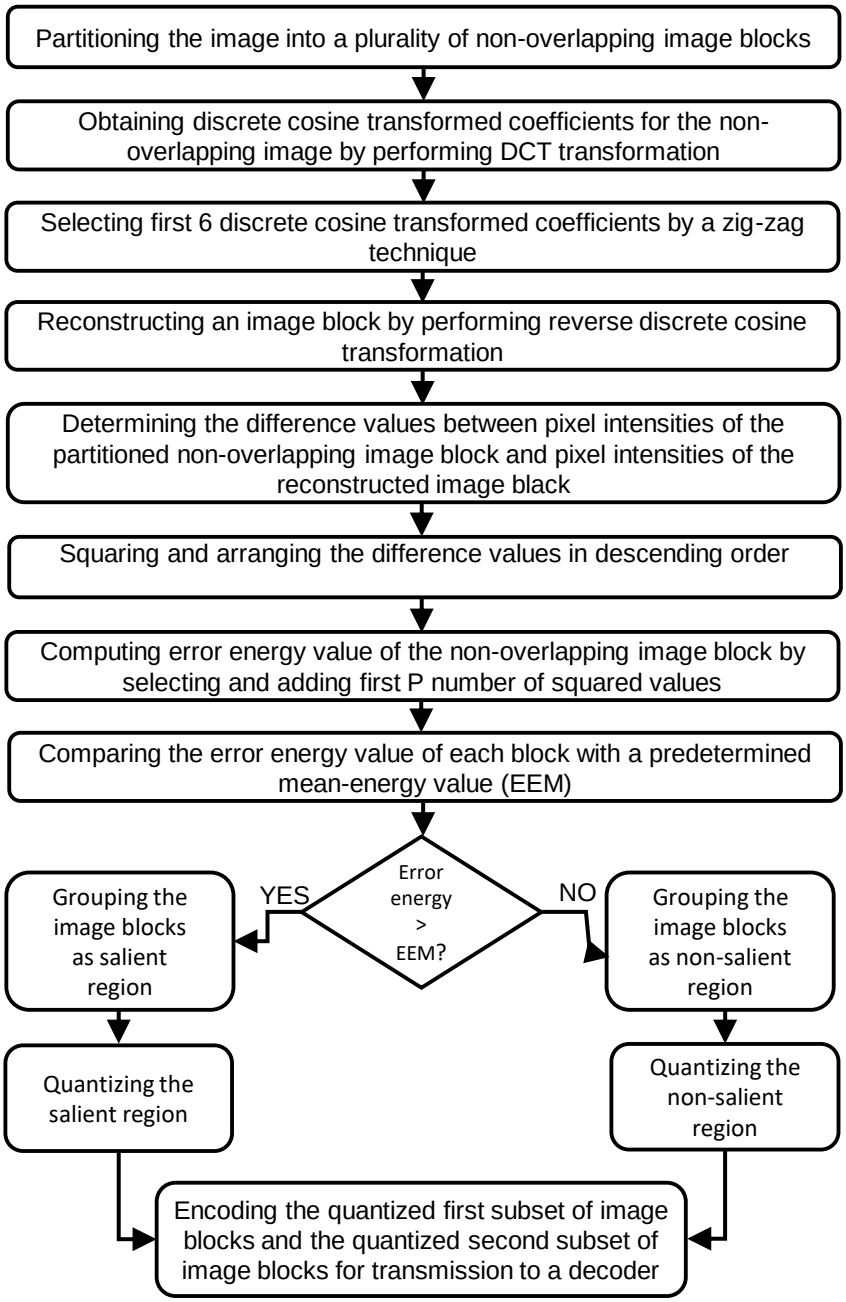

Figure 3.2 : Flow chart of the proposed method

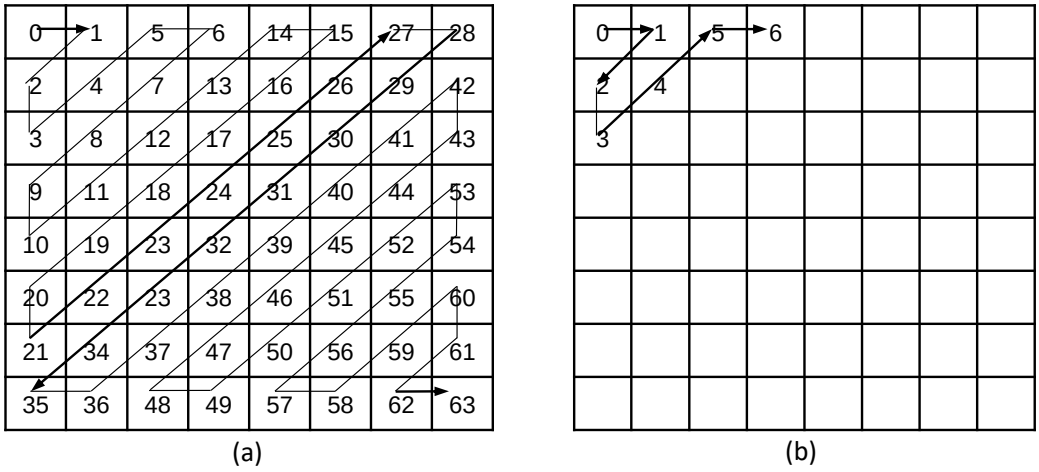

Figure 3.3 : (a) Traditional JPEG's zig-Zag scanning order, (b) Proposed scanning order

textual and non-textual components from a compound image are mainly based on, edge, absolute-deviation, variance, and DCT [Chaddha *et al.*, 1994]. The DCT based approach is found to offer the best accuracy and robustness in order to extract the textual or salient information from such images. Another advantage of using DCT based approach is that it is compatible with the standards such as JPEG, HEVC and can be flawlessly plugged in with a small computation overhead.

In order to segment the SCI into text and non-text regions, the input image is first divided into non-overlapping blocks of size 8×8. 2D-DCT is applied on each image block $B(x,y)$ to get $\hat{B}(x,y)$. All the basis functions or the 2-D DCT coefficients are ordered in conventional zig-zag order [Wallace, 1992]. To extract the smooth model, first, K coefficients are chosen as follows:

3.1.2 Error Energy Evaluation

The key step to efficiently find the salient region in the SCI is to choose an appropriate set of DCT coefficients. The proposed method evaluates the error energy between the original block and the reconstructed block by taking inverse DCT (I-DCT) with first K coefficients of conventional zig-zag order. So the problem boils down to choose a value of K which can efficiently detect the salient region. For this, statistical analysis has been conducted on the images available in two SCI databases [Yang *et al.*, 2015; Wang *et al.*, 2016].

The linear correlation between the error signal with K coefficients and with $K+1$ coefficients were evaluated. Firstly, the sum of squared error (SSE) for every non-overlapping 8×8 blocks of the error signal obtained using K coefficients are evaluated and represented as an error-energy matrix (EEM). For an error signal of size $M \times N$ the EEM has a size of $\frac{M}{8} \times \frac{N}{8}$. To analyze the effect of changing the value of K, the most widely used Pearson product-moment correlation coefficient (PPMCC) or the bi-variate correlation is applied. PPMCC helped in analyzing the linear correlation between the EEM's obtained using K and $K+1$ DCT coefficients where $K = 1, 2, 3, ..., 63$ as given in (3.1).

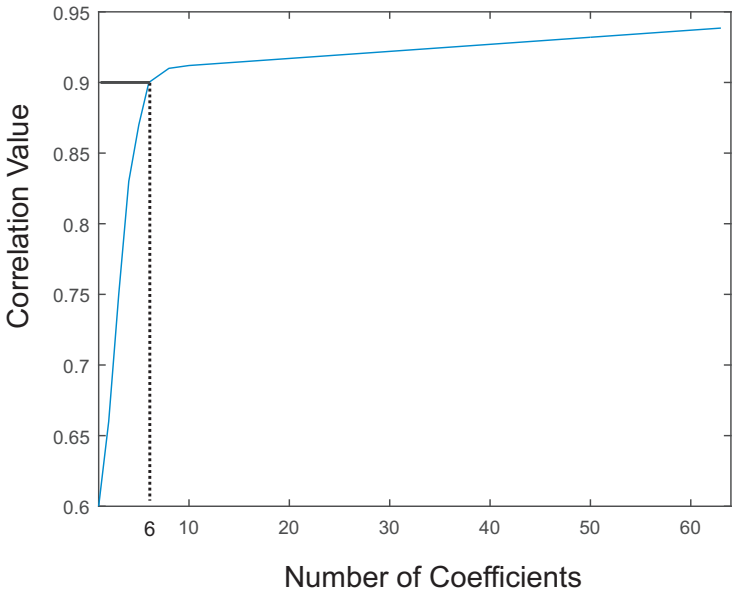

Figure 3.4 : Coefficients correlation analysis

$$corr_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}.\sqrt{var(y)}} \tag{3.1}$$

here the bi-variate correlation coefficient between the two variables x and y is denoted by $corr_{xy}$. There are two major reasons behind choosing this correlation coefficient in this context. The first reason is that as DCT coefficients are frequency dependent, and both frequency and bi-variate correlation works on first and second order moments [Yang *et al.*, 2016]. The other reason is that bi-variate correlation coefficient is invariant under a separate changes in both the variables. To illustrate this, it is possible to transform x to $a+b.x$ and transform y to $c+d.y$, where a,b,c, and d are constants with $b,d>0$, without the need to change the correlation coefficients.

The range of correlation coefficients $(corr_{xy})$ is between 0 (no linear relationship) to 1 (perfect positive linear relationship) or -1 (perfect negative linear relationship). Positive linear relationship signifies a direct relationship, i.e. if the value of one variable increases, then the other variable will also increase. On the other hand, a negative linear relationship indicates an indirect relationship, i.e. if the value of one variable increases, the other variable will decrease.

Figure 3.4 represents the plot between the number of coefficients (K) and the correlation between the EEM obtained with K coefficients and $(K+1)$ coefficients. For example, the correlation value correspond to number n represent the correlation between the error image with n and $n+1$ coefficients, where $n=1,2,3...,63$. It can be observed from Figure 3.4 that the rate of change in correlation for the number of coefficients more than 6 is considerably less than that of less than 6. On account of these observations, the value for K is set at 6.

3.1.3 Salient Region Identification

The proposed method presents an idea of evaluating block based two level saliency, i.e. instead of mapping saliency to each pixel of the image, the proposed method evaluates saliency map for the 8×8 pixel blocks. This idea enables the proposed method to be seamlessly plugged

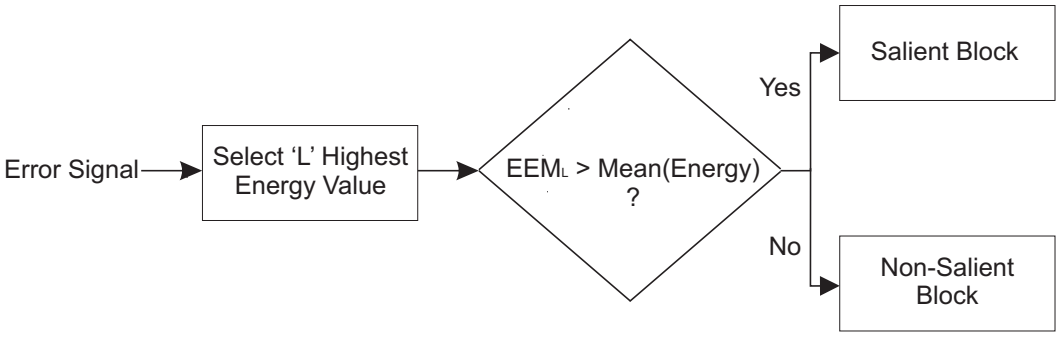

Figure 3.5 : Salient Region Identification Process

in with the JPEG baseline [Wallace, 1992] in order to upgrade it to do a judicious quantization for every 8×8 blocks. Figure 3.5 demonstrates the saliency identification process for every blocks.

The error signal obtained with an optimal K coefficients as discussed in Section 3.1.2 is first divided into non-overlapping 8×8 blocks. For every block, the highest L absolute energy values are chosen. There may be some cases when salient regions are partially present in the 8×8 block, i.e. a small border portion of the salient region falls within the block. If the influence of non-salient region is dominant on such blocks, it may get assigned as a non-salient block and further gets highly distorted. This may cause some textual part in the SCI to be illegible. Choosing L highest values for further analysis in order to decide saliency of the block helps to avoid such situations as it gives priority to high-frequency textural regions which are supposed to have higher error values than the non-textual regions. The choice to evaluate an optimal L value is carried out using bi-variate correlation or PPMCC on the similar lines as discussed in Section 3.1.2. The correlation is calculated between the error-energy matrix (EEM) obtained from all 64 energy values of an 8×8 block (EEM_{64}) and EEM with L highest values EEM_L. The analysis of average correlation values by changing the value of L for SCI's in [Wang *et al.*, 2016; Yang *et al.*, 2015] is shown in Figure 3.6. It is worth noting that after $L = 40$ the correlation between EEM_{64} and EEM_L achieves perfect positive linear relationship. In this way, $L = 40$ may be chosen as an optimal value for the further saliency identification process. It was observed that the segmentation results for $L > 40$ include all the textual blocks as salient. However, some of the non-textual blocks are also included as the salient blocks.

After getting optimal value for L, the saliency value of the block is evaluated on the basis its corresponding EEM_L value. If the EEM_L value of the block is more than that of the average EEM_{64} or the mean global error energy, then the block is mapped as salient. The rank value $R = 1$ is assigned to the salient blocks, and $R = 0$ to the non-salient blocks. The efficiency of the proposed saliency detection method is discussed in Section 3.2.2.

3.1.4 Adaptive Quantization

The block-wise saliency map is used for adaptive quantization. This is the key upgrade in the JPEG baseline framework which enables it to judiciously quantize the DCT blocks as per their importance, estimated in terms their rank values ($R = \{0, 1\}$). The base quantization table used in JPEG baseline (T_{50}) [Wallace, 1992] is proposed to be used for salient blocks ($R = 1$), and the same is scaled by a factor S ($whereS > 1$) for the non-salient blocks ($R = 0$). This means that the non-salient blocks are quantized higher than the salient blocks. The value of S can be changed according to the bit-rate requirements, i.e. the higher the value of S, the lower the bit-rate will be. Section 3.2.3 discusses the effect of change in S value.

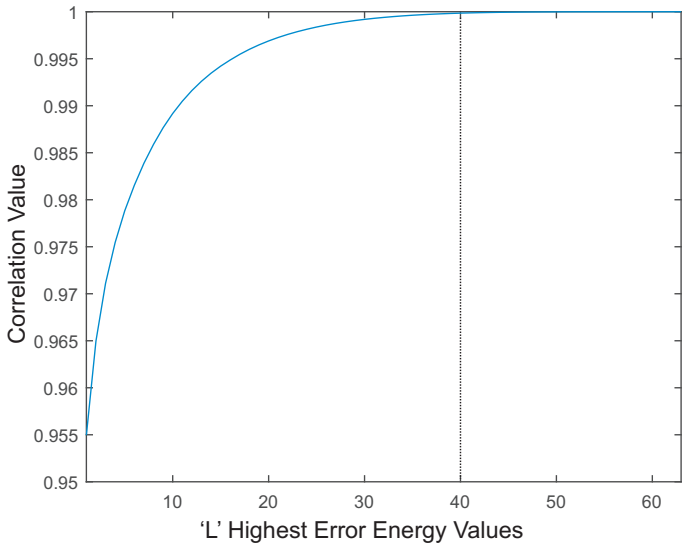

Figure 3.6 : Error Energy Correlation Analysis

Table 3.1 : Benchmark test databases for SCI IQA

Database	Reference Images	Distorted Images	Distortion Types	Image Type	Observers
QACS	24	492	2	Color	20
SIQAD	20	980	7	Color	96

3.1.5 Overhead Encoding

The number of bits required for sending the saliency map information to the decoder in terms of the block rank (R) will be 1 bits-per-block and its value in terms of bits-per-pixel (bpp) will be $\frac{1}{16}$ or 0.0156. This minimal overhead can easily be accommodated in order to achieve an upgraded performance in JPEG baseline.

3.2 RESULTS AND DISCUSSIONS

3.2.1 Protocol

In order to evaluate the robustness and efficiency of the proposed method, two publically available datasets namely, QACS [Wang *et al.*, 2016] and SIQAD [Yang *et al.*, 2015] are used. The QACS dataset contains 24 reference SCI's and 492 distorted SCI's after HEVC and HEVC-SCC [Xu *et al.*, 2016] designed for SCI compression and is helpful to analyze the performance of SCI compression techniques. The SIQAD contains 20 reference SCI's and 980 distorted SCI's after Gaussian noise, Gaussian blur, motion blur, contrast change, JPEG, JPEG 2000, and layer segmentation compression. The characteristics of these two databases are provided in Table 3.1. The performance of the proposed method is compared with JPEG baseline [Wallace, 1992], state-of-the-art HEVC, and HEVC screen-content-coding (HEVC_SCC) [Xu *et al.*, 2016]. To evaluate the quality of the reconstructed image PSNR [Wang and Bovik, 2002], and [Wang *et al.*, 2004] is used.

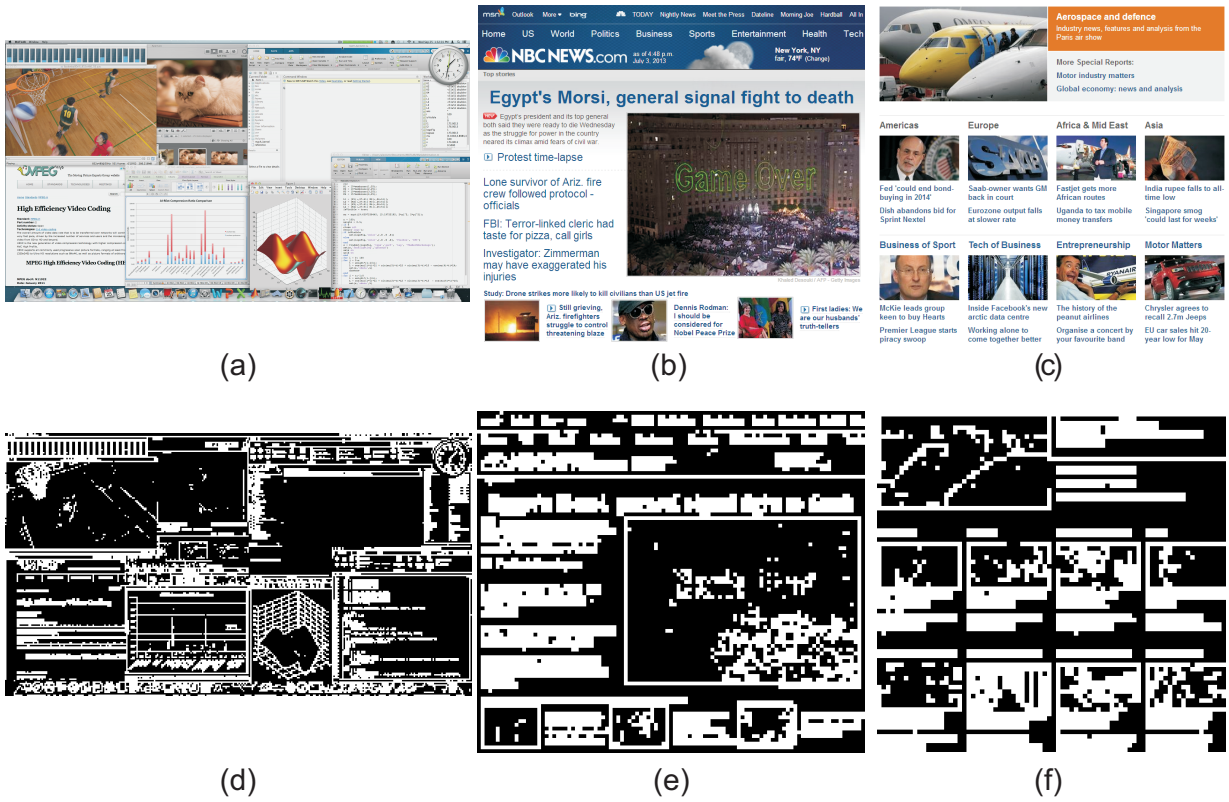

Figure 3.7 : Segmentation performance of the proposed method
(a)-(c) Input SCIs, (d)-(f) Segmented image after applying the proposed method on (a)-(c) respectively.

3.2.2 Performance of Saliency Detection

Figure 3.7 shows the performance of the developed two-level saliency ranking. The input SCIs are shown in Figure 3.7 (a) -(c), and the segmentation result is shown in Figure 3.7 (d) - (f). For the illustration purpose, the saliency of the regions is shown with a binary image where the white region implies the salient region and black implies non-salient. After applying the proposed saliency detection method on all the reference images in [Yang *et al.*, 2015; Wang *et al.*, 2016], it was observed that the proposed method was able to detect 100% textual region as salient. It was also observed that some of the other information like strong edges were also captured as salient region along with the textual one. This also helped the proposed method to preserve the edges in an image along with the textual region.

The comparison of performance between the proposed method and the state-of-the-art methods with respect to the quality of the reconstructed image at the same bitrate is shown in Figure 3.8. The uncompressed SCI is shown in Figure 3.8 (a), and subsequently the reconstructed images after applying JPEG baseline [Wallace, 1992], JPEG-2000 [Skodras *et al.*, 2001], HEVC [Sullivan *et al.*, 2012], HEVC_SCC [Xu *et al.*, 2016] and the proposed method at 0.8 bpp are shown from Figure 3.8 (b) to (f), respectively. For better visual comparison, the important textual regions from two different portions from the input SCI is also zoomed. The zoomed portion in the left side is a mixture of text with different font size, color and style, and the other portion is part of some codes with same fort size and style. It is observed that the developed method can judiciously compress the SCI without affecting the textual portions, regardless of the text size, color or font. On the other hand, JPEG and HEVC show almost similar results and can only save the large sized text. Moreover, JPEG-2000 and HEVC_SCC shows better performance than JPEG and HEVC, but

these are also unable for a legible reconstruction of the codes area or the right portion. The codes are only legible in the reconstructed image of the proposed method. This observation shows the superiority of the proposed method in order to reconstruct an SCI.

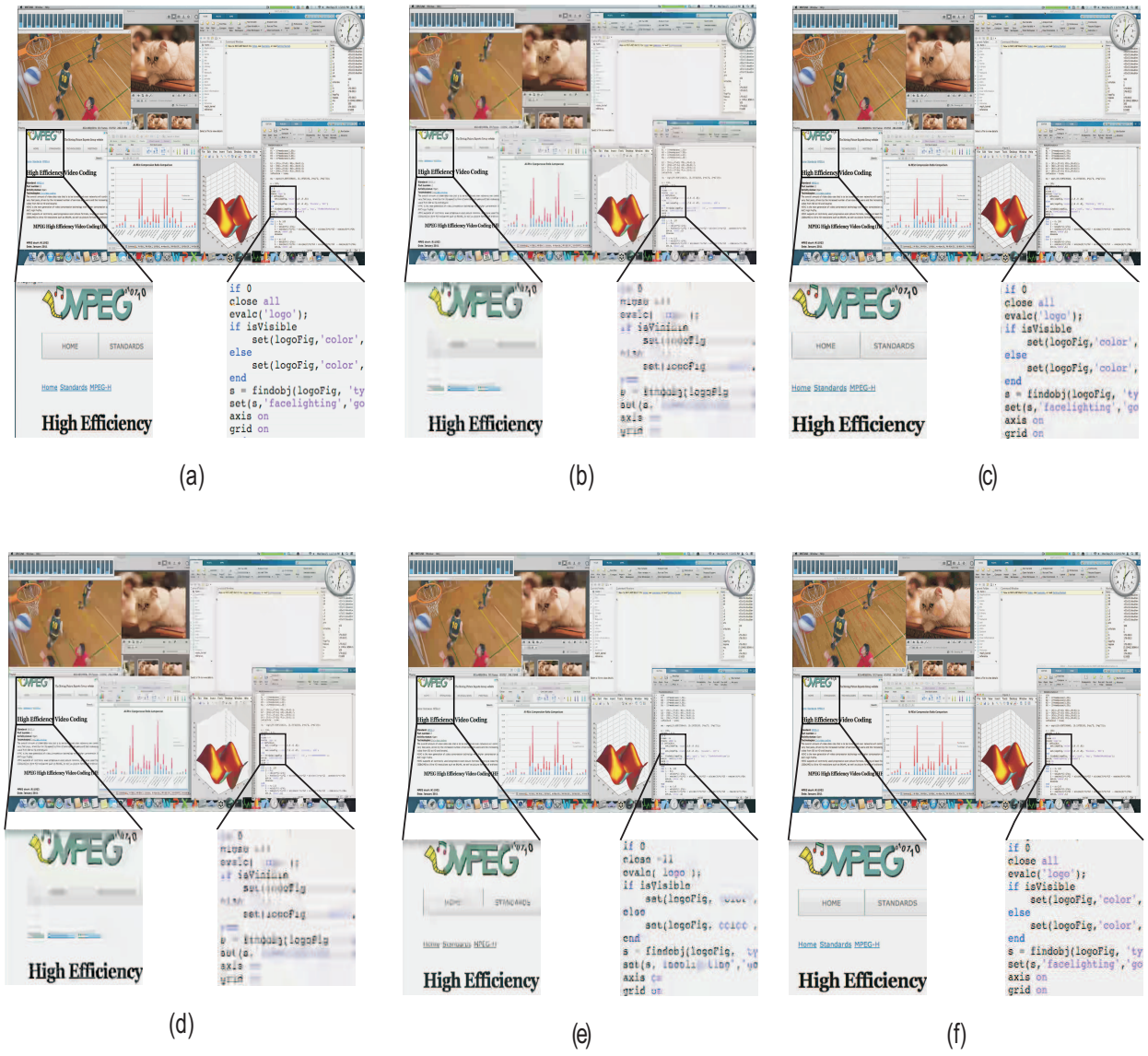

Figure 3.8 : Comparison on quality of the reconstructed image at 0.8 bpp by using different methods (a) Original SCI, (b) After applying JPEG [Wallace, 1992], (c) After applying JPEG 2000 [Skodras *et al.*, 2001], (d) After applying HEVC [Sullivan *et al.*, 2012], (e) After applying HEVC for SCC [Xu *et al.*, 2016], and (f) After applying the proposed method.

3.2.3 Rate Distortion Analysis

In order to analyze the overall performance of the proposed method, three different rate-distortion curve has been shown in Figure 3.9 to Figure 3.11. The rate-distortion curve in Figure 3.9, shows the performance comparison between the JPEG baseline [Wallace, 1992], JPEG-2000 [Skodras *et al.*, 2001], and the proposed method on the overall image, between bitrate 1 to 1.5. It is obvious that the PSNR of the proposed method will always be less than the JPEG baseline for a particular bitrate due to the overhead in sending the saliency map to the decoder. However, it can be observed that the performance of the proposed method approaches towards the JPEG baseline for lower bitrate. This is because of the fact that the proposed method is saving

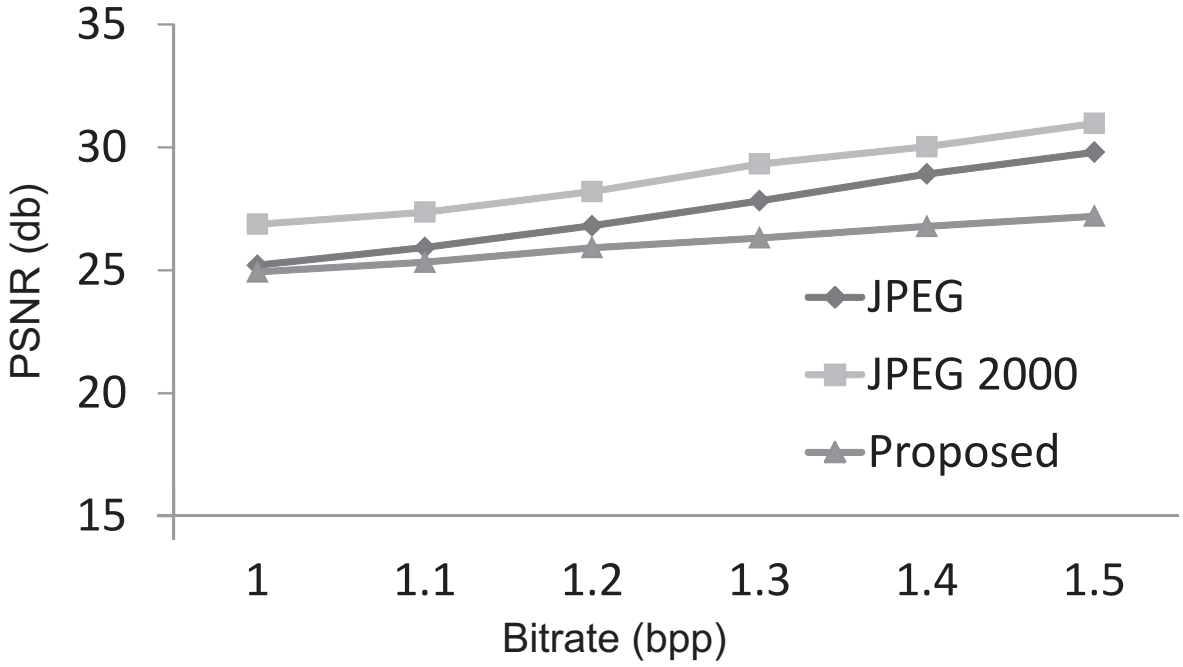

Figure 3.9 : Rate-distortion comparison between proposed method, JPEG baseline, and JPEG 2000.

the high-frequency textual regions from getting quantized heavily.

Figure 3.10 shows the region-wise performance comparison between the proposed method and JPEG baseline [Wallace, 1992]. It is observed that the PSNR of the proposed method is always higher compared to the JPEG for the image's salient regions $R1$ which constitutes an average of 35.4% of the total area (or the number of pixels) of the test SCIs. Moreover, the performance of the proposed method degrades at non-salient regions. However, since there is a significant improvement in quality at the salient regions i.e. the regions of perceptual importance, the overall visual quality of the images has improved, as shown in Figure 3.8.

Region-wise effect of changing the parameter S to achieve a bitrate between 1 to 1.5 can be observed in Table 3.2. Increasing the parameter S yields distortion only in the non-salient regions (i.e. $R2$) without affecting the most salient regions, which is about 35.4% of the image. Moreover, by decreasing the parameter S, the results behave like JPEG baseline as can also be observed from Figure 3.10. So, by changing the parameters S and Q_{am}, we get adequate flexibility in controlling quality and compression ratio than that can be obtained by simply applying the JPEG baseline method. The last three columns in Table 3.2 shows the effect of varying S on the whole image.

Due to the different properties of SCIs than CCIs as discussed in Chapter 1, many image quality assessment (IQA) model for SCI has been proposed over the past few years. The method FQI in [Rahul and Tiwari, 2019a] is found to have the most accurate distortion measure capability for compressed images than other such SCI-IQA methods. FQI is low-level feature based IQA method which evaluates the loss in feature for a distorted image. Figure 3.11 shows the rate-distortion curve with respect to FQI. It can be observed that the proposed method is always performing better than the JPEG. This is due to the fact that the low-level features are preserved in the reconstructed image after applying the proposed method.

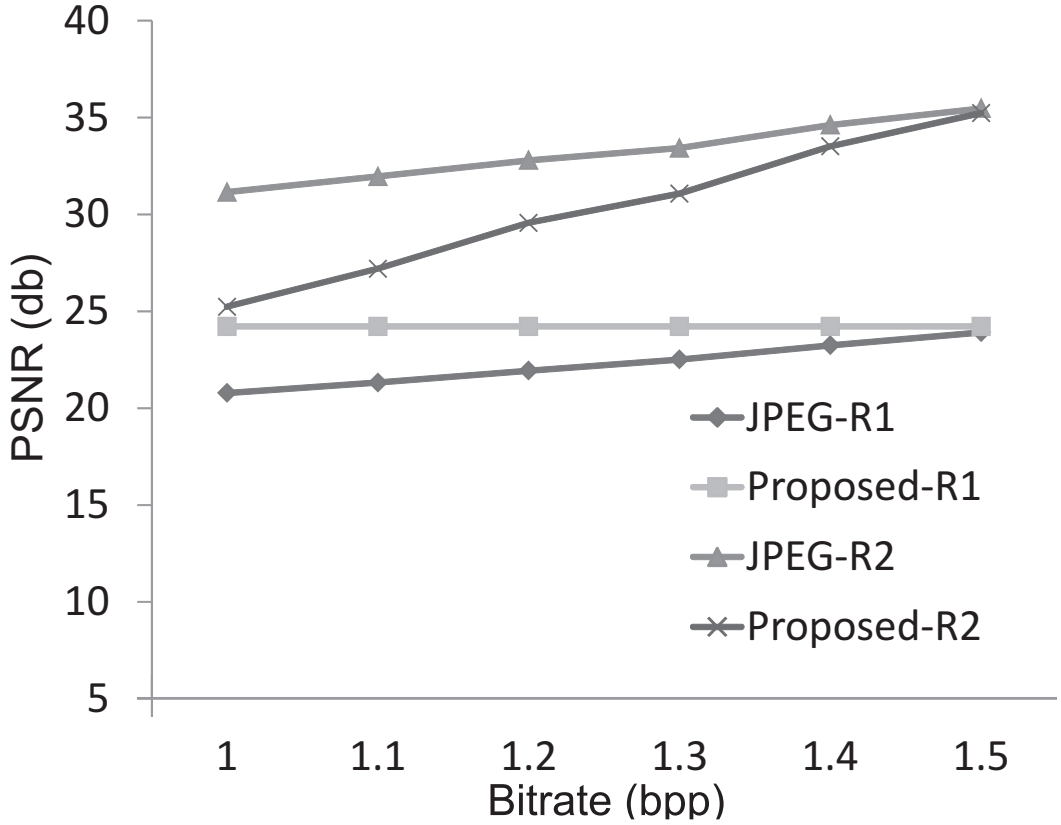

Figure 3.10 : Region wise rate-distortion comparison between proposed method and JPEG baseline. Here R1 indicates the salient region and R2 indicates non-salient region.

Table 3.2 : Effect of changing the parameter S in order to obtain different bitrate (bpp).

S *Value*	*Bitrate* *bpp*	**PSNR** $\mathbf{R_1}$	**PSNR** $\mathbf{R_2}$	**PSNR** **Overall**	**SSIM** **Overall**	**FQI** **Overall**
3.5	1.0	24.22	25.24	24.93	0.8778	0.5618
3.1	1.1	24.22	27.19	25.32	0.8868	0.5913
2.6	1.2	24.22	29.56	25.91	0.8951	0.6217
2.3	1.3	24.22	31.07	26.31	0.9063	0.6624
1.9	1.4	24.22	33.51	26.78	0.9214	0.6919
1.4	1.5	24.22	35.23	27.19	0.9321	0.7190
Region Weight (%)		35.4	64.6	100	100	100

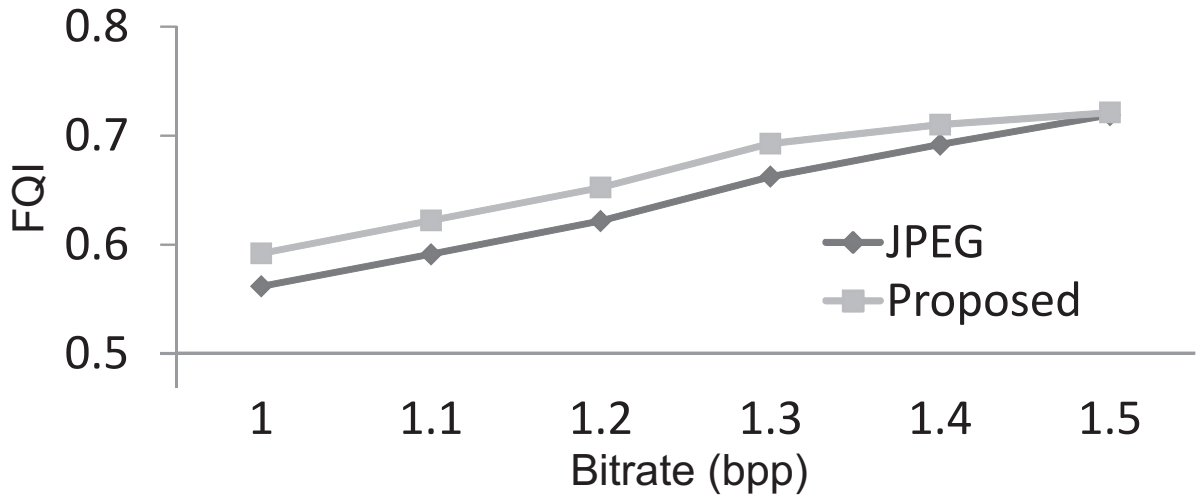

Figure 3.11 : Rate-distortion comparison with respect to feature quality index (FQI), between proposed method and JPEG baseline.

3.3 CONCLUSIONS

For the requirement of high compression ratio without any significant quality loss in salient textual regions of the reconstructed SCIs, we propose to classify a given image into two ranked regions and quantize the corresponding DCT coefficients judiciously. An error signal is produced with the difference of the original and the reconstructed image with K DCT coefficients. The saliency of the block is then evaluated by comparing the sum of L highest squared errors (EEM_L) in the corresponding block with the global mean of the squared error $mean(EEM_{64})$.

We achieved (average) 15.1% better quality at the most salient regions, which contained an average of 35.4% area in the test images of two SCI datasets, SIQAD and QACS. Due to improvement in the salient regions, the overall perceptual quality of the reconstructed image is better than the other state-of-the-art methods for SCIs. The experimental results obtained on different SCIs clearly showed that the proposed method outperforms the recently published similar method (HEVC_SCC) in terms of the perceptual quality of the reconstructed images. By ranking every 8×8 block instead of every pixel, we were able to send the saliency map to the decoder with only 0.0625 bpp overhead. The reconstruction of the saliency map at the decoder side is simple and more accurate than the recent state-of-the-art works where the ROI is approximated by a rectangular bounding box.

…