Hierarchical mask creation for intelligent image coding using saliency maps

R. Vargic and J. Polec

***Abstract*—In this paper we analyze basic mask creation methods for intelligent image coding using saliency maps. For saliency maps based image coding we use specific extension of SPIHT algorithm called SM SPIHT related to region of interest encoding but extending this approach further, ending with individual weight of importance for each pixel in image using the form of saliency map. This approach is proved to be effective. In this article we analyze impact of different basic hierarchical mask creation methods, which have impact on error separation between salient and not salient parts of the image. The results indicate that proposed mask creation method outperforms JPEG2000 based mask tree creation method.**

***Keywords*—saliency, wavelet, image compression, hierarchical mask, SPIHT.**

I. Introduction

In recent years, saliency information in video and image, its presence and exploitation is attracting the attention not only in image and video compression area but also in human computer interaction (HCI) and multimodal interfaces area, where the saliency can be one of inputs influencing the HCI. In the static image compression area, the wavelet based approaches are among the most successful. Well-known references are SPIHT [1] (Set Partitioning In Hierarchical Trees) and standardized JPEG 2000 algorithm [2]. Both of them were extended for classical region of interest (ROI) oriented coding. In classical ROI encoding one or more ROIs get certain advantage in sense of their bit budget over non-ROI areas but as the number of ROIs increases, the efficiency of the process decreases.

The SM SPITH method [3] [4] takes in account the individual significance of each pixel of the image in the encoding/decoding process. The significance information is expressed in the form of saliency map (SM). The "intelligence" of this approach lies in the generalization of ROI approach: first - describe what is important (significant) in the image with as much freedom as possible, second - encode the image accordingly, i.e. allocate from available bit budget more bits for more important pixels. The original SM SPIHT method was also extended to JPEG 2000 [5].

The paper focuses on detail, how the SM is embedded in the encoding process in original SM SPIHT in the form of hierarchical mask and what are the limits of the approach.

The paper is divided in four sections, first we discuss in more detail the differences between ROI and SM approach, then we explain in deep the steps in the SM SPIHT approach followed by details regarding the mask tree creation methods. Lastly the performance results are presented and discussed.

II. Region oriented coding and saliency maps

In the classical ROI coding one or more regions are defined and their importance is stated.

In JPEG 2000 there are defined 2 different ROI methods [2], maxshift and general scaling. In maxshift approach, spectral coefficients belonging to particular ROI are shifted in the sense of bit planes clearly over the other coefficients. Based on their value the decoder can distinguish them from other coefficients (and shift them back) so there is no need to encode the ROI shape information. The general scaling approach shifts the spectral coefficients belonging to particular ROI only particularly, so they overlap in value with other coefficients and encoding of the ROI shape is needed to distinguish into which group the particular coefficient belongs and shift the ROI coefficients back.

There are several extensions of these concepts [6] [7] [8] but none of them handles the significance of each image pixels separately in the sense of significance map. This approach can be simulated by many ROIs with different shift, but as the number of ROIs increases, the efficiency of the process decreases.

The aims of the proposed method are to possibly and effectively take in account the individual significance of each pixel of the image. The significance information of the pixels of the evaluated image is expressed in the form of SM, which is 8-bit gray scale image with the same dimensions as the evaluated image. Its pixel values contain the importance of the corresponding pixels of the evaluated image (0=no importance, ..., 255=highest importance).

III. Details of SM SPIHT approach

The SM SPIHT algorithm [3] [4] addresses the key question "how to pass to the encoder the side information about importance" of the particular pixels using the saliency map. The basis is the well-known and recognized SPIHT [1] algorithm, coupled with biorthogonal discrete wavelet transform (DWT) and famous 9/7 tap filter known as bior4.4 in the Matlab community (the filter pair has four zero moments in both, decomposition and reconstruction parts of

Manuscript received August 5, 2016; revised October 10, 2016.

This paper was partly supported by the Slovak Research Grant Agency VEGA under grant No. 1/0625/14 and grant No. 1/0789/15.

Radoslav Vargic works at Department of Telecommunications at Slovak Technical University in Bratislava. Tel: +421 2 60291415; e-mail: Radoslav.vargic@stuba.sk.

Jaroslav Polec works at Department of Telecommunications at Slovak Technical University in Bratislava. Tel: +421 2 60291409; e-mail: jaroslav.polec@stuba.sk

doi: 10.11601/ijates.v5i3.191

corresponding filter bank). The DWT uses nonstandard decomposition, i.e. one decomposition level to low pass (L) subband and high pass (H) subband is applied on each row, then in each column resulting in the LL, LH, HL and HH subbands. The decomposition is then repeated only on the LL subband. The basic steps in the the wavelet based SPIHT encoding algorithm are:

1. 2D DWT with nonstandard decomposition using 9/7 filter is performed
2. SPIHT encoding of the spectrum including arithmetic encoding of the resulting stream is performed

As in EBCOT (Embedded Block Coding with Optimized Truncation) in JPEG 2000 uses SPIHT bit plane coding, so the significance of the coefficients is given by his value. The significance information for the encoder and decoder in the SM SIPHT is provided as side information. The saliency map for the decoder is encoded using separate SPIHT path as normal image. The whole SM SPIHT process in encoding and decoding phase is depicted in Fig. 1.

In the encoding phase the saliency map has to be prepared e.g. as in [3] [4]. After that it is needed the saliency mask encoding and immediate back-decoding, to have the same information in image encoder as in the decoder. From the decoded saliency mask, the mask tree is derived. The derivation of the mask tree is the main topic of this article and will be separately discussed in the next chapter.

The form of hierarchical saliency mask (HSM) has to have the same form as the subbands in 2D nonstandard wavelet spectrum. This particular form of the hierarchical mask is important from the viewpoint that all spectral coefficients that influence the same pixel in the reconstructed image, shall reflect the importance of that pixel. The correspondence of the particular coefficients to the same spatial location is well known feature of the 2D nonstandard decomposition spectral structure.

The mask tree is applied to the image spectrum before the encoder starts the bit plane encoding process of the spectral coefficients. We shift the spectral coefficients $SC(i,j)$ depending on their significance (0-255) expressed in the form of saliency map HSM(i,j) in the using the formula

$$SC(i,j)_{new} = \frac{SC(i,j)_{old}}{1 + str\,\frac{255 - HSM(i,j)}{255}}, \qquad (1)$$

where the *str* is configurable strength parameter in the range 0 - 255. When strength is set to 0, the SM SPIHT algorithm is equivalent to normal SPIHT algorithm. When strength is set do 255, then the least significant coefficients are divided by 256, so they are shifted 8 bit planes lower. After the spectrum is masked, it can be encoded using the original SPIHT algorithm.

In the decoder the hierarchical mask tree shall be constructed and the weights created using the saliency mask shall be applied using the following reverse formula to approximate the original spectrum values

$$SC(i,j)_{new} = SC(i,j)_{old}\left(1 + str\,\frac{255 - HSM(i,j)}{255}\right). \qquad (2)$$

Fig. 1. Schematic picture of whole SM SPIHT encoding and decoding process

After that the inverse DWT can be applied to spectrum to finally get the decoded image.

IV. HIERARCHICAL MASK CREATION

There are many options how to create the hierarchical mask, with respect which sub band levels shall be suppressed and which amplified. In the SM SPIHT approach [3], [4] the hierarchical mask was prepared using the averaging mask with following rules:

1) *mask pixels in the lower subband N were created by averaging the corresponding 4 pixels in the subband N+1, i.e. as if the Haar low pass filter would be used.*
2) *the same mask dynamic (values 0-255) was preserved across the sub bands.*
3) *all the 3 spatial tree orientations were handled equally*

We refer to this mask tree creation method as Method A (MA). In optimal conditions the mask image can be delivered to the decoder losslessly (of course bigger bit budget would be needed). We refer to this case as MA LSM (MA with Lossless Saliency Mask).

The second considered basic method for hierarchical mask creation is inspired by the method used JPEG 2000 [9] (Part 2, Annex K). Here for both ROI based methods it is important to know, which spectral coefficient influences the ROI and which not. The rule is simple: Select all spectral coefficients that could have non zero influence to the ROI.

The implementation rule is straightforward: Perform the inverse 2D discrete wavelet transform back – doing all operations in reversed order, including order of steps in lifting scheme and directions of arrows in lifting scheme steps, …). Note:

1) This (rather complicated) kind of implementation implements the question "given the set of spatial points, by which set of spectral points is affected by that set of spatial points in the inverse DWT process?"
2) If we would just simply use forward transform of the saliency map, we would get the answer for the following questions: "given the set of spatial points, which set of spectral points in the DWT process it affects"?

These questions are not same and also the sets of spectral coefficients are not same. We can illustrate the difference by the example on Fig. 2.

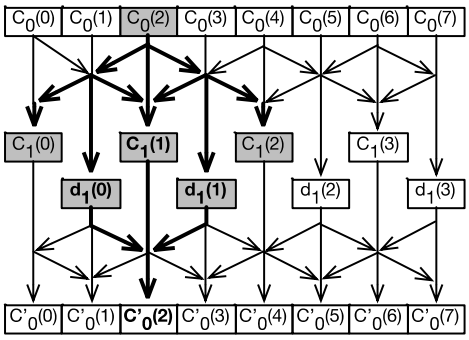

Fig. 2. Example of typical liftings steps in the DWT and inverse DWT. One can see that coefficient $c_0(2)$ affects 5 spectral coefficients, but in inverse DWT it is reconstructed from (affected by) only 3 spectral coefficients.

In JPEG 2000 conveys the resulting hierarchical mask just the binary information - which spectral coefficients could be needed and which not. In the saliency map approach this is not sufficient as the typical example are slowly changing values in the Saliency map. Resulting HSM has to be smooth as well, extending the binary information (important/not important) to how important the particular spectral coefficient is. This can be basically achieved by filtering using, where the 9/7 filters have the filter coefficients set equally to 1/7 in the case of 7 tap and 1/9 in the case of 9 tap filter. We refer to this mask creation method as Method B (MB).

From this method we derive also the MB LSM variant as it was the case in Method A.

In case we use lossless mask and the original saliency mask has binary form (not smooth), then we can use binary form of HSM and use exact the algorithm as in JPEG2000 (and abandon the filtering approach in MB). We refer to this method as MBO (MB optimized). Again – this method is applicable only for binary valued SM images and LSM mask has to be used. If the lossy mode for saliency mask in connection with MBO would be used, then even small nonzero values in non ROI area after mask reconstruction would indicate nonzero saliency and MBO declares them as significant. So we do not consider pure MBO method here.

We do not consider any further mask tree creation methods. In the next chapter pre provide performance results of abovementioned methods and outline the needed properties of another optimized method.

V. Results and Discussion

In this article we do not concentrate on situations with typical smooth saliency mask, they were evaluated in [3][4][5]. We concentrate on the border case, where the saliency mask has binary form – i.e. ROI with sharp edges exist. The best method shall have the best separation between the error in ROI compared to error in whole image.

a) Original Lena image b) very simple saliency map

Fig. 3. Image and saliency map used for experiments, a) Lena image b) very simple saliency map (equal to circular ROI in the Lena's face)

We compare the performance of all abovementioned masking methods. The experiments were performed using Lena image and typical circle shaped ROI as depicted on Fig. 3. We evaluated MSE and PSNR differences of the compressed image for the whole image and also taking in account she saliency information using the weighted MSE in the form:

$$MSE_{SM} = \frac{\sum_{i,j} \left(x(i,j) - \hat{x}(i,j) \right)^2 SM(i,j)}{\sum_{i,j} SM(i,j)}, \qquad (3)$$

where $x_{i,j}$ and $\hat{x}_{i,j}$ are pixels of the original and reconstructed image and $SM_{i,j}$ the pixels of the saliency map image. We apply this formula for the MSE computation in the ROI. The PSNR measure is derived from MSE using formula:

$$PSNR = 10 \log_{10} \frac{255^2}{MSE}. \qquad (4)$$

The result obtained for the method MA and its lossless variant MA LSM are depicted on Fig. 4. Both variants of MA method start to penalize the ROI at strength 10. Moreover, despite the expectation, the sharp saliency mask (available in the LSM mode) generally decreases the performance of the MA method then increasing the strength parameter.

The optimal strength value for the setup seems to be 7. The PSNR in ROI area is approximately the highest and by further increasing the strength the overall PSNR falls down.

The result obtained for the MB method its variants are depicted on Fig. 5. MB LSM and MB variants start to penalize the ROI at strength 10. The optimized variant continues to exploit the saliency of the ROI further. This leads us to important conclusion that the decrease of the performance in the ROI area at higher strengths is caused by low pass filtering during the hierarchical mask creation. The mask shape deformed and increasing the strength of the

process does not yield better results anymore (this explains also drop of the overall PSNR in the method A). The performance of the MBO LSM method is the best for the ROI area, however bigger strengths have to be used and we pay with heavy drop of the overall PSNR, so strength above 15 should be avoided.

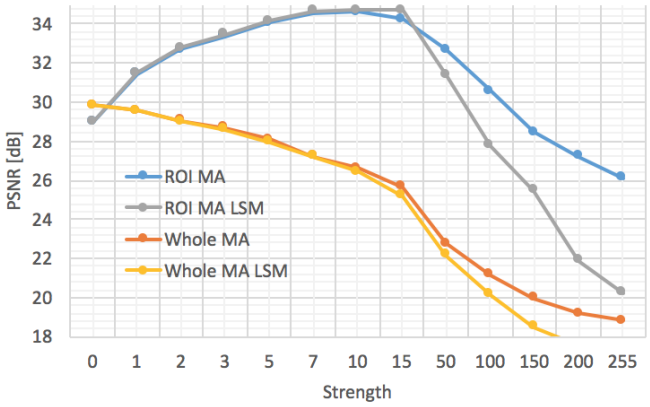

Fig. 4. Performance of the MA and MA LSM methods on Lena image compressed at 0.1 Bpp using different strength parameter. In MA is the mask lossy compressed to 0.01 Bpp. PSNR achieved in the ROI (weighted formula for MSE used) and in whole image can be compared.

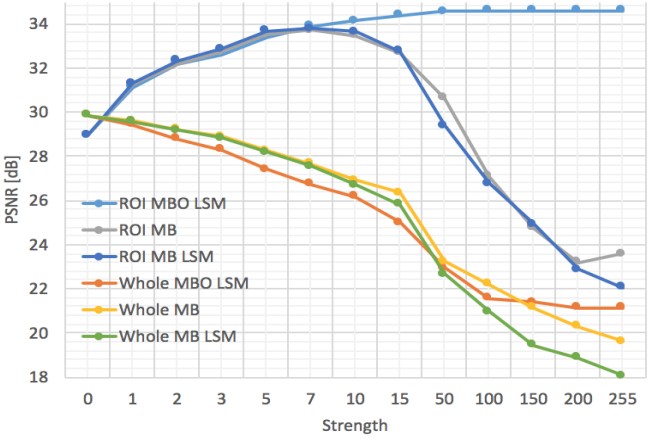

Fig. 5. Performance of the MB method and its variants on Lena image compressed at 0.1 bpp using different strength parameter. In MB is the mask lossy compressed to 0.01 bpp. PSNR achieved in the ROI (weighted formula for MSE used) and in whole image can be compared.

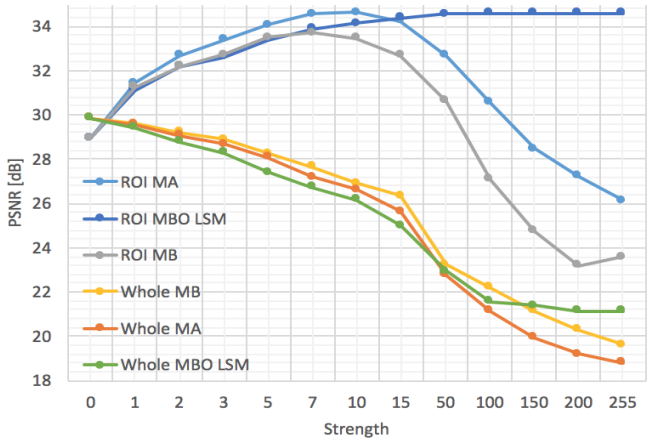

Fig. 6. Performance comparison of the MA and MB methods on Lena image compressed at 0.1 bpp using different strength parameter. In the lossy variants is the mask lossy compressed to 0.01 bpp.

Finally, the comparison of the best variants of MA and MB are given in the Fig. 6. As we can see, in the interesting part of the graph (strength <= 15) the MA clearly outperforms the MB methods – taking the same overall PSNR, the PSNR in the ROI is better for MA. Even more, the MA very slightly outperforms MBO LSM in the absolute PSNR achieved in the ROI. This leads us to important statement that is proven at least for tested image and bitrates: The JPEG 2000 algorithm makes sure that all spectral coefficients that could affect the ROI are taking in account, however this approach does not assure best performance from the rate/distortion sense. Simple averaging with as the MA method could achieve significantly better results.

Some representative results of compression are given in Fig. 7. All masking method enhance the facial area as expected. At strength = 7 the MA and MB perform visually similar, the 0,85 PSNR difference in the ROI is not visible. With strength set to maximum 255 the overall degradation is more visible in the MB than in MB LSM. Note, that though the MB method (as the MA method) has low PSNR in the face area, the facial details are preserved very good, there is notable only low pass distortion.

a) original image b) 0.1bpp, strength=0

a) 0.1bpp, MA, strength=7 b) 0.1bpp MB, strength=7

c) 0.1bpp MB, strength=255 d) 0.1bpp, MBO LSM, strength=255

Fig. 7. Examples of performance comparison of the presented mask creation methods using the mask from Fig. 5b, 0.1 bpp as target bitrate for the image and various strengths.

VI. CONCLUSION

In the presented contribution we demonstrate the performance of the selected basic hierarchical mask creation methods. The geometric/averaging principle is compared to "take all, that can have influence on" (JPEG 2000) principle. As the results show, the JPEG 2000 ROI approach can be outperformed in the rate/distortion sense by even simple averaging method. When the primary stress is really on the ROI area and overall PSNR is not important measure, there is probably space to further enhance the averaging method to better keep the ROI shape and do not drop down at strengths 7-15.

REFERENCES

[1] A. Said, and W. A. Pearlman, A new Fast and Efficient Image Codec Based on Set Partitioning in Hiererchical Trees, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 6, June 1996

[2] A. P. Bradley and F. W. M. Stentiford, JPEG 2000 and region of interest Coding, DICTA2002, 21-22. January 2002, Melbourne

[3] R. Vargic and J. Polec, Intelligent image coding using saliency map extended PSIHT algorithm, In *INES 2015: 19th International conference on intelligent engineering systems*. Bratislava, Slovakia. September 3-5, 2015. Danvers : IEEE, 2015, pp. 69-72.

[4] R. Vargic, J. Kučerová, J. Polec, Wavelet based image coding using saliency map, SPIE *Journal of Electronic Imaging*, Vol. 25, No. 6, 061610 (Nov/Dec 2016), published online

[5] R. Vargic, and Kačur, Image coding using saliency maps based on JPEG 2000, Redžúr 2016, *10th International workshop on multimedia and signal processing*. Bratislava, Slovakia. May 24, 2016.

[6] M. Penedo, W. A. Pearlman, P. G. Tahoces, M. Souto, J. J. Vidal, Region-Based Wavelet Coding Methods for Digital Mammography, *IEEE transactions on medical imaging*, Vol. 22, No. 10, October 2003, pp. 1288-1296.

[7] K. Martin, R. Lukac, K. N. Plataniotis, SPIHT-Based Coding of the Shape and Texture of Arbitrarily Shaped Visual Objects, *IEEE transaction on circuits and systems for video technology*, Vol. 16, No. 10, October 2006

[8] P. G. Tahoces, J. R. Varela, M. J. Lado, M. Souto, Image compression: Maxshift ROI encoding options in JPEG2000, *Computer Vision and Image Understanding*, Vol. 109, pg.139-145, 2008.

[9] JPEG 2000 image coding system: Extensions, ISO/IEC 15444-2:2004, International standard.

R**adoslav Vargic** was born in 1972 in Myjava. He has received the Ing and PhD Degree from Slovak Technical University in Bratislava in years 1995 and 1999 respectively. From that time, he works at Department of Telecommunications at Slovak Technical University in Bratislava. His research interests are wavelets and multimedia processing.

Jaroslav Polec was born in 1964 in Trstená, Slovakia. He received the Ing and PhD degrees in telecommunication engineering from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in 1987 and 1994, respectively. Since 1997 he has been associate professor and since 2007 professor at the Institute of Telecommunications of the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology and since 1998 at the Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics of the Comenius University. His research interests include Automatic Repeat Request (ARQ), channel modelling, image coding, interpolation and filtering.