Quality estimation based on interest points through hierarchical saliency maps

Michael Nauge, Mohamed-Chaker Larabi, Christine Fernandez-Maloigne

QUALITY ESTIMATION BASED ON INTEREST POINTS THROUGH HIERARCHICAL SALIENCY MAPS

Michael Nauge, Mohamed-Chaker Larabi senior member IEEE, Christine Fernandez

XLIM-SIC Lab., University of Poitiers, France

ABSTRACT

Quality of Experience (QoE) is a widely used notion nowadays because the end-user has been re-integrated in the quality loop. Subjective experiments are tedious and time consuming but to date they are the main way to have the human judgment. An important effort has been put on the development of metric estimating the QoE. So, in this paper, we propose a new image quality metric based on two concepts: the interest points and the objects saliency on color images. This metric is constructed by taking advantage of the variability of the interest points impact function of the image impairment and also the hierarchical attention of a human observer. This combination helps in giving more importance to the variability of the most salient regions and reducing the influence of regions having less visual importance.The results show a high correlation between the metric scores and the human judgment and a better quality range than well-known metrics.

***Index Terms*—** Image quality metric, Hierarchical saliency map, interest points.

1. INTRODUCTION

The need of tools for measuring quality of images/videos is an incontestable issue nowadays. This is due to the plethoric amount of visual data carried by users on hard disks and/or the Internet, and the emergent applications relying on transmission and compression for new display generations. From a consumer satisfaction point of view, one can notice the emergence of the notion of Quality of Experience (QoE) including the end-user in the loop at the opposite of what was done formerly. This new concept implies the use of quality assessment tools, not only objective (metrics) but subjective (testing) too, in order to monitor the systems.

The quality assessment literature is increasing in a considerable way. In addition to papers dealing with subjective paradigms, perceptual issues there are hundreds of papers introducing objective quality metrics dedicated to a large panel of image and video applications [1]. Basically, objective evaluation metrics can be categorized into three groups: full-reference (FR), no-reference (NR), and reduced-reference (RR) metrics. FR metrics often seen as fidelity metrics, need full information of the original images and demand ideal images as references which can be hardly achieved in practice. The traditional methods of FR (such as peak signal-to-noise-ratio PSNR) are based on pixel-wise error and have not always been in agreement with the perceived quality. Recently, some FR metrics based on human visual system (HVS) have been proposed like weighted signal-to-noise-ratio (WSNR) metric [2] by using the Contrast Sensitivity Function (CSF), to mimic the human early vision. Other metrics introduce complex HVS modeling, such as perceptual difference (Pdiff) metric [3]. So, in this metric the images are transformed from RGB to CIELAB. Then Barten's CSF and Daly's visual masking are applied. The pixel-wise difference between the original and impaired images is then computed, and the obtained difference is considered as perceptible if it is above a thresholds. Another approach consists in the use of natural scene statistics in the wavelet domain, such as Information Fidelity Criterion (IFC) metric [4]. This model captures two important, and complementary, distortion types: blur and additive noise. The IFC is the mutual information between the source and the distorted images. The IFC is not a distortion metric, but a fidelity criterion. It theoretically ranges from zero (no fidelity) to infinity (perfect fidelity). Wang et al. introduced in [5] a new framework for quality assessment based on the impairment of structural information. They developed a Structural SIMilarity metric (SSIM) and demonstrated its performance through a complete set of natural images.

At the opposite of FR metrics, NR metrics aim at evaluating distorted images without any cue from the source. However, most of the proposed NR quality metrics [6] are designed for one or a set of specific distortions and are unlikely to be generalized for evaluating images with other types of distortion. While RR [7] metrics are between FR and NR, they make use of a part of the information from original images in order to evaluate the visual quality of the distorted ones. The extracted features from the reference represent a small ratio with regards to the original size. This allows RR metrics to be used in very specific fields (transmission or monitoring) where the original data is not available.

In this paper, an image quality metric is constructed by taking advantage of the variability of the interest points impact function of the image impairment and also the hierarchical attention of a human observer. This combination helps in giving more impact to the variability of the most salient regions and reducing the influence of regions having less visual importance. For this, we construct a hierarchical saliency map that decomposes a given image into several layers thats sorts the content function of the visual attention. The obtained map is used to decompose the contribution of each layer in the quality assessment step. Then a weighting factor is assigned to each layer depending of its contribution to quality with the aim of increasing the correlation with human judgment. Thanks to construction of this metric, it can be oriented as a reduced-reference metric by providing the saliency layers and the scores of each one of them. The size of the reference is dependent on the number of layers.

The remainder of this paper is organized as follow: The next section is dedicated to the interest points and their evolution according to the impairment. Section 3 is dedicated to the construction of the hierarchical saliency map. The proposed metric is described in section 4 and, experimented and discussed in section 5. This paper ends with some conclusions and gives some future directions.

2. INTEREST POINTS EXTRACTION

As mentioned previously, it is very important to build an image metric having a behavior close to human judgment because, at the end, the visual content is intended to be used/manipulated/observed by the human observer. So it is very important, at a first stage, to understand the observers' response for different distortions. For very blurred images, it is difficult for a human observer to distinguish regions of interest because of them low acuity of contours. At the opposite, images with a significant blocking-effect, generate additional contours (vertical/horizontal) creating a confusion for the human observer. For the HVS, the previous artifacts may increase the cognitive task leading thus to a low judgment of the quality.

Furthermore, gaze points reflects what is important for a human observer. Several algorithms have been constructed following the concept of interest points, like Harris corner detector, SIFT detector or SURF detector even though there is no proof of correlation with observer's gaze points. Interest points are very important for the characterization of an object or a texture and allow the discrimination of shapes and objects. These algorithms are classically used in motion detection and object recognition. Few papers [8] studied the performance of these detectors in image matching task over various distortions. In general, these algorithms need to be invariant to scale, orientation and so on. However, they are lacking somehow for some impairments as compression artifacts like JPEG 2000 blurriness or JPEG blockiness effect. Figure 1 gives an example of the variation of the Harris interest points after a JPEG compression. It is easy to notice that their number decreases drastically when the compression ratio increases. This remark is very important for what will follow and is taken into account for the development of our metric. In our approach, Harris corner detector has been chosen for a sake of simplicity and for its higher sensitivity in comparison with the others detectors. It is based on the local auto-correlation function of a signal. The latter measures the local changes of the signal with patches shifted by a small amount in different directions. So, an interest point can be defined as the intersection of two edges. It can also be defined as a point for which there are two dominant and different edge directions in a local neighborhood of the point. The number and the position of detected corners can be tuned by different factors, such as neighborhood size, threshold on the corner response... As all the interest points detectors, Harris has several parameters to tune tune in order to address different problematics.

For the proposed metric, the objective was to select parameters that maximize the number of interest points and this because the score is based partially on their variation. For example, the loss of interest points in region of textures or contours (high activity) is useful from our quality estimation point of view. However, in flat regions (low activity), the interest points evolution is at the opposite because the compression artifacts may generate new interest points. From this remark, it seems very important to avoid to consider an image as one region but multiple regions with a given visual hierarchy. The next section is devoted to the construction of the perceptual hierarchy to deal with this issue.

3. HIERARCHICAL VISUAL SALIENCY

To predict salient regions, it is necessary to understand and model the HVS behavior with regards to local singularities. Saliency estimation tools are either based on biological informations or mathematical models. Some of them are based on both [9]. In general, it estimates the contrast of objects/regions using low-level criteria such as luminance, color and orientation. Two approaches exist in literature: the top-down and the bottom-up. One of the most used bottom-up model has been developed by Itti et al. [10]. It is based on a biological modeling and computes the saliency thanks to local features. To cope with the complexity of this model, Achanta et al. [9] proposed a very competitive and computational model based on color and luminance features. More precisely for each component (luminance and chroma), each pixel is compared to the average value of the whole image. This approach is very simple and gives acceptable results. For these reasons, we decided to extend the model for hierarchical visual saliency prediction based on HVS properties such luminance and color contrast sensitivity, central bias integration and temporal inhibition. In order to construct our saliency map (SM), the input image is transformed from RGB to CIELAB color space for its perceptual characteristics. After this step, Daly's CSF [11] is used by applying the achromatic CSF to the L channel and the chromatic CSF to a and b channels. The objective of this perceptual filtering is to apply the early vision properties on the achromatic and chromatics spatial frequencies and take into account the visualization conditions and simulate different viewing distances, screen sizes and resolutions. The saliency map is obtained by an Euclidean distance between a pixel and the mean of the filtered image (see Figure 2).

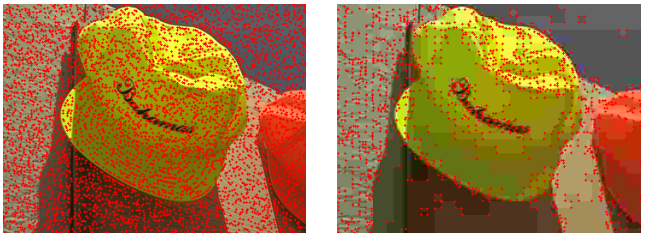


Fig. 1. Evolution of Harris Interest Points between the original image (left) and the JPEG compressed one (right)

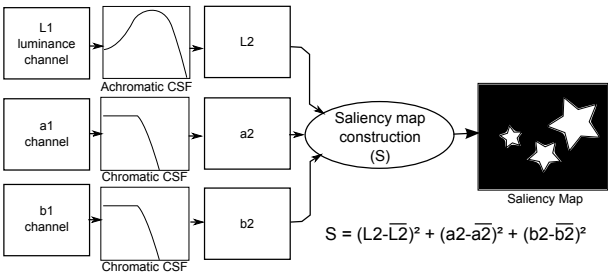
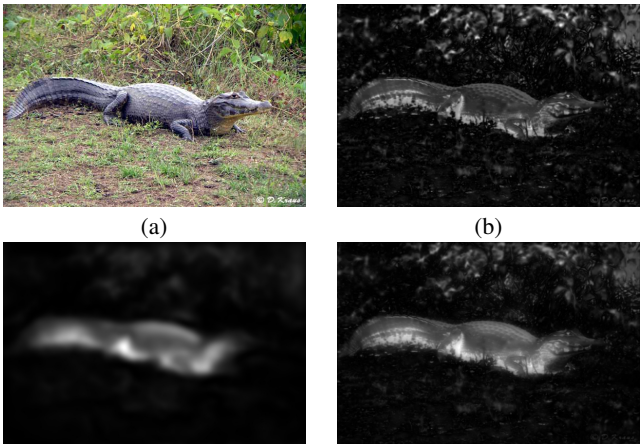


Fig. 2. Saliency map flowchart (part 1)

To obtain the hierarchical representation, we consider the near and far vision of a human observer. With the far vision, the shape of large objects is perceived without details. At the opposite, the near vision highlights the contours and the textures, but it extracts too many small regions and does not inform about the main perceived object. The idea then is to combine the near and far vision to aggregate contours and details to the main object of the scene. This procedure is illustrated on figure 3.

In addition to the far and near vision, the central bias property is a very important aspect to be included in the model. It assumes that the first gaze of an observer is always on the center of the image. To



(a) (b) (c) (d)

Fig. 3. Saliency map (d) based on pooling CSF near vision (b) and CSF far vision (c) of the original image (a).

simulate this effect for far vision, we used a spatial Gaussian model to increase the importance of the central part of the image.

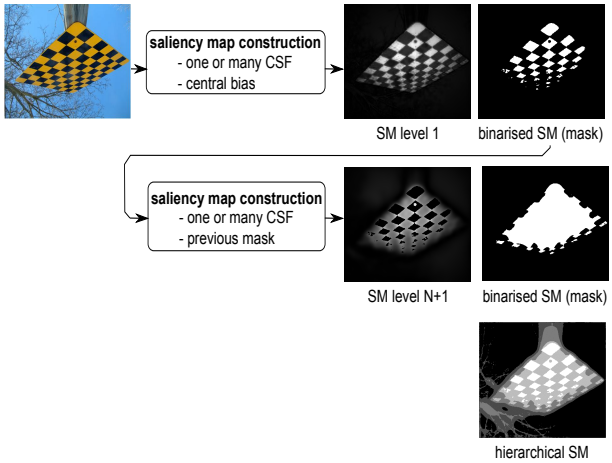
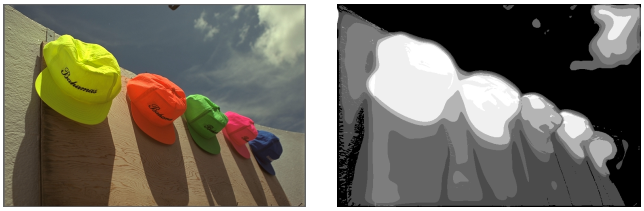


Fig. 4. Hierarchical Saliency map construction flowchart (part 2)

Results of the proposed model are given in figure for "Caps" image with 6 levels of hierarchy. It shows clearly that some objects are perceived before the others and thus the judgment of quality will be done in such a way.

In natural scene, there is not only one salient object, but many of them at different scales of importance. We can imagine that there is a kind of hierarchy in terms of objects. So, in order to extract objects of lower level it is necessary to mask all the previous extracted objects to make them more salient. Figure 4 shows the flowchart for multiple iterations of the SM algorithm to produce a hierarchy of saliency maps sorted by the importance of objects. The first level of saliency, applying on the original image two CSF filtering, one to simulate the far vision, and the other to simulate the near vision. We generate the saliency map for these two filtered images with the method detailed in Figure 2. After pooling these two saliency maps we simulate the central bias. The result is the SM level 1. For the



(a) (b)

Fig. 5. Hierarchical Saliency Map of 6 layers (b) from "Caps" original image (a).

following levels, the binarised SM of the previous level is used to mask the previous salient detected regions. The procedure above is re-iterated N times depending on the desired number of levels. The hierarchical SM is obtained by a fusion of the different layers where the gray level represents the importance of the layer.

The hierarchical saliency map obtained by the developed model is integrated in the metric scheme as described in the next section.

4. PROPOSED METRIC DESCRIPTION

The proposed metric tends to predict the perceived quality by using the evolution of interest points combined with a hierarchical saliency map. The correlation with human judgment improvement by using SM has been demonstrated by Tong et al. [12] on existing metrics. So the spatial distribution of distortions has an effect on perceived quality.

The proposed metric, named QIP-HSM (Quality by Interest Points based on a Hierarchical Saliency Map) can be summarized by the flowchart given in figure 6. The idea is to measure the evolution of interest points at each level of saliency. So the original image is used for constructing the hierarchical saliency map of n levels and extracting Harris interest points. Then, for each level, the number of interest points is computed. The impaired image uses the hierarchical saliency map of the original image and is processed in the same way to extract the interest points of each layer of saliency. The quality score is obtained by pooling the quality difference of each layer.

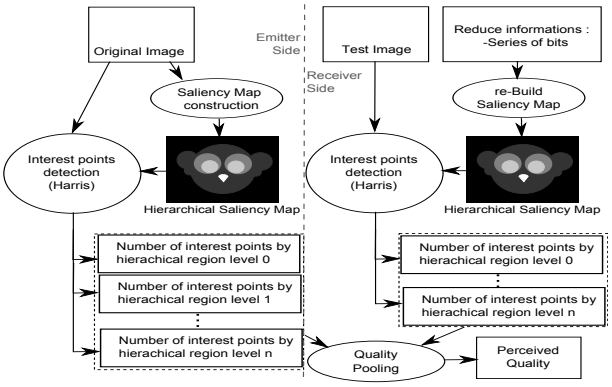


Fig. 6. Flowchart of the QIP-HSM (Quality by Interest Points based on a Hierarchical Saliency Map)

The pooling procedure takes in consideration the importance of

the information contained in the different saliency layers. The metric score is computed by assigning weighting factors for the foreground layer, the background layer and the intermediate layers as given by the equation 1

$$Q = w_f \times S_f + \sum_{i=2}^{n-1} w_{mi} \times S_{mi} + w_b \times S_b \quad (1)$$

where S_f and S_b are respectively the predicted quality for layer 1 (foreground) and layer n (background). The number of saliency levels (n) is variable and depends on the targeted application and the reduced information constraint. For the validation of the metric, we fixed this number to 6 empirically. Each S_* is computed with the formula 2. w_* are weighting factors for tuning the metric in such a way to mimic the HVS and thus give more or less importance for detected distortions.

$$S_* = \begin{cases} \frac{1-VDiff}{maxP} & if maxP > 0 \\ 1 & else \end{cases} \quad (2)$$

where $VDiff$ is the absolute difference of the number of interest points between the original and impaired layers and $maxP$ is the maximal number of detected points.

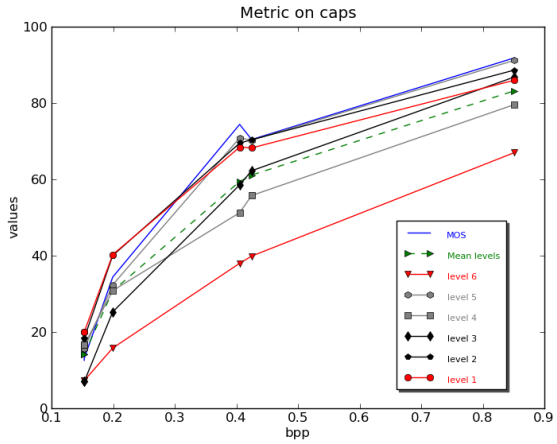
5. EXPERIMENTAL RESULTS AND DISCUSSION

Caps Sailing3

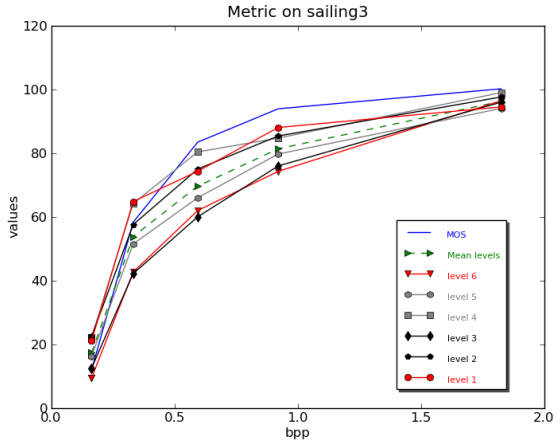
Fig. 7. Caps and Sailing3 images.

In order to prove the relevance of the interest points evolution on each layer of the hierarchical saliency map, it is necessary to run several experiments. We focused this study on JPEG compression distortions. The original and distorted images and their subjective scores (MOS) are obtained from the LIVE2 image database [13]. Two images (Caps and Sailing3) have been chosen for giving more details of the development of the metric because they have different types of content.

Figures 8-a and 8-b show the quality estimation (S_*) separately for each saliency layer from the background to the foreground. We can notice that each layer provides a quality score with the compression bitrates. However, the behavior of layer 1 (S_f) is very close to human judgment (MOS). These remarks are quite the same for both images (caps and sailings3). Intermediate layers (S_{mi}) give coherent results with some variations. The slope of layer 6 (S_b) is similar to the MOS but there is an important gap between them for some images. As a first conclusion, one can say that each saliency layer allows to predict the quality of an image but some of them are more suitable and accurate. However, the quality prediction cannot rely



(a)



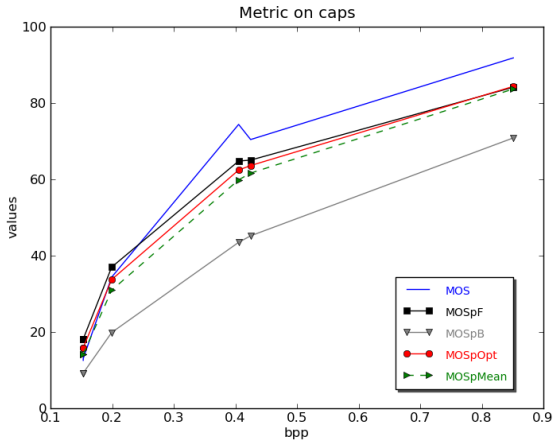
(b)

Fig. 8. Quality scores vs. bitrate for JPEG compression for two images "Caps" and "Sailing3": (a,b)- Quality prediction capabilities of each layer of the hierarchical saliency map MOS represents the human judgment.

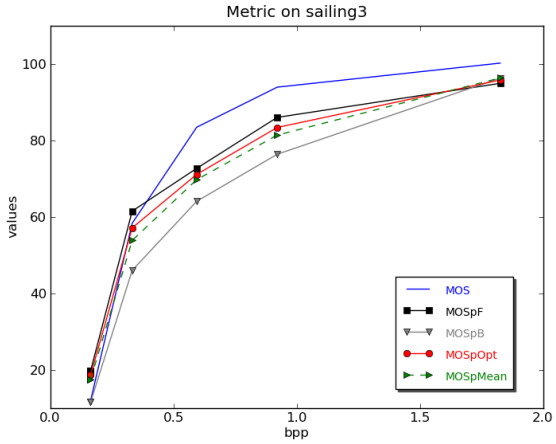
on one layer only because artifacts may affect specific areas whose are out of the selected one.

The adopted procedure consists in merging the scores of the 4 intermediate layers as described in equation 1. This means that we have to define 3 weighting factors (background, intermediate and foreground). Figures 9-a and 9-b show 4 different weighting approaches as defined by the Table 1. MOSpF, MOSpB and MOSpMean maximize respectively the first, the last and the intermediate layers of saliency. The MOSpF and MOSpMean give the best results. It means that the foreground and the intermediate layers play an important role in the prediction of the quality. From the previous remarks, we fixed the optimal weighting by giving more importance for the foreground and intermediate than the background. It corresponds to MOSpOpt row (Table 1). The selected set of weighting factors has shown on the whole Live2 database a strong coherence with human judgment.

Let us note that, MOSpOpt weighting factors depend on the hi-



(a)



(b)

Fig. 9. Quality scores vs. bitrate for JPEG compression for two images "Caps" and "Sailing3": (a,b)- comparison of different weighting strategies. MOS represents the human judgment.

erarchical saliency map and its number of layers. Some restrictions have been observed when constructing this map as the size and the compactness of the foreground layer. For example, the foreground area has to represent at least 5% of the total image size in order to have a representative layer. Knowing that the latter is the first gaze of the observer.

At this stage, all the major points of the metric are fixed. In order to illustrate the importance of using the hierarchical saliency map, we tested the Harris interest points with different configurations: Without Any information about the content, by separating the image into high activity and low activity areas (QIP) and by using the hierarchical saliency map (QIP-HSM). Figures 10-a and 10-b give the behavior of the three configuration in comparison with the human judgment (MOS).

The figures show that the usage of interest points for predicting quality of impaired images is relevant because the curves are close to the MOS and have a quite similar behavior. By separating interest points coming form flat regions and textured regions, it is clear that

Table 1. weights factors (w_* values)

Predicted score	w_f	w_m	w_b
MOSpF	0,8	0,1	0,1
MOSpB	0,1	0,1	0,8
MOSpOpt	0,4	0,5	0,1
MOSpMean	1/6	4/6	1/6

we improve the prediction capacity of the interest points. Of course, this aspect depends on the content of the image and the proportion of flat/textured regions. Finally, we can note than the use of a hierarchical saliency map improves the quality prediction and thus the performance of the proposed metric QIP-HSM.

Table 2. Performance on LIVE2 JPEG database

D\M	VSNR	PSNR	VIF	Pdiff
Corr.	0,951	0,905	0,949	0,937
RMSE	0,208	0,354	0,203	0,167
D\M	PSNR HVS	SSIM	IFC	QIP HSM
Corr.	0,956	0,976	0,915	0,979
RMSE	0,203	0,203	0,262	0,243

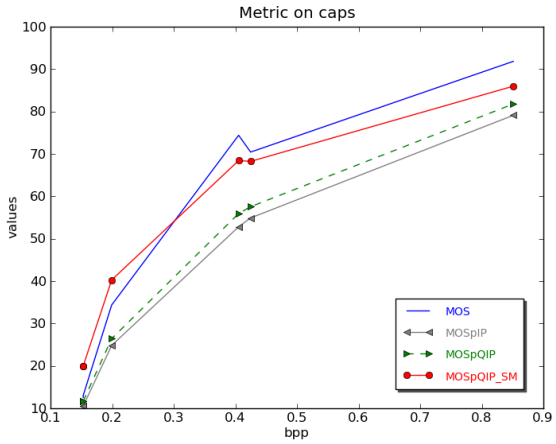
In order to study the performance of the proposed metric with regards to state-of-the-art metrics such as VSNR, PSNR, VIF, Pdiff, PSNR- HVS, SSIM, IFC . . . , we used two criteria: the prediction accuracy (RMSE) and prediction monotonicity (Pearson correlation) on LIVE2 image database for JPEG and JPEG 2000 compression with realigned subjective scores. The results of the performance study are given in Table 2 and Table 3. We can say that the proposed metric (QIP-HSM) has a good correlation and RMSE results in comparison to metrics from literature. Knowing the the proposed metric can be transformed in a reduced-reference metric by providing, at the receiver, the hierarchical saliency map and the respective scores.

Table 3. Performance on LIVE2 JPEG 2000 database

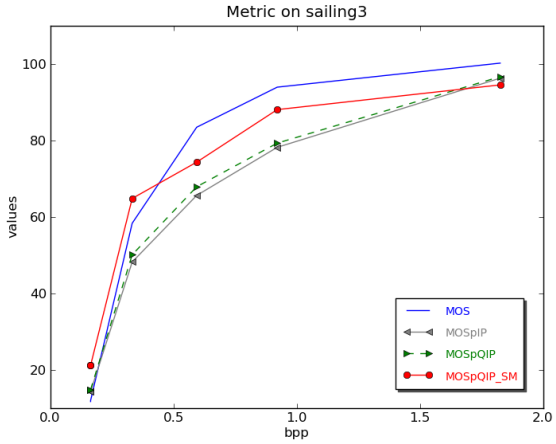
D\M	VSNR	PSNR	VIF	Pdiff
Corr.	0,960	0,917	0,959	0,972
RMSE	0,208	0,354	0,203	0,167
D\M	PSNR HVS	SSIM	IFC	QIP HSM
Corr.	0,970	0,961	0,938	0,978
RMSE	0,203	0,203	0,262	0,243

6. CONCLUSION

In this paper, we proposed a new image quality metric using interest points extracted from different regions obtained by using a hierarchical saliency map. The latter is a bio-inspired tool obtained by modeling several properties of the HVS, like central bias, achromatic and chromatic contrast sensitivity and temporal inhibition. The hierarchy of saliency is useful to decompose the content of an image into different importance layers. The major innovation lies in combining the use of interest points for the prediction of perceived image quality with the saliency hierarchy. This metric has a low complexity, the range of quality prediction is large; so each score is explicit



(a)



(b)

Fig. 10. Quality scores vs. bitrate for JPEG compression for two images "Caps" and "Sailing3": (a,b)- Comparison of several ways of using Interest Points for quality prediction. MOS represents the human judgment.

and understandable. The obtained results are very encouraging and demonstrate a high correlation with human judgment. One important conclusion of this study is that interest points represent a good way to predict image quality. Some future works can be considered as changing Harris detector by SIFT or SURF detector to deal with different properties. An important extension can be the integration of temporal information for video quality assessment.

7. REFERENCES

[1] Marius Pedersen and Jon Yngve Hardeberg, "Survey of full-reference image quality metrics," in *GCIS*, Gjovik, Norway, 2009.

[2] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," in *IEEE transactions on image processing*, 2000, pp. 636–650.

[3] H. Yee, "A perceptual metric for production testing," *Journal of Graphics Tool*, pp. 33–40, 2004.

[4] H. R. Sheikh, *Image Quality Assessment Using Natural Scene Statistics*, PhD thesis, University of Texas at Austin, 2004.

[5] 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 Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[6] D. Hands, D. Bayart, A. Davis, and A. Bourret, "No reference perceptual quality metrics: approaches and limitations," in *HVEI XIV*, feb 2009, vol. 7240.

[7] U. Engelke and H.-J. Zepernick, "Perceptual-based quality metrics for image and video services: A survey," in *3rd EuroNGI Conference on Next Generation Internet Networks*, Trondheim, Norway, 2007.

[8] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[9] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned Salient Region Detection," in *IEEE CVPR*, 2009.

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis, ," *IEEE Transactions on PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[11] S. Daly, "The visible differences predictor : An algorithm of image fidelity," *Digital Images and Human Vision*, pp. 19: 179–206, 1993.

[12] Y. Tong, H. Konik, F. Alaya Cheikh, and A. Tremeau, "Full reference image quality assessment based on saliency map analysis," *JIST*, pp. 54: 030503–(14), 2010.

[13] H.R. Sheikh, Z. Wang, L. Cormack, and A.C. Bovik, "Live image quality assessment database release 2," http://live.ece.utexas.edu/research/quality.