Issues in the Design of a No-Reference Metric for Perceived Blur

Hantao Liu*[a] and Ingrid Heynderickx[a,b]
[a]Department of Mediamatics, Delft University of Technology, Delft, The Netherlands
[b]Group Visual Experiences, Philips Research Laboratories, Eindhoven, The Netherlands

ABSTRACT

Developing an objective metric, which automatically quantifies perceived image quality degradation induced by blur, is highly beneficial for current digital imaging systems. In many applications, these objective metrics need to be of the no-reference (NR) type, which implies that quality prediction is based on the distorted image only. Recent progress in the development of a NR blur metric is evident from many promising methods reported in the literature. However, there is still room for improvement in the design of a NR metric that reliably predicts the extent to which humans perceive blur. In this paper, we address some important issues relevant to the design as well as the application of a NR blur metric. Its purpose is not to describe a particular metric, but rather to explain current concerns and difficulties in this field, and to outline how these issues may be accounted for in the design of future metrics.

Keywords: Image quality assessment, objective metric, perceived blur, edge, visual attention

1. INTRODUCTION

The past decades have witnessed a revolutionary growth in the development of reliable and effective image quality assessment techniques for current digital imaging systems. Many scientists have contributed to a better understanding of human vision based quality perception via subjective testing, and to the design of objective metrics for the automatic prediction of perceived image quality. Especially, in the latter area considerable progress in the development of reliable objective metrics for their respective application domains is reported in the literature [1].

Objective metrics can be classified into full-reference (FR) or no-reference (NR) metrics. FR metrics are based on measuring the similarity between the distorted image or video and its original version. This approach is rather straightforward and has been extensively investigated [2]-[5]. Its applicability, however, is limited by its inherent characteristic that access to the original image material is required. In many real-world applications, such as in the video chain of television (TV)-sets and in video streaming via the Internet, the original image material is mostly not available. Hence, for these applications NR metrics are highly needed.

In recent years, fast development of multimedia techniques has pushed the demand for reliable NR metrics, and a lot of research effort has been devoted to this field [6]-[15]. However, automatically assessing image quality without the use of the original image material is still very challenging, partly due to our limited understanding of how the human visual system (HVS) judges image quality. Fortunately, image distortions, such as blur caused during acquisition, sensor noise, and blockiness or ringing as a consequence of signal compression, are well studied and classified. Modeling a specific type of distortion is considered as a more realistic NR approach towards perceived quality assessment [7]-[13]. Furthermore, in practical imaging systems, the distortion processes involved are often known and fixed, and thus, the design of dedicated NR metrics accounting for a specific type of artifact created by a specific image distortion process is of sufficient added value. They can, for example, be combined to an overall perceived quality prediction. Various examples of this approach are given in literature; e.g., a ringing and a blur metric are often combined to assess the quality of wavelet based compressed images (see, e.g., [14] and [15]). In addition, these dedicated artifact metrics are beneficial individually for optimizing real-time digital imaging systems; e.g., in the video chain of current television-sets various NR metrics of individual artifacts are used to adapt the parameter settings of the video enhancement algorithms accordingly (see, e.g., [16] and [17]).

Blur is one of the most annoying artifacts, and thus, has an essential contribution to image quality assessment. Developing a NR metric that automatically quantifies perceived blur is of fundamental importance to a broad range of applications, such as the optimization of auto-focus systems, super-resolution techniques, and sharpness enhancement in displays. Existing blur metrics are formulated either in the spatial domain or in the frequency transform domain. The

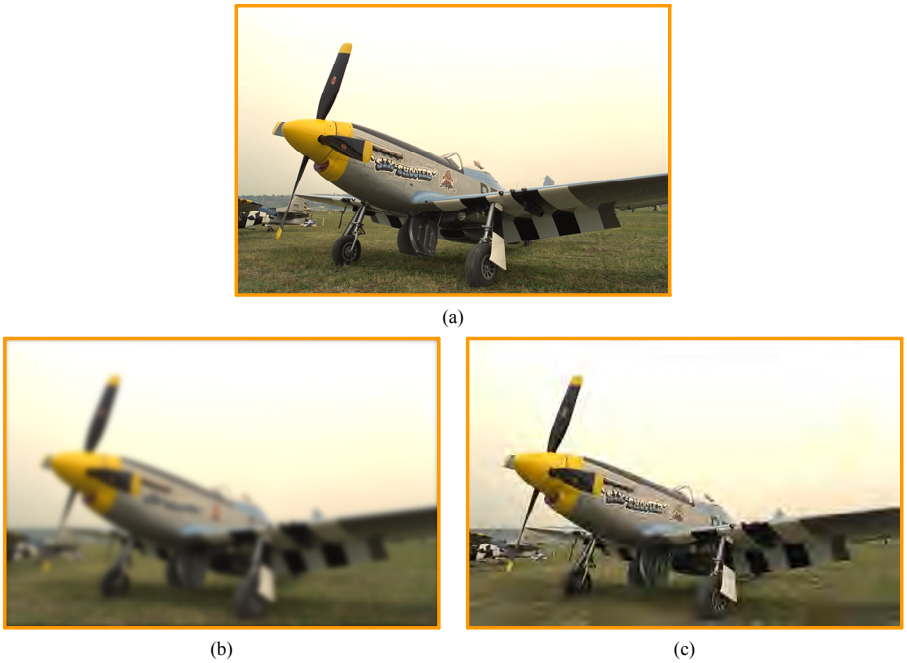
Image Quality and System Performance VIII, edited by Susan P. Farnand, Frans Gaykema, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7867, 78670C · © 2011 SPIE-IS&T · CCC code: 0277-786X/11/$18 · doi: 10.1117/12.873277

metrics implemented in the transform domain (see, e.g., [18]-[20]) are motivated by the fact that blur is intrinsically caused due to the attenuation of high spatial frequencies. They usually involve a rather complex calculation of energy falloff in the DCT or wavelet transform domain. Moreover, they often require access to the encoding parameters, which are, however, not always available in practical applications. A blur metric defined in the spatial domain generally relies on measuring the spread of edges in an image (see, e.g., [7]-[9]), since blur is perceptually apparent along edges or in textured regions. This approach is quite straightforward and has its potential in estimating image quality locally or in its incorporation in an overall quality assessment metric in combination with other artifacts. Despite the many contributions to NR blur metrics in literature, there is still room for improvement in their reliability, and so, this research area is still far from mature. In this paper, we intend to address some interesting issues relevant to the development and application of NR blur metrics. To our opinion, some of these issues are insufficiently highlighted in the literature, and so, addressing them can be beneficial for future research in the design of more reliable NR blur metrics.

2. ISSUES IN THE DESIGN OF A NR BLUR METRIC

Since our main focus is on simple, computationally inexpensive metrics that can be used in real-time applications, we limit ourselves to the spatial domain approach for the design of a NR blur metric. Based on state-of-the-art metrics and data of perceived blur available in the literature, we attempt to explain current concerns and difficulties in the design of a NR blur metric. In addition, we discuss in either quantitative or qualitative terms how these issues may be accounted for.

2.1 Classification of NR blur metrics



(a)

(b) (c)

Figure 1. Illustration of various types of blur: (a) a source image extracted from the LIVE image quality assessment database [21], (b) a Gaussian blurred version of image (a) (with a standard deviation of the Gaussian kernel of σ=4.92), and (c) a JPEG2000 compressed version of image (a) (with a bit rate of bits/pixel=0.054).

The performance of a NR metric may be quite application dependent, which implies that a metric designed for a specific distortion type is not necessarily able to reliably predict the quality related to other types of distortions. More specifically, we want to explicitly point out that a blur metric that is designed to predict the quality degradation of Gaussian blurred images may not be useful (at least not without adaptations) for the overall quality assessment of JPEG2000 compressed images, and vice versa. Figure 1 illustrates typical differences in perceived blur between a Gaussian blurred image and a JPEG2000 compressed image; in both cases blur is perceived, albeit in a different way. In a Gaussian blurred image, the degradation in image quality is solely induced by blur, and consequently, using a blur metric to approximate the overall perceived quality is reasonable and feasible. For example, the metric described in [7] shows that the perceived quality of Gaussian blurred images can be simply estimated from calculating the averaged local edge width. In a JPEG2000

compressed image, two annoying artifacts i.e. blur and ringing, are coexisting, and their individual annoyance varies depending on the image content and the compression level. Therefore, directly applying a blur metric to predict the overall quality of JPEG2000 compressed images becomes difficult, and in general, unreliable. This explains why, as in [14], a generally designed blur metric yields a lower correlation for JPEG2000 compressed images than for Gaussian blurred images. Unfortunately, this issue of different types of perceived blur often gets insufficient attention in the design and evaluation of a blur metric.

The concern mentioned above actually raises the question whether a dedicated blur metric for Gaussian blur should necessarily be tested against JPEG2000 compressed images, since in the latter case the human perception of blur might be strongly affected by the presence of ringing artifacts. In this respect, two research questions become relevant in the case of JPEG2000 compression: (1) how to design a dedicated blur metric for JPEG2000 compressed images, and (2) how to assess the overall quality of JPEG2000 compressed images. The former question implies that a dedicated metric for perceived blur in JPEG2000 compressed images should take the presence of ringing artifacts into account, and its performance should be tested by requesting subjects to assess blur annoyance (instead of overall quality) of JPEG2000 compressed images. The latter question can be addressed by combining a blur metric with a ringing metric, assuming that they both are well designed and validated individually for JPEG2000 compressed images (see, e.g., [15]).

Thus, to make a metric design more efficient, and to ensure a fair comparison between metrics, we should be aware of the targeted application of the NR blur metric. Last, but not least, it is important to note that when comparing alternative metrics, the performance of these metrics needs to be evaluated in a comparative setting (e.g. on the same database and with the same fitting strategy), so that all metrics' strengths and weaknesses can be fairly analyzed.

2.2 Effect of edge detection on NR blur metrics

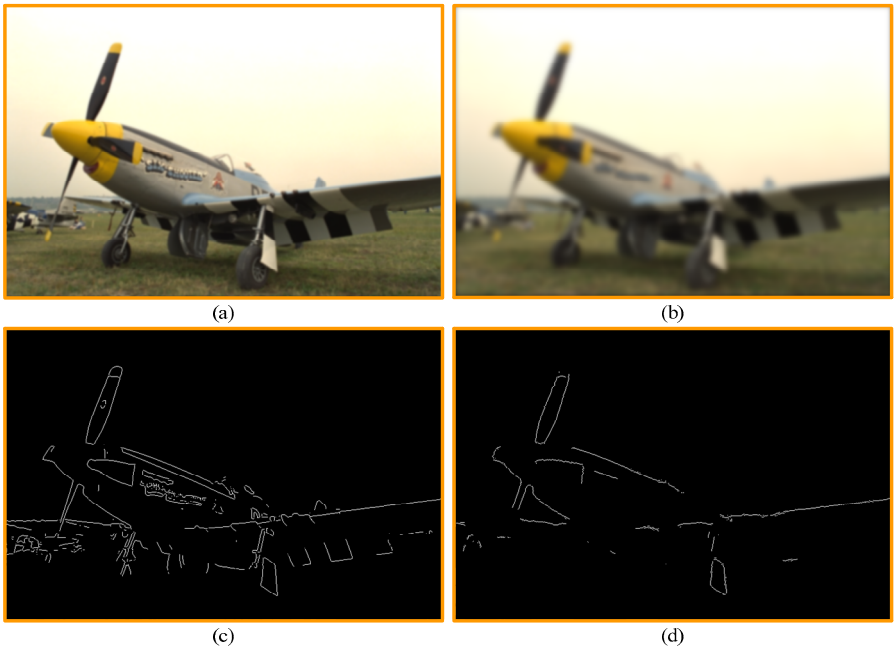


Figure 2. Illustration of the use of a Sobel edge detector on Gaussian blurred images: (a) a Gaussian blurred image extracted from the LIVE database [21] (with the standard deviation of the Gaussian kernel σ= 1.708), (b) a Gaussian blurred image extracted from the LIVE database [21] (with the standard deviation of the Gaussian kernel σ=4.917), (c) Sobel edge map of (a), and (d) Sobel edge map of (b). The Sobel edge detection method is implemented by the MATLAB's "edge" function, in which the default sensitivity threshold (i.e. using the mean of the gradient magnitude squared image) is used.

Most, if not all, of the existing NR blur metrics defined in the spatial domain start from detecting prominent edges in an image (see, e.g., [7]-[9]). The difference between the resulting metrics comes from how the visibility of blur is measured around these edges, ranging from the very simple method described in [7] to the rather complicated concept of JNB (Just Noticeable Blur) described in [9]. The methods employed in [7]-[9] to capture strong edges use an ordinary edge detector, such as a Sobel operator. It implies that a certain threshold is applied to the gradient magnitudes to remove noise and

insignificant edges. Figure 2 illustrates typical performance of a Sobel edge detector applied to Gaussian blurred images: as the level of blur increases, the amount of edges detected by the Sobel detector decreases. This may affect the performance of a NR blur metric, since it reduces the number of local blur measurements, as a consequence of which the measurement may become more sensitive to noise, and may lose some of its prediction accuracy in overall perceived image quality.

In addition, the reduction in the amount of detected edges may degrade the reliability of the NR blur metric in applications where local values of blur induced quality degradation in the images are needed. In some practical applications, a spatially varying quality degradation profile is used to drive local signal processing, such as de-blurring or sharpness enhancement algorithms. These algorithms in general are able to locally adapt their parameter settings in order to optimize the overall perceived quality for the viewer. In such a scenario, preserving all perceptually relevant edges for the local blur estimation is rather important. For example, in Figure 2(d) the Sobel edge map eliminates a number of perceptually important edges on the contour of the airplane, which obviously would benefit from sharpness enhancement of the object. Thus, developing a more advanced edge detection method particularly for a NR blur metric would be greatly beneficial for improving its robustness and for its use in more demanding applications. It, however, should be noted that including a more reliable edge detector inevitably increases the computational complexity of a blur metric. Thus, it is important to be aware of the targeted application of a blur metric, and to adapt its design accordingly.

2.3 Content independency of NR blur metrics

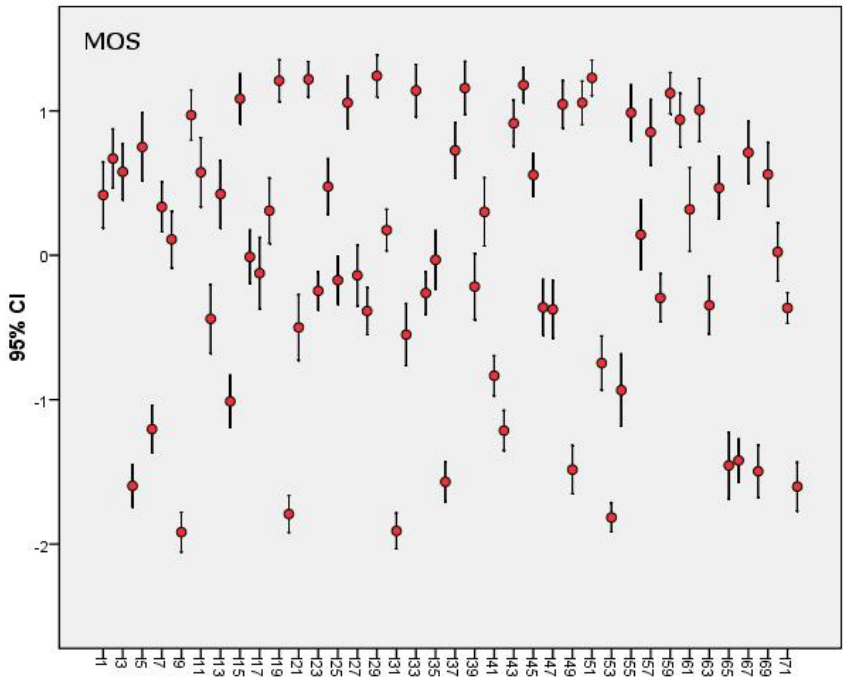
Robustness against image content is one of the essential requirements for an objective metric. Understanding how humans perceive image quality attributes in an adequate diversity of image content is essential for further improving the performance of objective metrics. Robustness against image content, however, is also the most difficult aspect in the metric design, especially in the case of a NR metric. In the past five years, scientists in the area of objective metric design devoted part of their research effort on the collection and distribution of more experimental data from subjective tests (see, e.g., [21]-[24]). These data can now be used as benchmark for the performance of objective metrics on a wide variety of content. In the evaluation of blur metrics, the sub-set of the LIVE image quality database containing Gaussian blurred images (i.e. hereafter referred to as LIVE blur database) [21] is extensively used for the evaluation of objective blur metrics. This database contains a set of twenty-nine source images that attempts to cover sufficient diversity in image content, including pictures of faces, people, animals, close-up shots, wide-angle shots, natural scenes, man-made objects, images with distinct foreground/background configurations, and images without any specific object of interest. All these source images are filtered using a circular-symmetric 2-D Gaussian kernel of standard deviation σ_B, ranging from 0.42 to 15 pixels, which results in a set of 174 stimuli (including the source images). In addition, this database contains for each stimulus a quality score, obtained with a single-stimulus set-up, in which subjects are requested to assess the quality of all test stimuli, including the original images considered as being the reference. The resulting quality scores were further processed subtracting the quality scores for the test stimuli of the quality scores for the corresponding reference, yielding a difference mean opinion score (DMOS). The database is generally used to evaluate the overall performance of blur metrics; however, it has its limitation in evaluating metrics' performance on more demanding images. For example, the database contains a very limited number of highly textured images. Hence, evaluating the performance of an objective blur metric on this particular type of images with the LIVE blur database may be insufficiently reliable. Therefore, we created an additional database (i.e. first published in [25]) specifically with highly textured images blurred at different levels.

To this end, a perception experiment was conducted at the Delft University of Technology [24]. A set of 12 source images of highly textured content was collected. The source images are illustrated in Figure 3(a). These images were high quality colored images of size 512×768 (height×width) pixels. They were blurred in the same way and with the same range of σ_B as for the LIVE blur database. Thus, each source image was blurred at five different levels, yielding a test database of 72 stimuli (including the source images). The stimuli were displayed on a Dell 24" LCD screen with a native resolution of 1920×1200 pixels. The experiment was conducted in a standard office environment and the viewing distance was approximately 70cm. Eighteen participants, being ten male and eight females, were recruited for the experiment. We followed the single-stimulus (SS) protocol as described in [25] to carry out the image quality assessment and the method described in [11] to produce the mean opinion scores (MOS). This new database is named as the highly textured image (HTI) database, and the resulting MOS are shown in Figure 3(b).

To give an example of the use of the HTI database, we implemented the NR blur metric described in [7], which is based on the calculation of local edge width (i.e. hereafter referred to as NRPB). We tested this objective metric on both the LIVE blur database and the HTI database. Figure 4 illustrates the scatter plot of the (D)MOS versus the NRPB for both

databases. The quantitative evaluation of the metrics' performance in terms of Pearson linear correlation coefficient (CC), Spearman rank order correlation coefficient (SROCC), and root mean squared error (RMSE) are: (1) for the LIVE blur database: CC=0.78, SROCC=0.92 and RMSE=3.02, and (2) for the HTI database: CC=0.89, SROCC=0.94 and RMSE=1.93. In general, NRPB exhibits a poor correlation with perceived quality induced by blur, as also shown here for the LIVE blur database. It, however, shows a relatively high prediction accuracy when handling the highly textured images in the HTI database. This interesting finding motivates us to further investigate which types of image content the metric is not able to handle very well, to improve the reliability of the metric further based on that information. In this respect, we have decided to make the subjective data freely available to the research community so that other researchers can easily use it in their design and evaluation of blur metrics. Meanwhile, we expect that more experimental data will be collected and shared in the research community.

(a)



(b)

Figure 3. The highly textured image (HTI) database: (1) source images, and (2) the mean opinion score (MOS) of the 72 Gaussian blurred images of the HTI database (the error bars indicate the 95% confidence intervals).

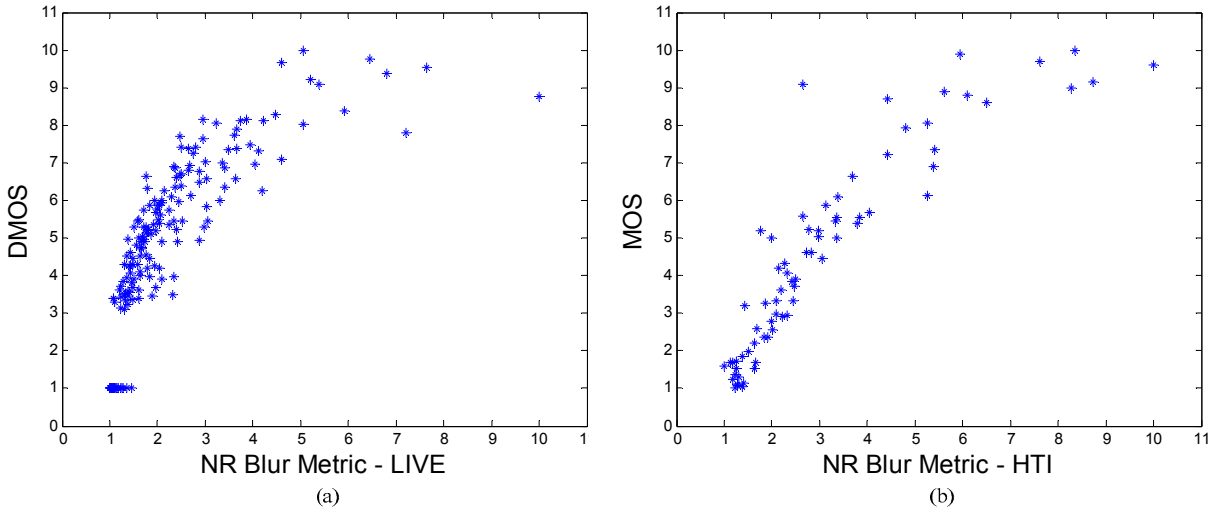
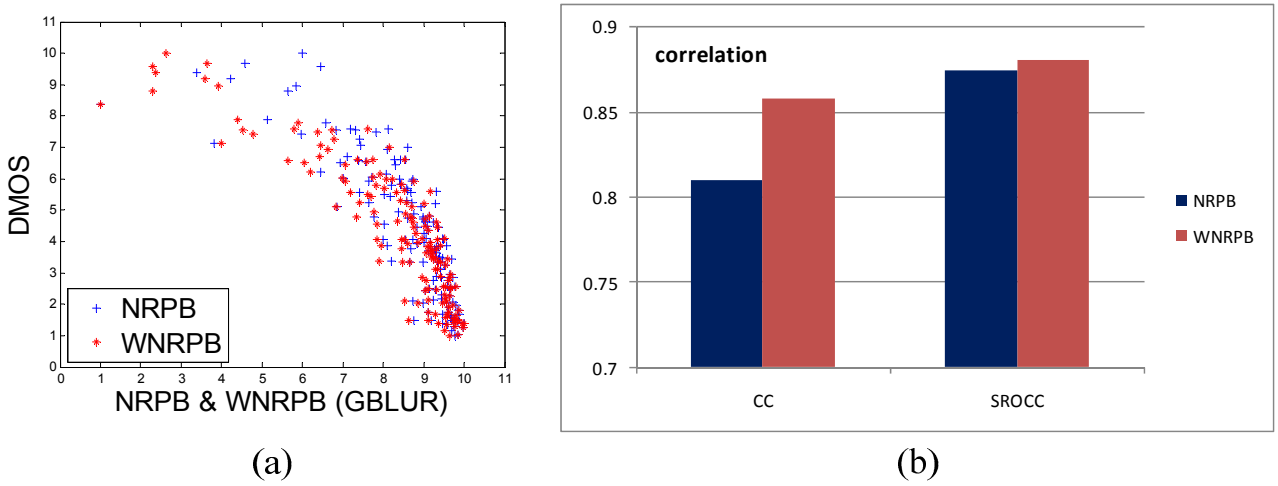


Figure 4. Scatter plot of (D)MOS vs. the NR blur metric as described in [7] for the LIVE blur database (a) and the HTI database (b).

2.4 Added value of visual attention in NR blur metrics

Current research on image quality assessment tends to include visual attention in objective metrics to further enhance their performance in predicting perceived quality. To this end, a variety of computational models of visual attention is implemented in different metrics by weighting local distortion maps with local saliency maps, a process referred to as "visual importance pooling" (see. e.g., [26] and [27]). The attention models used in these studies, however, are either specifically designed or chosen for a specific domain, and their accuracy in predicting human attention in general terms is not always fully proved yet. To circumvent this issue, we decided to use "ground truth" visual attention data instead of using a computational model. These "ground truth" visual attention data were obtained from eye-tracking experiments, thus making the evaluation of adding visual attention in objective metrics independent of the reliability of an attention model. First results adding saliency to the PSNR and SSIM metrics showed an improvement in their performance predicting perceived image quality [28].

In addition, we want to investigate here whether adding visual attention data obtained from eye-tracking measurements is beneficial to the performance of the NRPB metric. We thus follow the same approach as described in [28]. Natural scene saliency (NSS), obtained from asking people to freely look to the images, is included in the metric by locally weighting its distortion map (i.e. the local blur values calculated on the edges of an image). This results in an attention-based metric, which is referred to as WNRPB. The two metrics, NRPB and WNRPB are applied to the LIVE blur database. Figure 5 shows the scatter plot of the DMOS versus the two metrics, and the corresponding correlation coefficients CC and SROCC. It shows that there is indeed a gain in performance including visual attention in the blur metric, e.g. the gain of WNRPB over NRPB corresponds to an increase in the Pearson correlation coefficient of 5%. The experimental results intrinsically indicate that the blurred edge in the salient areas is more annoying than in the non-salient areas of an image. Hence, further research on developing a saliency model and incorporating its results in the edge detection part of a blur metric looks valuable. However, it should be noted that modeling visual attention is computationally demanding, and in many real-time applications the performance gain of the attention-based metric should be balanced against the additional cost for the rather complex saliency model.



(a) (b)

Figure 5. Illustration of the performance of the NRPB and WNRPB metrics: (a) scatter plot of DMOS vs. NRPB and WNRPB for the LIVE blur database, and (b) correlation coefficients of NRPB and WNRPB for the LIVE blur database.

3. CONCLUSIONS

In this paper, we address some important issues in the design of a NR blur metric: (1) the classification of blur metrics depending on their targeted application, by which the metric can be made more efficient, and it ensures a fair comparison between metrics, (2) the effect of the method used to detect edges on the performance of a blur metric, (3) the sensitivity in performance of a blur metric in terms of content independency, and (4) the added value of including "ground truth" visual attention data in the design of a blur metric. Based on state-of-the-art metrics and data of quality assessment experiments available in the literature, we intend to discuss these issues in either quantitative or qualitative terms, in the hope that the concerns raised are beneficial for the future research in designing a more reliable NR blur metric.

REFERENCES

[1] Wang, Z. and Bovik, A. C., [Modern Image Quality Assessment], Synthesis Lectures on Image, Video & Multimedia Processing, Morgan & Claypool Publishers (2006).

[2] Wang, Z. and Bovik, A. C., "Mean squared error: love it or leave it? - A new look at signal fidelity measures," IEEE Signal Processing Magazine, vol. 26, no. 1, pp. 98-117 (2009).

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

[4] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," IEEE Transactions on Image Processing, vol.15, no.2, pp. 430- 444 (2006).

[5] Sheikh, H. R., Sabir, M. F. and Bovik, A. C., "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," IEEE Transactions on Image Processing, vol. 15, no. 11, pp, 3440-3451 (2006).

[6] Hemami, S. S. and Reibman, A. R., "No-reference image and video quality estimation: Applications and human motivated design," Signal Processing: Image Communication, vol. 25, no. 7, pp. 469-481 (2010).

[7] Marziliano, P., Dufaux, F., Winkler, S. and Ebrahimi, T., "A no-reference perceptual blur metric," in Proc. IEEE International Conference on Image Processing, vol. 3, pp. 57-60 (2002).

[8] Liang, L., Chen, J., Ma, S., Zhao, D. and Gao, W., "A no-reference perceptual blur metric using histogram of gradient profile sharpness," in Proc. IEEE ICIP, pp. 4369-4372 (2009).

[9] Ferzli, R. and Karam, L. J., "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," IEEE Transactions on Image Processing, vol. 18, pp. 717–728 (2009).

[10] Feng, X. and Allebach, J.P., "Measurement of Ringing Artifacts in JPEG Images," in Proc. SPIE, vol. 6076, pp. 74-83 (2006).

[11] Liu, H., Klomp, N. and Heynderickx, I., "A No-Reference Metric for Perceived Ringing Artifacts in Images," IEEE Trans. on Circuits and Systems for Video Technology, vol. 20, pp. 529-539 (2010).
[12] Wu, H. R. and Yuen, M., "A generalized block-edge impairment metric for video coding," IEEE Signal Processing Letters, vol. 4, no. 11, pp. 317–320 (1997).
[13] Liu, H. and Heynderickx, I., "A Perceptually Relevant No-Reference Blockiness Metric Based on Local Image Characteristics," EURASIP Journal on Advances in Signal Processing, vol. 2009 (2009).
[14] Marziliano, P., Dufax, F., Winkler, S. and Ebrahimi, T., "Perceptual blur and ringing metrics: Application to JPEG2000," Signal Processing: Image Communication, vol. 19, pp. 163-172 (2004).
[15] Liang, L., Wang, S., Chen, J., Ma, S., Zhao, D. and Gao, W., "No-reference perceptual image quality metric using gradient profiles for JPEG2000," vol. 25, no. 7, pp. 502-516 (2010).
[16] Koh, C. C., Mitra, S. K., Foley, J. M. and Heynderickx, I., "Annoyance of Individual Artifacts in MPEG-2 Compressed Video and Their Relation to Overall Annoyance," in SPIE Proceedings, Human Vision and Electronic Imaging X, vol. 5666, pp. 595-606 (2005).
[17] Kirenko, I. O., Muijs, R. and Shao, L., "Coding artifact reduction using non-reference block grid visibility measure," in Proc. IEEE International Conference on Multimedia and Expo, pp. 469–472 (2006).
[18] Marichal, X., Ma, W. and Zhang, H., "Blur determination in the compressed domain using DCT information," in Proc. IEEE ICIP, pp. 386-390 (1999).
[19] Caviedes, J. and Oberti, F., "A new sharpness metric based on local kurtosis, edge and energy information," Signal Processing: Image Communication, 18(1), pp. 147-161 (2004).
[20] Hassen, R., Wang, Z. and Salama, M., "No-reference image sharpness assessment based on local phase coherence measurement," in Proc. IEEE ICASSP, pp. 2434-2437 (2010).
[21] Sheikh, H. R., Wang, Z., Cormack, L. and Bovik, A. C., "LIVE Image Quality Assessment Database Release 2," http://live.ece.utexas.edu/research/quality
[22] VQEG (2003, Aug.): Final report from the video quality experts group on the validation of objective models of video quality assessment. Available: http://www.vqeg.org
[23] Winkler, S., "Image and Video Quality Resources," http://stefan.winkler.net/resources.html
[24] Liu, H., Redi, J. and Heynderickx, I., "TUD Image Quality Database: Perceived blur," http://mmi.tudelft.nl/iqlab/blur.html
[25] Redi, J., Liu, H., Alers, H., Zunino, R. and Heynderickx, I., "Comparing subjective image quality measurement methods for the creation of public databases", IS&T/SPIE Electronic Imaging 2010, Image Quality and System Performance VII (2010).
[26] Sadaka, N. G., Karam, L. J., Ferzli, R. and Abousleman, G. P., "A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling," in Proc. IEEE Int. Conf. ICIP, pp. 369-372 (2008).
[27] Moorthy, A. K. and Bovik, A. C., "Perceptually significant spatial spooling techniques for image quality assessment," in Proc. Electronic Imaging (2009).
[28] Liu, H. and Heynderickx, I., "Studying the Added Value of Visual Attention in Objective Image Quality Metrics Based on Eye Movement Data", in Proc. IEEE International Conference on Image Processing, pp. 3097-3100 (2009).